Molecular Modeling of Enzymes Application to the Study of Phosphoryl Transfer Reactions and the Dynamics-Function Relationship

Enrique Marcos Benteo

Universidad de Barcelona

Facultad de Química

Departamento de Química Física

Programa de doctorado en Química Teórica y Computacional (bienio: 2006-2008)

Memoria para optar al título de doctor por la Universidad de Barcelona

Molecular Modeling of Enzymes
Application to the Study of Phosphoryl Transfer Reactions and the Dynamics-Function Relationship

Enrique Marcos Benteo

Director de tesis:
Ramon Crehuet Simon

Tutor de tesis:
Albert Solé Sabaté

Para Javier,

quien también inicia una etapa de crecimiento

Agradecimientos

Con mucha frecuencia nos concentramos en el aspecto técnico de nuestra formación. Aprendemos numerosos conceptos y metodologías para abrirnos camino en una profesión, pero ¿qué hay de las experiencias que hemos vivido durante este aprendizaje? ¿Qué nos han transmitido las personas que nos hemos cruzado? Desde luego, el doctorado no sólo se caracteriza por ser una etapa de formación muy técnica. He podido comprobar cómo, al mismo tiempo, conlleva un cúmulo de experiencias con personas que dan forma a tu manera de hacer las cosas. Este es el aspecto más humano de la formación de un investigador: reconocer como maestros a aquellas personas que te guían, influyen, apoyan o inspiran. ¿Acaso no deberíamos movernos hacia una educación que, además de focalizarse en aspectos técnicos, también enseñe actitudes y valores que optimicen la generación y el uso del conocimiento? A todas las personas que han contribuido en esta etapa os quiero dar las gracias.

A Ramon, mi director de tesis, quiero agradecerle su paciencia y el tiempo tan extenso que me ha dedicado, así como las actitudes que me ha transmitido. Su curiosidad científica, rigor y capacidad para escuchar me han proporcionado libertad e iniciativa para explorar campos muy diferentes dándome una visión muy amplia de lo que se puede hacer con la química teórica. Su vocación de transmitir conocimientos trasciende la química y, sin duda, ha potenciado mi gusto por la montaña. Gracias por ser un magnífico consejero y colega de excursiones.

A Albert Solé como tutor de tesis por la buena disposición que siempre ha tenido para ayudarme en todos los aspectos burocráticos que conlleva el proceso de una tesis.

Quiero agradecerle a Josep por haberme recibido tan amablemente en su grupo al principio de los tiempos y haberme guiado durante la primera parte de esta tesis. Fue mi primer contacto con la química teórica y ha sido un referente del trabajo bien hecho. También agradecerle a Santiago su apoyo al inicio de la tesis y su ejemplo de rigor científico.

A mis compañeros del CSIC por el buen rollo que siempre he respirado en el despacho. A Aurora por ser una compañera ideal durante casi cinco años. Por su buen humor y apoyo siempre que hacía falta. A Álex por ser un ejemplo muy claro de lo que significa el orden. Mucha felicidad en vuestra nueva etapa con el pequeño Gabriel. A Javi por su agradable compañía. A Miquel por la alegría que aporta al grupo. Muy pronto lograrás la RyC. Te lo mereces! A Melchor que ha cogido el testigo tan eficientemente y, que con su gran capacidad de trabajo, seguro que hará una muy buena tesis. A otros que pasaron por

menos tiempo: Pau y Miriam. Gracias por vuestra ayuda en el último trozo de esta tesis. Y también a Miquel de Banyoles con quien fue un placer compartir despacho. Del CSIC también quiero agradecer a Patricia su simpatía y las muchas comidas que hemos compartido con Aurora en el bar. Un ejemplo de decisión y valentía para emprender aquello que deseaba hacer.

Al CESCA por haberme acogido cuando más difícil lo tenía para empezar la tesis. Quiero agradecer a todos los compañeros que tuve el ambiente tan bueno de trabajo que lograban crear con su buen humor y las ganas de trabajar. Especialmente a Ingrid y Alfred por ser unos supervisores estupendos.

Al CSIC quiero agradecerle la concesión de una beca predoctoral durante los últimos cuatro años que además me ha dado la oportunidad de hacer estancias breves que han supuesto enormes experiencias de vida.

To Ivet Bahar for her warm welcome in Pittsburgh and for believing in our project. I am indebted to her ongoing support to my scientific career. I am also grateful to all her group members for valuable discussions during my stay.

To Martin J. Field for his guidance during my stay in Grenoble. I also want to thank his support to my scientific career. To my colleagues at IBS. Thanks to Alexey for interesting scientific discussions and his kindness sharing the office. I also really appreciate the comradeship of Arijit and Pankaj.

It was also very instructive to share scientific discussions with neutron scatterers in Grenoble (G. Zaccai, F. Gabel and M. Weik) and Heidelberg (J.C. Smith). Thanks for their insight into the last part of this thesis. I would like to acknowledge all advices of Pau Bernadó in our incursion in neutron scattering.

To David Baker for gladly accepting me in Seattle and his support to my interest in enzyme design. I look forward to joining your lab in this new stage of my career. I also want to thank Modesto for supporting this joint project.

A mis compañeros de los cursos de doctorado interuniversitarios que hicimos en Santander. Gracias por aquel inolvidable mes de concentración.

A Nayra, Manuel y Marta por ser mi familia en Pittsburgh.

A mis amigos de Dale Carnegie. Habéis sido una gran inspiración en el último trayecto de esta etapa que termino. Gracias por inyectarme vuestro entusiasmo y recordarme que el mundo cambia cuando tú cambias. Sois geniales. Que todas vuestras visiones se hagan realidad.

Por último quiero darles las gracias a mis incondicionales.

A Dani, Casadó y Baltà por su larga amistad y con quienes he compartido tantas discusiones filosóficas.

A Manu, Ragàs, Berzosa y Juan por ser ejemplos de paciencia, del trabajo bien hecho y por lo bien que nos lo hemos pasado estos años. Gracias Manu por tu asesoramiento estadístico y al resto de "Piltrafillas" que conforman este grupo de buena gente y corazón sano.

A mi familia por creer en mí aún sin entender lo que estaba haciendo. Quiero agradecer a mis padres la gran inspiración que han supuesto para mí y regalarme la oportunidad de construir mi propio camino. Sois mi referencia. A mis hermanos Carolina y Javier por ser ejemplos de fuerza y valentía. A Jorge por su tutela y confianza desde que empecé la universidad. A mis queridos sobrinos: Álvaro, Irene, Diego y Álex por la alegría que me aportan.

A ti Blanca que iluminas el camino siendo como eres.

CONTENTS

1. Introduction 1

1.1. Enzyme catalysis 1

1.1.1. Experimental techniques to study protein dynamics 3

1.1.2. Computational techniques to study protein dynamics 5

1.1.3. Controversy on the coupling between enzyme dynamics and catalysis 5

1.2. Phosphorylation reactions by enzymes 7

1.2.1. Mechanisms of phosphorylation 7

1.2.2. Hypervalency in pentacoordinated phosphorus 10

1.2.3. Pentacoordinated phosphorus in enzymes: controversy on β-phosphoglucomutase 11

1.3. Amino Acid Kinase family: large-amplitude motions mediating function 13

1.3.1. N-Acetyl-Glutamate Kinase (NAGK) 13

1.3.2. Carbamate Kinase (CK) 16

1.3.3. Uridine Monophosphate Kinase (UMPK) 17

1.4. Thermostability in enzymes 19

1.4.1. Sequence and structural requirements for protein thermostability 19

1.4.2. Dynamical requirements for protein thermostability 21

1.4.3. Exchange between thermostability and catalytic activity 22

1.5. References 23

2. Objectives 29
3. Methodology 31
3.1. Quantum-Mechanical methods 31
3.1.1. Hartree-Fock method 33
3.1.2. Post Hartree-Fock methods 35
3.1.3. Semi-empirical methods 38
3.1.4. Basis sets 40
3.1.5. Density Functional Theory methods 42
3.1.6. Wave function analysis 46
3.2. Molecular Mechanics 49
3.2.1. Bonding interactions 49
3.2.2. Non-bonding interactions 50
3.3. Hybrid Quantum Mechanical / Molecular Mechanics methods 53
3.4. Potential energy surface: stationary points and conformational sampling 55
3.4.1. Location of stationary points 55
3.4.1.1. Energy minimization methods 56
3.4.1.2. Transition state structure and reactions paths 58
3.4.2. Sampling methods 59
3.4.2.1. Molecular dynamics 59
3.4.2.2. Elastic Network Models 64
3.4.2.3. Brownian dynamics 68
3.5. Calculation of Elastic Incoherent Neutron Scattering properties 72
3.6. Bibliography and references 76
4. Results 81
4.1. Phosphoryl transfer reactions 82
4.1.1. Pentacoordinated phosphorus: structure, reactivity and biological implications 82

4.1.2. Pentacoordinated Phosphorus: methodologies to describe polarization effects 100

4.1.3. Pentacoordinated Phosphorus in β-phosphoglucomutase 112

4.2. Dynamical properties of the Amino Acid Kinase family 127

4.2.1. NAGK as a paradigm of large-amplitude motions in the Amino Acid Kinase family 127

4.2.2. Oligomerization effecs on large-amplitude dynamics 143

4.3. Thermal stability of enzymes 160

4.3.1. Flexibility and diffusion observed by neutron scattering 160

4.3.2. Intramolecular dynamics of the thermo-mesophilic pair of enzymes ... 182

5. Conclusions 225

6. Sumario 229

CHAPTER 1

INTRODUCTION

1.1. Enzyme catalysis

Enzymes make up the biological machinery responsible for catalyzing the broad diversity of chemical reactions that fuel the biological processes of every living organism. Every enzymatic function is focused on accelerating any chemical reaction to time scales that are biologically relevant which fall into the micro-milliseconds time range [1]. Since most of non-catalyzed biochemical reactions would otherwise take place in time scales ranging from minutes to millions of years, enzymes are outstandingly efficient catalysts achieving rate-enhancements, with respect to aqueous solution, that can be superior to 15 orders of magnitude under mild conditions [2,3] (see Figure 1). Indeed, enzymes represent the most refined expression of chemical pre-organization. This allows to optimally orient a variety of amino acid functional groups in the active site cavity for efficient substrate binding and stabilization of the reaction transition state. Understanding the origin of such efficient catalysis thus has a tremendous potential for elucidating factors responsible for diseases as well as finding inhibitors of non-desired enzymatic functions. Furthermore, the extraordinary efficiency, versatility and enantioselective control of enzymes finds wide use in a variety of biotechnological applications, which redesign natural enzymes by mutagenesis in order to catalyze non-naturally occurring chemical reactions of interest.

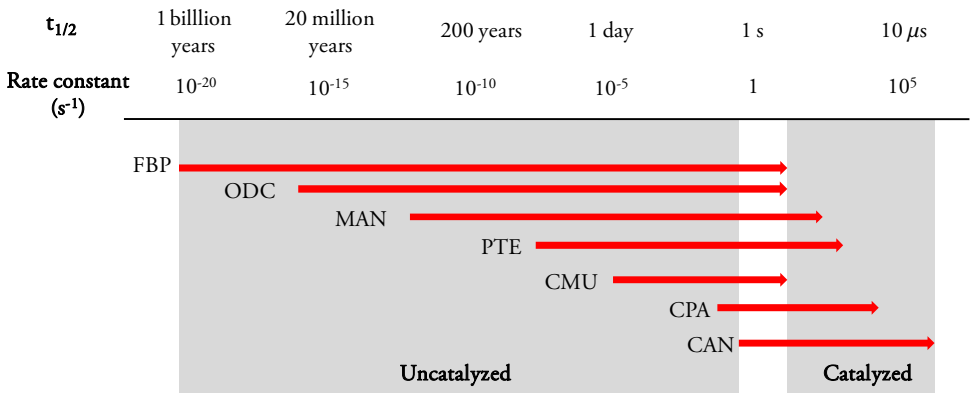


Figure 1. Representation of the rate-enhancement of some enzyme-catalyzed reactions (FBP, fructose-bisphosphatase ; ODC, orotidine 5'-phosphate decarboxylase ; MAN, mandelate racemase ; PTE , phosphotriesterase ; CMU, chorismate mutase ; CPA, human cyclophylin A ; CAN, carbonic anhydrase). Data extracted from references [1-3].

Specificity is the hallmark of enzymes. The original Lock and Key model by Fischer, invoked to provide an explanation for enzyme specificity, regarded the enzyme as a rigid molecule that is complementary in shape to its substrate. With the advent of X-ray crystallography, the fascinating pictures of the precise pre-organization of the enzyme active site rapidly spread the view of enzymes as static entities with a well defined native structure that entirely determines the substrate binding mode and, ultimately, the catalytic efficiency. However, the increasing number of X-ray structures for unbound (apo) and bound (holo) states of enzymes rapidly changed this paradigm. The experimental evidence showed that enzymes, indeed, are easily deformable structures that require changes in conformation to bind substrates in the optimal position for efficient catalysis. On the one hand, the induced fit model by Koshland [4] considers some degree of plasticity in the enzyme and proposes that the enzyme deformations observed upon ligand binding are "induced" by the ligand to optimize the network of interactions between the enzyme and ligand. In the conformational selection model [5], on the other, the ligand selects a conformation from a pre-existing equilibrium of conformations accessible by the enzyme. The latter model takes one step further toward the dynamic view emphasizing that enzymes are *intrinsically* flexible. It is believed that both the induced fit and the conformational selection model contribute to describe the origin of functional conformational changes. Nowadays the dynamic nature of enzymes is commonly accepted and turns out to be the subject of study by many experimental and computational groups aimed at characterizing the vast range of dynamic events involved in the enzymatic function.

The dynamic nature of enzymes is described by the concept of multidimensional free energy surface as originally introduced by Frauenfelder and co-workers [6]. This rough energy landscape is characterized by different conformational states, which in turn comprise many conformational substates. The transitions between different conformational states are rare and thus occur at slow time scales (beyond the microsecond time scale), whereas faster fluctuations between substates within the same conformational state take place in the range from picoseconds to nanoseconds. Slow dynamics correspond to large-amplitude and collective conformational changes which are typically associated to substrate binding and allosteric events, e.g. domain motions or relative motions between subunits in oligomeric structures. Faster time scales, on the other hand, involve more local fluctuations between closely similar structures such as sidechain rotations as well as subdomain and loop motions (see reference [7] for review).

1.1.1. Experimental techniques to study protein dynamics

Nowadays protein dynamics can be studied by a wide range of complementary experimental techniques suitable for different time scales and resolutions (Figure 2).

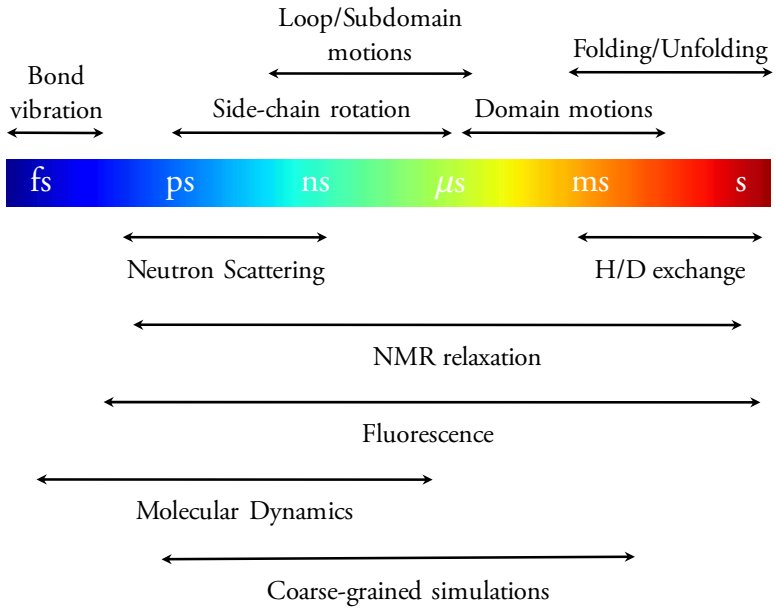


Figure 2. Time scales of dynamical events in proteins and techniques sensitive to different time scales.

In NMR, the relaxation of nuclei, after excitation with magnetic fields, detect conformational transitions, at atomic resolution, spanning time scales from picoseconds to seconds. Appealing NMR dispersion experiments on the cyclophilin A enzyme by Kern and co-workers [8] unveiled that large-amplitude motions can be rate-limiting. The observation of a global conformational change in the course of catalysis and the fact that these motions are detected in the free enzyme with frequencies similar to the reaction turnover indicated that the slow dynamics necessary for catalysis is both rate-limiting and an intrinsic property of enzymes. Put it more succinctly, what limits the catalytic activity of the enzyme is the ability to move.

Hydrogen-Deuterium (H/D) exchange [9] measure the number of hydrogen atoms, from amide groups in the protein backbone, being exchanged by deuterium atoms when the protein is dissolved in D_2O. Studying the time and temperature dependence of the number of exchange atoms provides insight into changes in solvent accessibility that are intimately related to conformational fluctuations occurring at slow time scales (beyond milliseconds).

Single-molecule fluorescence experiments illustrate very well the flexible nature of enzymes, as pioneered by Xie and co-workers in the cholesterol oxidase study [10]. This kind of experiments shows large fluctuations of the turnover rate constant with time, which gives rise to a distribution of catalytic rate constants for a single enzyme. Interestingly, such oscillations are observed to be correlated with fluctuations between long-lived conformations exhibiting different reactivity, i.e. on-off states. Therefore the large size of the activity fluctuations (the minimum and maximum catalytic rate can differ by an order of magnitude) highlights the marked influence of the enzyme conformation in determining the reactive properties (see references [11,12] for reviews).

Neutron scattering techniques provide dynamic information at faster time scales (from picoseconds to few nanoseconds) by measuring the mean-square-displacement, which quantifies the amplitude of motions accessible at a given time scale [13]. Even though the amount of dynamic information that one can extract with neutron scattering is more limited than with other techniques, such as NMR, it has found wide use in comparing the local flexibility of different proteins. For instance, different studies on thermo-mesophilic pairs of proteins showed that thermophilic enzymes are more flexible than their mesophilic homologues at the fast time scales probed by neutron scattering [14,15]. This completely changed the traditional paradigm that thermophilic proteins are more rigid than their mesophilic homologues and amazingly suggested that higher flexibility at fast time scales does not necessarily imply higher catalytic activity. In section 1.4 this will be extended.

1.1.2. Computational techniques to study protein dynamics

Protein dynamics can be explored by all-atom simulations such as Molecular Dynamics (MD) that describe the conformational fluctuations of the system at time scales ranging from picoseconds to hundreds of nanoseconds. MD allows to follow the atomic positions with time, identify the most relevant conformations as well as characterize conformational transitions. In general the description of dynamical events in the microseconds time scale or beyond is inaccessible by conventional MD with current computational power. In this regard, impressive progress by Shaw and co-workers is being done to cover extremely longer time scales by using a special-purpose machine for MD. They recently reported the first 1-millisecond simulation for the bovine pancreatic trypsin inhibitor [16]. Despite the increasing computing power, a wealth of more approximate methods are ongoingly developed to cover longer time scales, e.g. coarse-graining. This will be covered in the *Methods* section.

1.1.3. Controversy on the coupling between enzyme dynamics and catalysis

Despite the large amount of evidence from experimental and computational studies with regard to the active role of enzyme motions in catalysis, nowadays there is an intense debate in the biochemical and biophysical communities about the actual contribution of dynamical effects in catalysis. The core of the controversy lies in what is understood by "dynamical effects". For many authors, such as Kern and co-workers, "dynamical effects" in catalysis merely refer to "any time-dependent change in atomic coordinates" [7]. This definition includes the concept of "coupled promoting motions" disseminated by Hammes-Schiffer and Benkovic [17] which refer to a network of coupled protein motions that occur in the progress along the chemical reaction coordinate. However, Warshel and co-workers argue that the observation of correlated protein motions in the course of the reaction is trivial [18], since any chemical reaction in solution is also subjected to coupled motions of the environment and that what fully accounts for the enzyme catalytic effect is the electrostatic pre-organization of the active site. What Warshel and co-workers understand by "dynamical effects" is a transfer of energy from a conformational coordinate to the chemical reaction coordinate in an inertial way [18]. Several studies quantifying this contribution to lowering the barrier height conclude that the dynamical effect is minimal [19,20], whereas others found that this is an important contribution [21]. Schwartz and co-workers coined the term "rate-promotion vibrations" [21] to describe these protein motions that transfer energy to the reaction coordinate and find, instead, that protein compression modes can significantly reduce the energy barrier [22].

On balance, part of the origin of this controversy lies in semantic issues and thus an exact and uniquely used definition of dynamical effects in enzyme catalysis is urgently needed.

This will provide a consensus view approaching both experimental and computational communities in a concerted effort to precisely determine the actual role of protein motions in determining the outstanding efficiency of enzymes.

1.2. Phosphorylation reactions by enzymes

The phosphorylation of proteins is fundamental in many processes of the cell including division, differentiation and development as well as metabolic pathways. This is borne out by the observation that one third of all proteins in the cell are phosphorylated [23]. Besides proteins, phosphate groups are present in a wide diversity of biomolecules, ranging from nucleic acids, enzyme cofactors, metabolites and ATP [24]. The latter, in particular, is frequently used by enzymes for phosphorylating other proteins. These enzymes are termed *kinases*, whereas those removing the phosphate group are called *phosphatases*. ATP is regarded as an abundant storage of chemical energy due to the high exothermicity of its hydrolysis reaction that is in turn exploited for assisting endothermic reactions. This chemical energy can also be transformed into mechanical energy for muscular contraction or into osmotic energy for transporting ions against an electric potential gradient across the membrane. Given the central role of phosphates, it is not surprising that a broad range of malfunctions in the cell stem from the anomalous behavior of phosphorylating enzymes. For instance, alterations in signaling kinases involved in cellular division processes promote the formation of cancerous cells [23].

From the chemical point of view, studying the mechanisms of enzymatic phosphorylation reactions is particularly necessary for unraveling how this kind of enzymes achieve their tremendous catalytic power. For instance, fructose-bisphophatase and inositol-phosphatase are estimated to enhance the reaction rate 10^{21} fold [3], which is the highest rate-enhancement reported so far. The key of such outstanding rate-enhancement lies in the ability to perform reactions that are extremely slow, due to the extraordinary kinetic stability of phosphates in solution, at time scales suited for biological processes. Anions of phosphoesters are very stable as they repel nucleophiles, being ~$1.1 \cdot 10^{12}$ years, for instance, the half-life time for attack of water on alkyl phosphate dianions [3]. Given the observed dominance of phosphorylating enzymes in biological processes and their salient catalytic power, a very extensive literature exists on mechanistic investigations of enzymatic phosphoryl transfers. These investigations are carried out in combination with mechanistic studies on uncatalyzed phosphoester hydrolysis in water to give insight into strategies for enzyme catalysis (see references [25-27] for reviews).

1.2.1. Mechanisms of phosphorylation

The reaction mechanisms of phosphoryl transfer are more complex than those of nucleophilic substitutions on carbon atoms as a consequence of the hypervalent character of phosphorus. A nucleophilic substitution on phosphorus has three different mechanisms: dissociative, concerted and associative [28] (Figure 3).

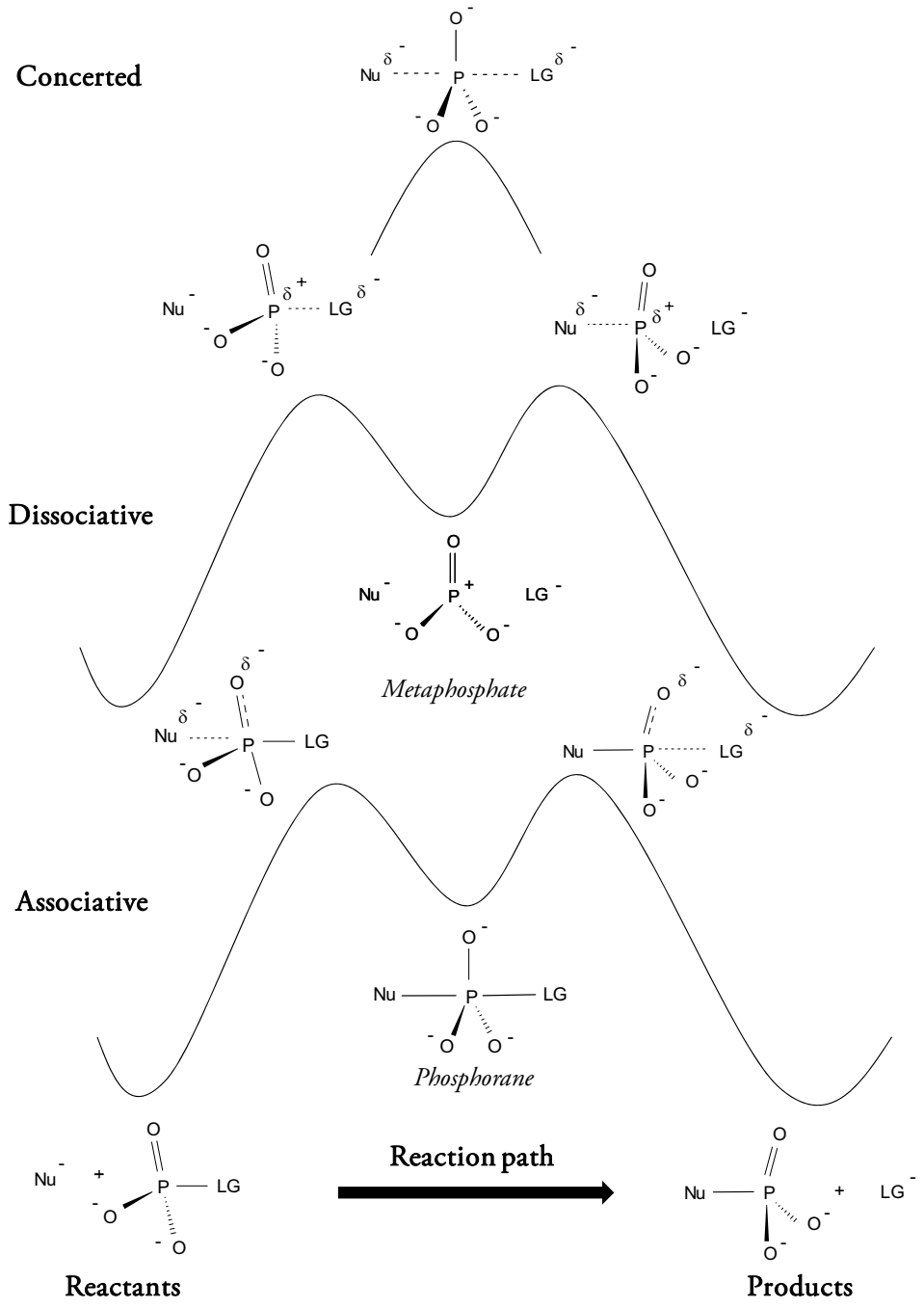


Figure 3. Phosphoryl transfer mechanisms: concerted, dissociative and associative.

Dissociative mechanism

As a first step, the bond between P and the leaving group is broken leading to the formation of a *metaphosphate* intermediate, which in a second step forms a bond with the nucleophile. This intermediate species is an anion with a planar trigonal structure (Figure 4A). It is unstable in solution and has only been observed in a crystallographic structure of fructose-1,6-bisphosphatase [29].

Concerted mechanism

The bond-breaking of the leaving group is carried out in concert with the bond-formation of the nucleophile giving rise to a pentacoordinated phosphorus species that is the transition state of the reaction. Depending on how synchronous is the formation and breaking of the two bonds the transition state is defined as *loose*, when it has more dissociative character, or *tight*, which is more associative [30]. It is widely believed that most of enzymatic phosphorylations proceed in a concerted fashion [27].

Associative mechanism

As opposed to the dissociative mechanism, the bond-formation of the nucleophile occurs prior to the bond-breaking of the leaving group leading to the formation of a pentacoordinated phosphorus intermediate, also known as *phosphorane* (Figure 4B). What provides stability to this species is the hypervalency of P and for this reason a wide variety of synthetic phosphorane compounds have been obtained [31].

This pentacoordinated species can be stable enough to undergo a pseudorotation process, in which both axial substituents are exchanged by two equatorial ones, while the third equatorial position remains at its position as a pivot [31] (Figure 4C). This pseudorotation process is of great importance with regard to reactions involving chiral phosphates. In general nucleophilic substitutions lead to an inversion of the configuration of the chiral center, but a retention of configuration can occur in case of pseudorotation.

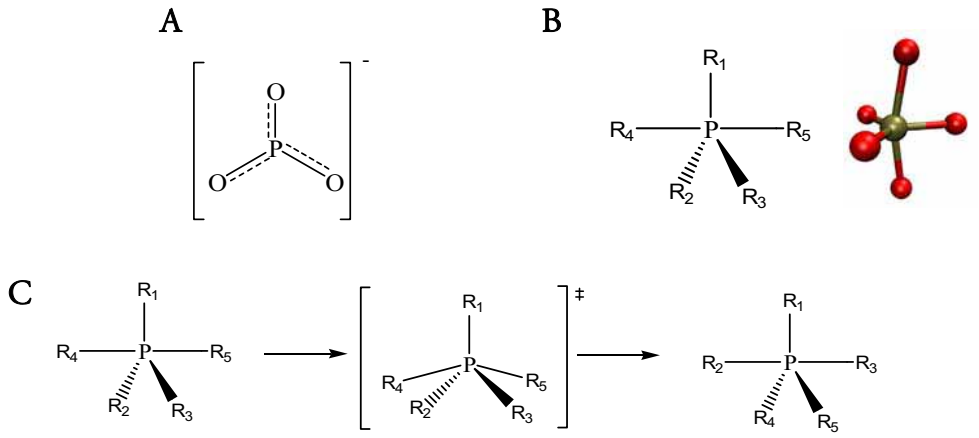


Figure 4. Intermediate species in phosphoryl transfer reactions. (A) Metaphosphate is anion with a planar trigonal structure with three P-O bonds (B) 2D- (l*eft*) and 3D- (*right*) representations of a phosphorane. The structure is a trigonal bipyramid centered at the P atom, where the two substituents at axial positions form an angle close to 180º with bond distances longer than those corresponding to the other three equatorial bonds at the central plane of the molecule. (C) Pseudorotation process of a phosphorane.

1.2.2. Hypervalency in pentacoordinated phosphorus

The observation of pentacoordinated phosphorus compounds, and other hypervalent molecules made by sulfur, silicon or halogen atoms, motivated a modification of Lewis' rules, which establish that each element fulfills the octet by forming chemical bonds between atom pairs that share two electrons. Historically two possible modifications were suggested [32]. Valence Bond Theory first proposed an expansion of the valence shell by allowing the promotion of electrons from occupied orbitals to empty *d* orbitals increasing the number of electrons participating in chemical bonds. In this framework, the trigonal bipyramid structure of pentacoordinated phosphorus compounds, like the paradigmatic PF_5, is explained by a sp^3d hybridization of the phosphorus atom. The linear combination of the 3s orbital, three 3p orbitals and one 3d orbital (with an electron promoted from the 3s orbital) yields five sp^3d orbitals pointing to the vertices of a trigonal bipyramid. An alternative to this hybridization scheme is the formation of chemical bonds with high ionic character or fractional contribution of electrons not violating the octet rule. With the advent of accurately computed wave functions along with wave function analysis methods, such as the Natural Population Analysis by Weinhold and co-workers [33], the hybridization scheme was shown to be incorrect as it implies a participation of *d* orbitals in the chemical bonds that is too high. Instead, chemical bonds in these molecules are

shown to be delocalized and distinguished by a high ionic character, which implies a bond order lower than one. These kinds of bonds thus turn out to be weaker than covalent bonds. The 3-center 4-electrons bond scheme (3c4e) is in line with this description. A central atom shares a bonding orbital with two atoms that provide an electron pair each one (Figure 5). This model provides an explanation for the higher coordination number of the central atom avoiding the necessity to expand the octet. It is worth pointing out that the hybridization scheme, which is used in many chemistry text books, is obsolete and do not represent properly the interactions responsible for hypervalency [32]. *In this thesis we have studied factors that influence the stability of pentacoordinated phosphorus compounds.*

Figure 5. Lewis structure of a 3c-4e bond model, which is represented by two resonance structures.

1.2.3. Pentacoordinated phosphorus in enzymes: controversy on β-phosphoglucomutase

A phosphorane formed in the course of an enzymatic reaction is a high-energy intermediate that, if isolated, could provide a revealing picture of the enzyme in a high-energy state in which the interactions between the substrate and the enzyme are the key factors of catalysis. For this reason, it was of great interest the report in *Science* by Allen and co-workers [34] of the first crystallographic structure of a pentacoordinated phosphorus intermediate in the active site of an enzyme, *β*-phosphoglucomutase (*β*-PGM), which catalyzes the isomerization of *β*-glucose-1-phosphate (G1P) to *β*-glucose-6-phosphate (G6P). This structure aroused much controversy from the very beginning [35,36], since some authors argued that the phosphorane might have been wrongly identified and that, instead, a MgF_3^- salt formed under crystallization conditions was likely to mimic the phosphoryl moiety of the phosphorane. In support of this idea, theoretical studies by Webster [37] showed that a pentacoordinated phosphorus in the active site of this enzyme could not be a stable species, but a transition state. Subsequently Allen and co-workers defended their thesis by showing with quantitative analytical methods that the enzyme does contain a phosphorus species in the active site [38]. Following the idea of the wrong assignment, NMR and kinetic assays by Waltho and co-workers [39,40] convincingly demonstrated that the MgF_3^- salt is indeed a potent inhibitor of *β*-PGM that acts as a transition state analogue in similar conditions to those of the original experiment. These were clear indications that the formation of the

phosphorane intermediate is unlikely in this enzyme, but it still lacked a detailed understanding of the phosphorylation mechanism. *In this thesis we have studied the reaction pathway of this phosphoryl transfer.*

1.3. Amino Acid Kinase family: large-amplitude motions mediating function

The Amino Acid Kinase family of enzymes (AAK) comprises a series of enzymes that catalyze a phosphorylation reaction and have important similarities in terms of sequence and structure. The family members are: N-acetyl-L-glutamate kinase (NAGK), Carbamate kinase (CK), Glutamate-5-kinase (G5K), UMP kinase (UMPK), Aspartokinase (AK) and the fosfomycin resistance kinase (FomA). Rubio and co-workers [41] have exhaustively studied this family and proposed that the shared fold among family members is likely to give rise to a similar mechanism of substrate binding and catalysis. These enzymes share the same α/β fold of the monomeric subunit and, among them, NAGK is regarded as the structural paradigm of the AAK family, since it is the most widely studied family member. Studies on the catalytic mechanism and dynamics of NAGK thus can shed light on how the rest of family members perform their function [41].

1.3.1. N-Acetyl-Glutamate Kinase

NAGK uses ATP to catalyze the phosphorylation of the amino acid N-Acetyl-L-Glutamate (NAG) in the biosynthesis of arginine from glutamate in microorganisms and plants (see Figure 6). NAG is one of the N-acetylated intermediates that are produced in this anabolic route and its phosphorylation by NAGK turns out to be the key regulatory step of the route in many organisms, since NAGK is feedback inhibited by the end product arginine. What makes this route interesting from the medical point of view is the fact that in mammalian cells the arginine biosynthesis proceeds through non-acetylated intermediates. Therefore NAGK activity may be selectively inhibited and, given the regulatory role of this enzyme in bacteria, it can be a potential target for antibacterial drugs. From the chemical point of view, the reaction catalyzed by NAGK is relatively uncommon as the phosphoryl group is transferred from ATP to a carboxylate group of N-acetyl-glutamate, whereas most of kinases phosphorylate alcohol groups from protein residues, e.g. serine or tyrosine, and metabolites.

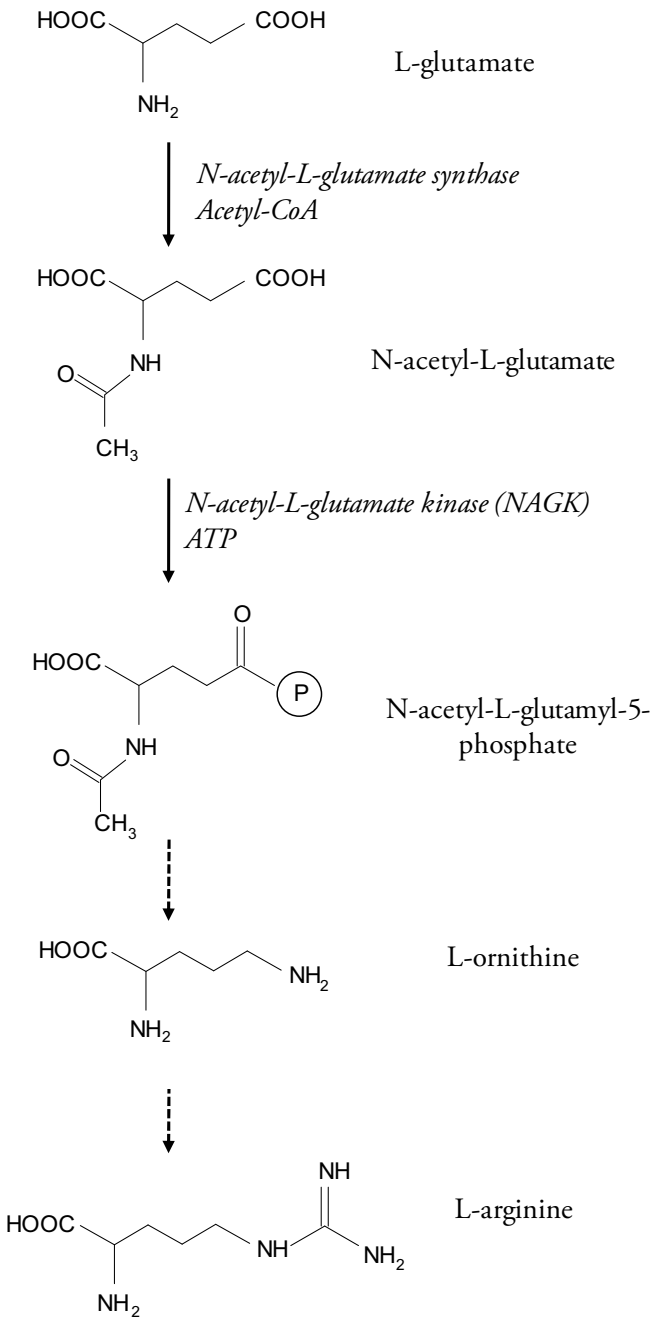


Figure 6. Bacterial route of arginine biosynthesis from glutamate. Dashed arrows denote more than one chemical step.

The regulation of NAGK activity by arginine inhibition has been thoroughly studied in hexameric NAGKs from *Thermotoga Maritima* and *Pseudomonas Aeruginosa* [42,43]. The inhibition occurs allosterically, which means that arginine binding in one of the subunits enhance the affinity of other subunits for binding a subsequent arginine molecule. The transition from the unbound state to the inhibited state with six bound arginine molecules involves a large conformational change that modifies the active site cavity as well as the size of the central ring of the hexameric structure. Apart from arginine inhibition, NAGK activity has been observed to be positively regulated in photosynthetic organisms. This enhancement of NAGK activity occurs by forming a complex with the signaling protein PII, which stabilizes the more active conformation and therefore competes with arginine to bind NAGK [44]. The formation of this complex is promoted in conditions of nitrogen abundance so as to enhance the production of arginine for subsequent storage.

NAGK from *Escherichia Coli* (*Ec*NAGK), on the other hand, is an example of an arginine-insensitive NAGK and turns out to be the best characterized enzyme among all NAGKs and Amino Acid Kinase family members. Its mechanism of phosphoryl transfer has been subjected to a wide range of crystallographic [41,45,46] and site-directed mutagenesis studies [47]. It is a homodimer of 258 residues in each monomeric subunit, which consists of a N domain that hosts the NAG binding site (NAG lid) and a C domain that binds ATP. It is at the interface between these two domains that the phosphoryl transfer reaction takes place (see Figure 7). This dimeric arrangement provides only thermodynamic stability to the monomeric fold that has been evolutionary selected to perform the catalytic function. This is borne out by kinetic studies that have not shown evidence of cooperativity between both subunits [47].

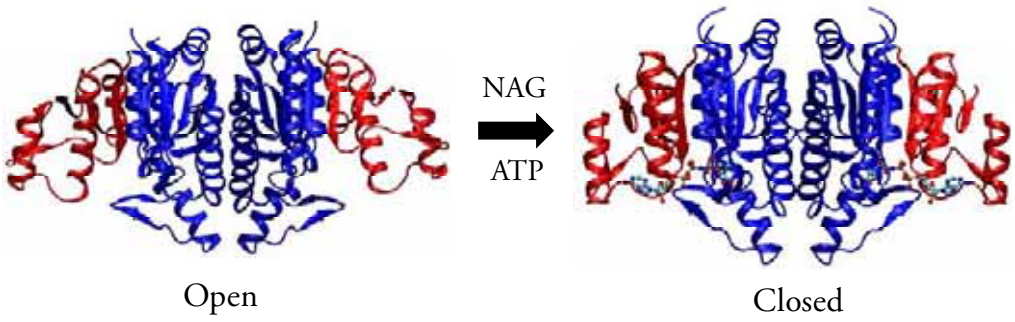


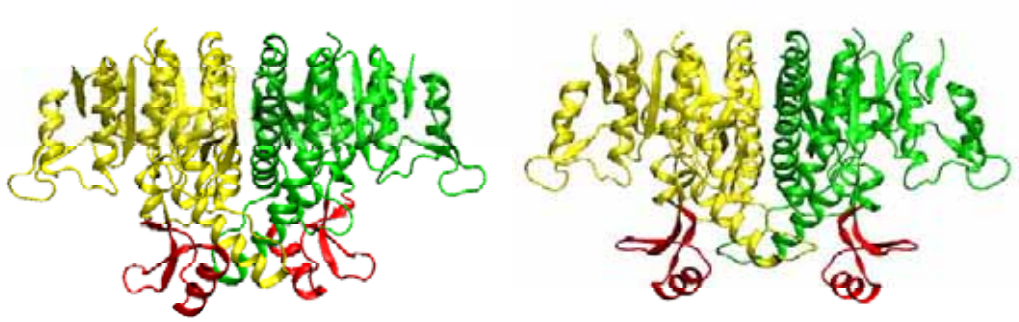
Figure 7. Structure of the open (unbound; PDB code 2WXB) and closed (with the ATP analogue AMPPNP and the amino acid NAG; PDB code 1GS5) structures of the *Ec*NAGK dimer. The substrates ATP and NAG (ball and sticks) bind to the C (red) and N domains (blue) respectively.

The first crystallographic structures of *Ec*NAGK were solved for the bound state of the enzyme (PDB codes 1GS5, 1OH9, 1OHA and 1OHB) and revealed an active site that was too narrow to allow the substrates bind directly without a conformational rearrangement. In the X-ray structure with code 1GS5, the ATP analogue AMPPNP and NAG were optimally positioned in the active site to perform the phosphoryl transfer. The reactive atoms are so closely positioned that continuous electron density was observed between both atoms. A subsequent structure with a transition state analogue (AlF_4^-) provided insight into the actual interactions in the putative transition state structure. To understand how the substrates achieve such optimal position in the closed active site it was hypothesized that a more open conformation would be more accessible by the enzyme in the absence of substrates. Several years after determining the first X-ray structure, a more open conformation of the enzyme in the apo state was crystallized (PDB code 2WXB) [46]. This confirmed the former hypothesis and shed light into a large-amplitude conformational change enabling substrate binding. *In this thesis we have studied the intrinsic dynamics of EcNAGK, regarding large-amplitude motions linked to substrate binding. We have also explored the extent of conservation of these intrinsic dynamical features among AAK family members.*

1.3.2. Carbamate Kinase

Carbamate kinase (CK) catalyzes the formation of carbamate by using ATP and carbamoyl phosphate (CP). CP plays an active role in the biosynthesis of pyrimidines, arginine and urea. CK in other organisms makes the reverse reaction, which involves the synthesis of ATP and carbamate from CP. CKs are homodimers that share the same α/β fold of the AAK family but adds a distinctive structural feature, which is a protruding

subdomain close to the intersubunit surface. The X-ray structures of different bound states of the enzyme in different organisms reveal that this subdomain undergoes important conformational changes that open and close the CP site [48-50] (Figure 8). A recent hypothesis states that the flexibility of this subdomain might play a role in enhancing the specificity of this enzyme for CP with respect to more abundant analogues, such as bicarbonate or acetate. *In this thesis we have analyzed the flexibility of the protruding subdomain.*



Open Closed

Figure 8. Structures of Carbamate Kinase from Pyrococcus Furiosus (left) and Enterococcus Faecalis (right). The protruding subdomain of each subunit is colored in red. The two structures show different conformations of this subdomain (open and closed).

1.3.3. Uridine Monophosphate Kinase

UMPK catalyzes the phosphorylation of UMP using ATP to yield ADP and UDP. It mediates one of the multiple steps of the biosynthesis of UTP, which is a precursor of RNA, DNA and phospholipids. Bacterial UMPK is a hexamer regarded as a trimer of dimers in which each subunit has the typical fold of the AAK family [51]. The interesting structural feature of UMPK is the assembly of the subunits in each dimer. In contrast to NAGK or CK, where helices that build hydrophobic contacts between the two subunits are crossed (in an angle of ~65º), in UMPK the analogue helices are parallel (Figure 9A). This leads to a hexameric assembly that is significantly different to that observed in hexameric NAGKs.

The activity of hexameric UMPKs is regulated by the nucleoside triphosphates GTP and UTP. GTP is an allosteric activator, whereas UTP, the end product of the biosynthetic route, is an inhibitor. This mechanism of activity regulation is known to balance the

synthesis of purine and pyrimidine bases. The X-ray structures determined for the UTP [52] and GTP [53] bound states of UMPK from *Eschericia Coli* reveal a large conformational change, in which GTP triggers the opening of each dimer (Figure 9B). *In this thesis we have studied how this conformational change is related to the different assembly of this enzyme.*

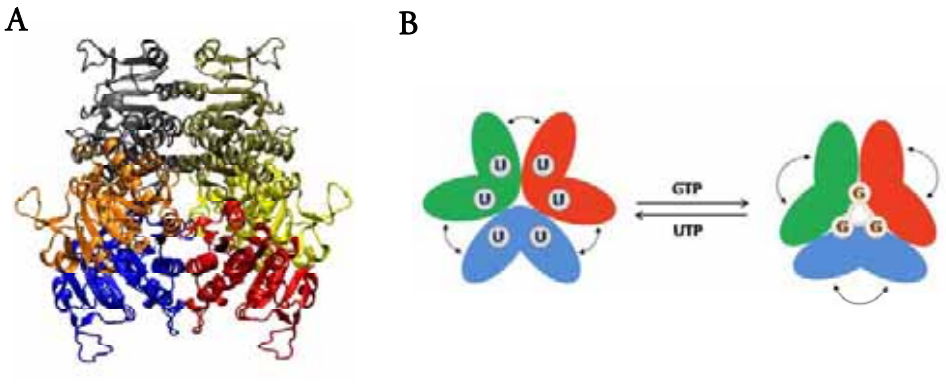


Figure 9. UMPK and its allosteric regulation. (A) Ribbon diagram of *Ec*UMPK structure (B) Schematic representation of the conformational transition between UTP and GTP bound states of UMPK. Each area with the same color refers to one of the dimers. This figure has been reproduced from reference [53].

1.4. Thermostability in enzymes

Of particular importance is the fact that enzymes achieve their outstanding efficiency under very specific environmental conditions. Factors like temperature, pressure or salt concentration have a tremendous impact on enzyme activity, but the broad versatility of the enzyme machinery allows finding enzymes adapted to extremely diverse environments. Of particular interest are thermophilic and hyperthermophilic proteins from organisms that grow at very high temperatures ranging from 50 to 120ºC. Such adaptation to high temperatures requires very resistant proteins to heat denaturation and elucidating the origin of this resistance attracts the attention for designing proteins with enhanced thermal stability for a wide range of applications.

There has been a surge of experimental and computational studies comparing thermophilic proteins with their corresponding homologues working at room temperature, known as mesophilic. From studies in the last 20 years, it seems that there is not a unique strategy adopted by evolution to thermostabilize proteins. One of the main problems when interpreting the differences observed between thermo-mesophilic pairs of proteins is that not all these differences necessarily have to be related to differences in thermal adaptation and, perhaps, other properties derived from the different nature of the organisms may be involved.

Structural and amino acid sequence comparisons between a vast range of thermophilic and mesophilic proteins point to some features that correlate with increased thermostability. In the following, an overview of the several features that, in general, are required for protein thermostability is provided.

1.4.1. Sequence and structural requirements for protein thermostability

The observation in thermophilic proteins of a larger proportion of charged residues (arginine, lysine, aspartate, glutamate and histidine) at the expense of uncharged polar residues, with respect to mesophilic homologues, has strongly supported the primary role of electrostatic interactions in protein thermostability (see references [54,55] for reviews). Charged residues are important for thermal stability because they form salt bridge interactions between oppositely charged residues. These are very intense non-bonding interactions that are known to strengthen with temperature [56-58]. This is borne out by the fact that the desolvation penalty required to form a salt bridge decreases with temperature. At low temperature, salt bridge interactions are destabilizing because the interactions of solvent-exposed charged residues with water are more favorable than within a salt bridge. In other words, the cost of disrupting the interactions between charged residues and water molecules (desolvation penalty) is not balanced by the

formation of an electrostatically favorable salt bridge. Water molecules are less efficient screening electrostatic interactions at high temperature, e.g. water dielectric constant decreases from 86 to 70 when increasing the temperature from 0ºC to 50ºC. This implies that upon raising the temperature the desolvation penalty decreases to the extent that charged residues interacting through a salt bridge become more stable than being highly exposed to the solvent. For this reason, salt bridge interactions are considered to be *uniquely suited* for protein thermostability [59].

Salt bridge interactions are observed to participate in networks of electrostatic interactions making cross-links across the whole protein that stabilize the tertiary structure. Computational studies also indicate that these networks are further optimized in thermophilic proteins [60]. Obviously, in a network of electrostatic interactions, not only the attractive interactions between pair-wise salt bridges are present, but also repulsive interactions as a result of nearby residues with the same charge. Interestingly, what makes the networks of thermophilic proteins more stable is the fact that the magnitude of repulsive interactions is minimized. The importance of such optimization in thermophilic proteins was impressively demonstrated by site-directed mutagenesis studies [61] showing that a single mutation Arg → Glu at the surface of a thermophilic cold-shock protein destabilizes electrostatic interactions to the extent that the unfolding temperature dramatically drops in ~20ºC. Overall, both the number of charged residues and their spatial distribution within the protein are determinants of the impact of electrostatic interactions on thermal stability. Based on this idea, computational protein design methods [62] have been successful in thermostabilizing different enzymes by optimizing charge-charge interactions at the surface.

A complementary view of the stabilizing effect of charged residues is based on the fact that the higher content of charged residues at the surface of thermophilic proteins provides the surface with enhanced hydrophilic character. Findings from simulations [63,64] indicate that such increased affinity for water gives rise to a more dense water shell surrounding the protein that protects the protein core against water penetration, which ultimately drives protein unfolding.

Related to the increased number of charged residues at the surface of thermophilic proteins is the fact that hydrophobic residues are more densely packed in the protein core with respect to mesophilic homologues. From the structural point of view, thermophilic proteins also exhibit shortened loops at the surface. These two features are consistent with the idea that thermophilic proteins are able to maintain the robustness of the structure to a higher extent.

1.4.2. Dynamical requirements for protein thermostability

The dynamical requirements for protein thermostability are more controversial. Since thermophilic proteins unfold at higher temperature and are less active than their mesophilic homologues at lower temperature, thermophilic proteins have been traditionally considered more rigid. According to the *corresponding states* hypothesis, thermophilic and mesophilic proteins achieve similar flexibility at their respective temperatures for maximum activity.

Nevertheless, experimental and simulation techniques able to explore atomic motions at different time scales have indicated that the panorama is more complex and that there is not a unique strategy for protein thermostability. Some of these studies found thermophilic proteins to be more rigid than their mesophilic homologues [65-69], whereas others showed the opposite [14,15,70-73]. This lack of consensus stems from the absence of a unique mechanism of thermostability and, on the other hand, from the fact that these techniques explore different aspects of protein dynamics among the vast diversity of dynamic events that occur in a broad range of time scales. For instance, in the case of enhanced flexibility in thermophilic proteins, as shown by neutron scattering at fast time scales [14,15], the larger structural fluctuations can entail an increase in conformational entropy of the native state that provides more stability [74]. On the other hand, NMR relaxation experiments [69] probing slow time scales show that large-amplitude motions linked to substrate binding and catalysis in a thermophilic adenylate kinase do not occur as frequently as in a mesophilic homologue at low temperature, which supports the corresponding states hypothesis. This is a clear indication that a proper definition of flexibility requires the specification of time scale and type of motion. For this reason, the flexibility regarding different dynamic events is not directly comparable and, ultimately, their linkage to stability and function do not have to be necessarily the same. Moreover, it is worth recalling that differences in dynamics between thermophilic and mesophilic homologues are not completely related to thermostability, since adaptation to other required properties are likely to alter the dynamics.

One of the most intriguing dynamical mechanisms of protein thermostability was proposed few years ago by Zaccai and co-workers based on neutron scattering experiments [15]. They observed that the mean-square-displacement at short time scales of a thermophilic enzyme was less sensitive to temperature changes than the corresponding mesophilic homologue. These results motivated the authors to suggest that this can be a plausible mechanism for thermophilic proteins to control the structural fluctuations at high temperature to avoid unfolding. *In this thesis we have explored the dynamical basis of these results.*

1.4.3. Exchange between thermostability and catalytic activity

When a thermophilic enzyme exhibits more restricted motions than a mesophilic homologue, a dynamical reason is usually considered to be at the origin of the lower activity at room temperature. However, computational studies by Warshel and co-workers [75] suggest, instead, that a thermophilic enzyme is less active at room temperature because the active site is less pre-organized than in a mesophilic homologue. This implies that the energy barrier for the chemical step is higher requiring more elevated temperatures to be surmounted. They argue that a mesophilic enzyme invest more folding energy in a highly pre-organized active site at the expense of lower stability. The reason for such tradeoff between stability and activity is that the formation of an active site cavity is thermodynamically unfavorable. Thus it is clear that the study of the relationship between enzyme thermostability and activity offers a convenient way to shed light into the aforementioned controversy on the role of enzyme dynamics in catalysis.

1.5. References

1. Wolfenden R, Snider MJ (2001) The depth of chemical time and the power of enzymes as catalysts. *Acc Chem Res* 34: 938-945.
2. Radzicka A, Wolfenden R (1995) A proficient enzyme. *Science* 267: 90-93.
3. Lad C, Williams NH, Wolfenden R (2003) The rate of hydrolysis of phosphomonoester dianions and the exceptional catalytic proficiencies of protein and inositol phosphatases. *Proc Natl Acad Sci U S A* 100: 5607-5610.
4. Koshland DE (1958) Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci U S A* 44: 98-104.
5. Monod J, Wyman J, Changeux JP (1965) On nature of allosteric transitions - A plausible model. *J Mol Biol* 12: 88-118.
6. Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. *Science* 254: 1598-1603.
7. Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. *Nature* 450: 964-972.
8. Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE, Kern D (2005) Intrinsic dynamics of an enzyme underlies catalysis. *Nature* 438: 117-121.
9. Katta V, Chait BT (1993) Hydrogen-deuterium exchange electrospray-ionization mass-spectrometry - a method for probing protein conformational-changes in solution. *J Am Chem Soc* 115: 6317-6321.
10. Lu HP, Xun LY, Xie XS (1998) Single-molecule enzymatic dynamics. *Science* 282: 1877-1882.
11. Min W, English BP, Luo GB, Cherayil BJ, Kou SC, Xie XS (2005) Fluctuating enzymes: Lessons from single-molecule studies. *Acc Chem Res* 38: 923-931.
12. Smiley RD, Hammes GG (2006) Single molecule studies of enzyme mechanisms. *Chem Rev* 106: 3080-3094.
13. Gabel F, Bicout D, Lehnert U, Tehei M, Weik M, Zaccai G (2002) Protein dynamics studied by neutron scattering. *Q Rev Biophys* 35: 327-367.
14. Fitter J, Heberle J (2000) Structural equilibrium fluctuations in mesophilic and thermophilic alpha-amylase. *Biophys J* 79: 1629-1636.
15. Tehei M, Madern D, Franzetti B, Zaccai G (2005) Neutron scattering reveals the dynamic basis of protein adaptation to extreme temperature. *J Biol Chem* 280: 40974-40979.
16. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* 330: 341-346.

17. Agarwal PK, Billeter SR, Rajagopalan PTR, Benkovic SJ, Hammes-Schiffer S (2002) Network of coupled promoting motions in enzyme catalysis. *Proc Natl Acad Sci U S A* 99: 2794-2799.
18. Kamerlin SCL, Warshel A (2010) At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins: Struct, Funct, Bioinf* 78: 1339-1375.
19. Roca M, Oliva M, Castillo R, Moliner V, Tunon I (2010) Do Dynamic Effects Play a Significant Role in Enzymatic Catalysis? A Theoretical Analysis of Formate Dehydrogenase. *Chem Eur J* 16: 11399-11411.
20. Pisliakov AV, Cao J, Kamerlin SCL, Warshel A (2009) Enzyme millisecond conformational dynamics do not catalyze the chemical step. *Proc Natl Acad Sci U S A* 106: 17359-17364.
21. Antoniou D, Schwartz SD (2001) Internal enzyme motions as a source of catalytic activity: Rate-promoting vibrations and hydrogen tunneling. *J Phys Chem B* 105: 5553-5558.
22. Quaytman SL, Schwartz SD (2007) Reaction coordinate of an enzymatic reaction revealed by transition path sampling. *Proc Natl Acad Sci U S A* 104: 12253-12258.
23. Cohen P (2001) The role of protein phosphorylation in human health and disease - Delivered on June 30th 2001 at the FEBS Meeting in Lisbon. *Eur J Biochem* 268: 5001-5010.
24. Westheimer FH (1987) Why nature chose phosphates. *Science* 235: 1173-1178.
25. Hengge AC (2005) Mechanistic studies on enzyme-catalyzed phosphoryl transfer. In: Richard JP, editor. Advances in Physical Organic Chemistry, Vol 40. pp. 49-108.
26. Cleland WW, Hengge AC (2006) Enzymatic mechanisms of phosphate and sulfate transfer. *Chem Rev* 106: 3252-3278.
27. Lassila JK, Zalatan JG, Herschlag D (2011) Biological Phosphoryl-Transfer Reactions: Understanding Mechanism and Catalysis. *Annu Rev Biochem* 80: 669-702.
28. Allen KN, Dunaway-Mariano D (2004) Phosphoryl group transfer: evolution of a catalytic scaffold. *Trends Biochem Sci* 29: 495-503.
29. Choe JY, Iancu CV, Fromm HJ, Honzatko RB (2003) Metaphosphate in the active site of fructose-1,6-bisphosphatase. *J Biol Chem* 278: 16015-16020.
30. Jencks WP (1972) General acid-base catalysis of complex reactions in water. *Chem Rev* 72: 705-718.
31. Swamy KCK, Kumar NS (2006) New features in pentacoordinate phosphorus chemistry. *Acc Chem Res* 39: 324-333.

32. Reed AE, Schleyer PV (1990) Chemical bonding in hypervalent molecules - the dominance of ionic bonding and negative hyperconjugation over d-orbital participation. *J Am Chem Soc* 112: 1434-1445.
33. Reed AE, Curtiss LA, Weinhold F (1988) Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chem Rev* 88: 899-926.
34. Lahiri SD, Zhang GF, Dunaway-Mariano D, Allen KN (2003) The pentacovalent phosphorus intermediate of a phosphoryl transfer reaction. *Science* 299: 2067-2071.
35. Blackburn GM, Williams NH, Gamblin SJ, Smerdon SJ (2003) Comment on "The pentacovalent phosphorus intermediate of a phosphoryl transfer reaction". *Science* 301: 1.
36. Allen KN, Dunaway-Mariano D (2003) Response to comment on "The pentacovalent phosphorus intermediate of a phosphoryl transfer reaction". *Science* 301: 1.
37. Webster CE (2004) High-energy intermediate or stable transition state analogue: Theoretical perspective of the active site and mechanism of beta-phosphoglucomutase. *J Am Chem Soc* 126: 6840-6841.
38. Tremblay LW, Zhang GF, Dai JY, Dunaway-Mariano D, Allen KN (2005) Chemical confirmation of a pentavalent phosphorane in complex with beta-phosphoglucomutase. *J Am Chem Soc* 127: 5298-5299.
39. Baxter NJ, Olguin LF, Golicnik M, Feng G, Hounslow AM, Bermel W, Blackburn GM, Hollfelder F, Waltho JP, Williams NH (2006) A Trojan horse transition state analogue generated by MgF3- formation in an enzyme active site. *Proc Natl Acad Sci U S A* 103: 14732-14737.
40. Golicnik M, Olguin LF, Feng GQ, Baxter NJ, Waltho JP, Williams NH, Hollfelder F (2009) Kinetic Analysis of beta-Phosphoglucomutase and Its Inhibition by Magnesium Fluoride. *J Am Chem Soc* 131: 1575-1588.
41. Ramon-Maiques S, Marina A, Gil-Ortiz F, Fita I, Rubio V (2002) Structure of acetylglutamate kinase, a key enzyme for arginine biosynthesis and a prototype for the amino acid kinase enzyme family, during catalysis. *Structure* 10: 329-342.
42. Ramon-Maiques S, Fernandez-Murga ML, Gil-Ortiz F, Vagin A, Fita I, Rubio V (2006) Structural bases of feed-back control of arginine biosynthesis, revealed by the structures of two hexameric N-acetylglutamate kinases, from Thermotoga maritima and Pseudomonas aeruginosa. *J Mol Biol* 356: 695-713.
43. Fernandez-Murga ML, Rubio V (2008) Basis of arginine sensitivity of microbial N-acetyl-L-glutamate kinases: Mutagenesis and protein engineering study with the Pseudomonas aeruginosa and Escherichia coli enzymes. *J Bacteriol* 190: 3018-3025.

44. Llacer JL, Contreras A, Forchhammer K, Marco-Marin C, Gil-Ortiz F, Maldonado R, Fita I, Rubio V (2007) The crystal structure of the complex of P-II and acetylglutamate kinase reveals how P-II controls the storage of nitrogen as arginine. *Proc Natl Acad Sci U S A* 104: 17644-17649.
45. Gil-Ortiz F, Ramon-Maiques S, Fita I, Rubio V (2003) The course of phosphorus in the reaction of N-acetyl-L-glutamate kinase, determined from the structures of crystalline complexes, including a complex with an AlF4- transition state mimic. *J Mol Biol* 331: 231-244.
46. Gil-Ortiz F, Ramon-Maiques S, Fernandez-Murga ML, Fita I, Rubio V (2010) Two Crystal Structures of Escherichia coli N-Acetyl-L-Glutamate Kinase Demonstrate the Cycling between Open and Closed Conformations. *J Mol Biol* 399: 476-490.
47. Marco-Marin C, Ramon-Maiques S, Tavarez S, Rubio V (2003) Site-directed mutagenesis of Escherichia coli acetylglutamate kinase and aspartokinase III probes the catalytic and substrate-binding mechanisms of these amino acid kinase family enzymes and allows three-dimensional modelling of aspartokinase. *J Mol Biol* 334: 459-476.
48. Marina A, Alzari PM, Bravo J, Uriarte M, Barcelona B, Fita I, Rubio V (1999) Carbamate kinase: New structural machinery for making carbamoyl phosphate, the common precursor of pyrimidines and arginine. *Protein Sci* 8: 934-940.
49. Ramon-Maiques S, Marina A, Uriarte M, Fita I, Rubio V (2000) The 1.5 angstrom resolution crystal structure of the carbamate kinase-like carbamoyl phosphate synthetase from the hyperthermophilic archaeon Pyrococcus furiosus, bound to ADP, confirms that this thermostable enzyme is a carbamate kinase, and provides insight into substrate binding and stability in carbamate kinases. *J Mol Biol* 299: 463-476.
50. Ramon-Maiques S, Marina A, Guinot A, Gil-Ortiz F, Uriarte M, Fita I, Rubio V (2010) Substrate Binding and Catalysis in Carbamate Kinase Ascertained by Crystallographic and Site-Directed Mutagenesis Studies: Movements and Significance of a Unique Globular Subdomain of This Key Enzyme for Fermentative ATP Production in Bacteria. *J Mol Biol* 397: 1261-1275.
51. Marco-Marin C, Gil-Ortiz F, Rubio V (2005) The crystal structure of Pyrococcus furiosus UMP kinase provides insight into catalysis and regulation in microbial pyrimidine nucleotide biosynthesis. *J Mol Biol* 352: 438-454.
52. Briozzo P, Evrin C, Meyer P, Assairi L, Joly N, Barzu O, Gilles AM (2005) Structure of Escherichia coli UMP kinase differs from that of other nucleoside monophosphate kinases and sheds new light on enzyme regulation. *J Biol Chem* 280: 25533-25540.

53. Meyer P, Evrin C, Briozzo P, Joly N, Barzu O, Gilles AM (2008) Structural and Functional Characterization of Escherichia coli UMP Kinase in Complex with Its Allosteric Regulator GTP. *J Biol Chem* 283: 36011-36018.
54. Kumar S, Nussinov R (2001) How do thermophilic proteins deal with heat? *Cell Mol Life Sci* 58: 1216-1233.
55. Karshikoff A, Ladenstein R (2001) Ion pairs and the thermotolerance of proteins from hyperthermophiles: a 'traffic rule' for hot roads. *Trends Biochem Sci* 26: 550-556.
56. de Bakker PIW, Hunenberger PH, McCammon JA (1999) Molecular dynamics simulations of the hyperthermophilic protein Sac7d from Sulfolobus acidocaldarius: Contribution of salt bridges to thermostability. *J Mol Biol* 285: 1811-1830.
57. Danciulescu C, Ladenstein R, Nilsson L (2007) Dynamic arrangement of ion pairs and individual contributions to the thermal stability of the cofactor-binding domain of glutamate dehydrogenase from Thermotoga maritima. *Biochemistry* 46: 8537-8549.
58. Vinther JM, Kristensen SM, Led JJ (2011) Enhanced Stability of a Protein with Increasing Temperature. *J Am Chem Soc* 133: 271-278.
59. Thomas AS, Elcock AH (2004) Molecular simulations suggest protein salt bridges are uniquely suited to life at high temperatures. *J Am Chem Soc* 126: 2208-2214.
60. Spassov VZ, Karshikoff AD, Ladenstein R (1994) Optimization of the electrostatic interactions in proteins of different functional and folding type. *Protein Sci* 3: 1556-1569.
61. Perl D, Mueller U, Heinemann U, Schmid FX (2000) Two exposed amino acid residues confer thermostability on a cold shock protein. *Nat Struct Biol* 7: 380-383.
62. Gribenko AV, Patel MM, Liu J, McCallum SA, Wang CY, Makhatadze GI (2009) Rational stabilization of enzymes by computational redesign of surface charge-charge interactions. *Proc Natl Acad Sci U S A* 106: 2601-2606.
63. Melchionna S, Sinibaldi R, Briganti G (2006) Explanation of the stability of thermophilic proteins based on unique micromorphology. *Biophys J* 90: 4204-4212.
64. Sterpone F, Bertonati C, Briganti G, Melchionna S (2009) Key Role of Proximal Water in Regulating Thermostable Proteins. *J Phys Chem B* 113: 131-137.
65. Wrba A, Schweiger A, Schultes V, Jaenicke R, Zavodszky P (1990) Extremely thermostable D-glyceraldehyde-3-phosphate dehydrogenase from the Eubacterium Thermotoga-Maritima. *Biochemistry* 29: 7584-7592.
66. Lazaridis T, Lee I, Karplus M (1997) Dynamics and unfolding pathways of a hyperthermophilic and a mesophilic rubredoxin. *Protein Sci* 6: 2589-2605.

67. Zavodszky P, Kardos J, Svingor A, Petsko GA (1998) Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc Natl Acad Sci U S A* 95: 7406-7411.
68. Butterwick JA, Loria JP, Astrof NS, Kroenke CD, Cole R, Rance M, Palmer AG (2004) Multiple time scale backbone dynamics of homologous thermophilic and mesophilic ribonuclease HI enzymes. *J Mol Biol* 339: 855-871.
69. Wolf-Watz M, Thai V, Henzler-Wildman K, Hadjipavlou G, Eisenmesser EZ, Kern D (2004) Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat Struct Mol Biol* 11: 945-949.
70. Hernandez G, Jenney FE, Adams MWW, LeMaster DM (2000) Millisecond time scale conformational flexibility in a hyperthermophile protein at ambient temperature. *Proc Natl Acad Sci U S A* 97: 3166-3170.
71. Grottesi A, Ceruso MA, Colosimo A, Di Nola A (2002) Molecular dynamics study of a hyperthermophilic and a mesophilic rubredoxin. *Proteins: Struct, Funct, Genet* 46: 287-294.
72. Wintrode PL, Zhang DQ, Vaidehi N, Arnold FH, Goddard WA (2003) Protein dynamics in a family of laboratory evolved thermophilic enzymes. *J Mol Biol* 327: 745-757.
73. Colombo G, Merz KM (1999) Stability and activity of mesophilic subtilisin E and its thermophilic homolog: Insights from molecular dynamics simulations. *J Am Chem Soc* 121: 6895-6903.
74. Stone MJ (2001) NMR relaxation studies of the role of conformational entropy in protein stability and ligand binding. *Acc Chem Res* 34: 379-388.
75. Roca M, Liu H, Messer B, Warshel A (2007) On the relationship between thermal stability and catalytic power of enzymes. *Biochemistry* 46: 15076-15088.

CHAPTER 2

OBJECTIVES

The main objective of the present thesis is to provide a global picture of different properties important for the enzymatic function: *reactivity*, *dynamics* and *thermostability*. By means of a variety of computational methods we have examined these properties in a set of enzymes.

1) Reactivity

The first part examines the *pentacoordinated phosphorus* species that can be either an intermediate or a transition state in the widely catalyzed phosphoryl transfer reactions in enzymes. Taking into account the difficulty in distinguishing the different types of phosphoryl transfer mechanisms, the main objectives of this part are:

- Find a methodology properly describing the interactions present in pentacoordinated phosphorus compounds for subsequent application in QM/MM studies.
- Characterize the electronic structure of pentacoordinated phosphorus compounds.
- Analyze systematically the inductive effects exerted by phosphorane substituents on the structure as well as polarization effects mediated by external electric fields.
- Evaluate the viability of the pentacoordinated phosphorus observed in the controversial structure of the β-phosphoglucomutase enzyme and calculate the reaction mechanism.

2) Dynamics

The second part is focused on the role of *large-amplitude motions* in the enzymatic activity of the Amino Acid Kinase (AAK) family. The objectives of this part are:

- Describe the most accessible modes of motion of the *Ec*NAGK enzyme
- Identify dynamical features common to AAK members.
- Study the effects of the oligomeric assembly on the dynamics associated to ligand binding processes.

3) Thermostability

The last part aims to gain insights into the dynamical properties associated to thermal stability as observed by neutron scattering experiments on a thermo-mesophilic pair of enzymes:

- Characterize the intramolecular dynamics of the two enzymes at different time scales.
- Approach the complex diffusional behavior of these enzymes in the crowded solution studied taking into account inter-protein interactions

CHAPTER 3

METHODOLOGY

3.1. Quantum-Mechanical methods

Quantum Mechanics (QM) is the most rigorous framework for the development of a computational method aimed to describe a molecular system at the atomic level. *Ab initio* methods aim to solve the time-independent Schrödinger equation to find the wave function which concentrates all information of the microscopic system.

$$\hat{H}\Psi = E\Psi \tag{3.1}$$

$\hat{H}$ is the non-relativistic *Hamiltonian* operator, which consists of five contributions: kinetic energy of nuclei and electrons, nuclei-electrons attraction, nuclei-nuclei repulsion and electron-electron repulsion:

$$\hat{H} = -\sum_k \frac{\hbar^2}{2m_k}\nabla_k^2 - \sum_i \frac{\hbar^2}{2m_e}\nabla_i^2 - \sum_i \sum_k \frac{e^2 Z_k}{r_{ik}} + \sum_{k<l} \frac{e^2 Z_k Z_l}{r_{kl}} + \sum_{i<j} \frac{e^2}{r_{ij}} \tag{3.2}$$

where i and j run over electrons, k and l over nuclei, $\hbar$ is the Planck's constant divided by 2π, m_e is the mass of the electron, m_k is the mass of nucleus k, ∇^2 is the Laplacian operator, e is the charge of the electron, Z_k is the atomic number of nucleus k and r_{ab} is the distance between particles a and b. The complexity of this problem lies in the inter-dependence or *correlation* of nuclei and electrons of the system. An analytical solution of this differential equation is not possible and some approximations have been devised to circumvent this problem leading to a wealth of methods that widely differ in terms of accuracy and computational cost.

Born-Oppenheimer approximation

Since the electron mass is much lower than that of the nuclei, the motion of the electrons will be fast enough to instantaneously adapt to the nuclei motion. In most of the situations, it is reasonable to assume that the motion of electrons and nuclei are decoupled. Only in those systems in which the fundamental and excited states are close in energy, this approximation fails as the motion of electrons and nuclei can be strongly coupled. The assumption of the Born-Oppenheimer approximation separates the wave function into a nuclear and an electronic part.

$$\Psi(\mathbf{r}_i\,;\,\mathbf{R}_j) = \Psi_{\mathrm{e}}(\mathbf{r}_i\,;\,\mathbf{R}_j)\Psi_{\mathrm{N}}(\mathbf{R}_j) \tag{3.3}$$

Now the problem lies in solving the Schrödinger equation for each nuclei configuration, i.e. the *electronic* Schrödinger equation, which is written as:

$$(\hat{\mathrm{H}}_{\mathrm{e}} + \hat{\mathrm{V}}_{\mathrm{n}})\Psi_{\mathrm{e}}(\mathbf{r}_i\,;\,\mathbf{R}_j) = \mathrm{E}_{\mathrm{e}}\Psi_{\mathrm{e}}(\mathbf{r}_i\,;\,\mathbf{R}_j) \tag{3.4}$$

where $\hat{\mathrm{H}}_{\mathrm{e}}$ includes the terms of Eq. (3.1) that depend on the electrons and $\hat{\mathrm{V}}_{\mathrm{n}}$ is the nuclear-nuclear repulsion for a fixed nuclear configuration $\mathbf{R}_j$. Note that the nuclei kinetic energy term of the full Hamiltonian is cancelled and that the inter-nuclei coulombic repulsion energy here becomes a constant that adds to the electronic energy.

From the Born-Oppenheimer approximation then emerges the concept of *Potential Energy Surface* (PES), which is the surface defined by the electronic energy for all nuclear coordinates (potential energy). This concept is fundamental and sets the grounds of many aspects of chemistry. This will be addressed in section 3.4. At this stage, an accurate representation of the molecular system requires the use of two families of methods. The first one aims to solve the electronic problem, whereas the second one explores the configurational space of the nuclei over the PES, the so-called *conformational space*. As we will see, in general, an accurate solution of the electronic problem is generally achieved at the cost of a reduced knowledge of the conformational space and *vice versa*. Therefore,

one always seeks a balance between the accuracy of the calculation of the potential energy and the extent of sampling of the PES.

The Electronic Problem

The simplification of the Schrödinger equation made by the Born-Oppenheimer approximation is not enough to analytically find a wave function of the molecular system. After having neglected the electron-nuclear correlation, now it is the correlation between the electrons that makes the problem particularly troubling. Of course, this is an exception for atoms with only one electron (hydrogenoid atoms), whose analytical solution, as we will see, will be very useful to find a solution for poly-electronic systems. The basis of many computational methods aimed at finding a solution to the electronic problem is the Hartree-Fock method.

3.1.1. Hartree-Fock Method

Because the complexity in solving the electronic problem lies in the inter-electronic interactions, approximations in this direction are needed. Let us first assume a system of N non-interacting electrons. In such a system, the Hamiltonian is separable and can be expressed as a sum of one-electron Hamiltonians, in which the electron-electron interaction term represents a Coulombic interaction potential between the electron and the electrostatic field generated by the rest of electrons. The eigenfunction of the corresponding Hamiltonian becomes the product of N monoelectronic wave functions, known as *Hartree product.* This product, however, does not fulfill the antisymmetry principle that describes the behavior of fermions, such as electrons. The most compact and simple way of expressing an antisymmetric function is the use of a Slater determinant, where each row correspond to an electron and each column to a monoelectronic orbital with a given spin, known as spin-orbital χ_i.

From the application of the exact Hamiltonian to a Slater determinant with a closed-shell configuration, the energy takes the following form:

$$E = 2\sum_{i}^{N/2} H_{ii} + \sum_{i}^{N/2}\sum_{j}^{N/2}\left(2J_{ij} - K_{ij}\right) \tag{3.5}$$

where H_{ii} corresponds to the kinetic and potential energy of each electron moving in the field of the nuclei, J_{ij} is the electrostatic repulsion between a pair of electrons and K_{ij} is the exchange interaction between electrons of the same spin. The exchange interaction has

not a classical counterpart as arises from the antisymmetry of the wave function fulfilling the Pauli principle that establishes that electrons with the same spin have a reduced probability of being close to one another.

According to the Variational principle, the better the approximation to the exact wave function, the lower the energy. In order to find the best wave function of a poly-electronic system described with a single Slater determinant, the energy, as expressed in Eq. (3.5), requires to be minimized. By imposing the condition of minimizing the energy with respect to the molecular orbitals, subject to the constraint that the molecular orbitals are orthonormal, the Hartree-Fock equations are obtained.

$$\hat{f}_i \chi_i = \varepsilon_i \chi_i \tag{3.6}$$

where $\hat{f}_i$ is the monoelectronic Fock operator and takes the following form:

$$\hat{f}_i = -\frac{1}{2}\nabla_i^2 - \sum_k^M \frac{Z_k}{r_{ik}} + \sum_j^{N/2} \left(2J_j(i) - K_j(i)\right) \tag{3.7}$$

where J_j and K_j are the one-electron Coulomb and Exchange operators respectively. This monoelectronic Hamiltonian includes a potential that accounts for the interaction between an electron and the rest of electrons in an average way. With this Hamiltonian, the electrons only *feel* an effective potential created by the rest of electrons and do not interact instantaneously, i.e. their motion is not correlated. In practice, to solve the Hartree-Fock equations it is convenient to expand the molecular orbitals as a linear combination of basis functions: $\phi = \sum_i^n a_i \varphi_i$. This approach is known as the *Linear Combination of Atomic Orbitals* approximation (LCAO). This expansion must be as reduced as possible to minimize the computational cost, but requires enough flexibility to reproduce the exact wave function. The resulting equations are known as the Roothaan-Hall equations.

As the effective potential of the Fock operator requires to know the target wave function beforehand, the solution of the equations is achieved iteratively starting from an initial guess of the solution. The solution is converged when a self-consistent field (SCF) is achieved. Because the correlation is not considered in the calculation of the inter-electronic interaction, the energy difference between that obtained with the Hartree-Fock method, in the limit of an infinite basis set, and the exact energy is named *correlation energy*.

The Hartree-Fock method provides the best possible solution for a wave function described with a single Slater determinant. Subsequent improvement of the wave function can be achieved by including more Slater determinants, thus recovering part of the

correlation energy. One can differentiate two types of correlation: *dynamical* and *static*. The dynamical correlation arises from a Hartree-Fock wave function that is improved by small contributions of many other determinants representing alternative configurations. This reflects the inter-dependence of the motion of electrons. The static correlation, on the other hand, is related to wave functions in which the contributions of few determinants dominate the description of the wave function. This is typical of molecules with nearly degenerate Slater determinants, in which different electronic configurations are necessary for the description of the molecule.

Based on the Hartree-Fock method there are two different ways for computing the wave function. The first one simplifies the Hartree-Fock calculations by parameterizing integrals with the aim to make the calculations much faster. Indeed, the Hartree-Fock method requires the calculation of a high number of Coulomb and exchange integrals. This is the computational bottleneck, whose computational cost scales as N^4, where N is the number of basis functions. The parameterization of the integrals is made against experimental data and in some cases can increase the accuracy of Hartree-Fock. The methods following this philosophy are called *semi-empirical*. The second possibility represents a family of methods that use the Hartree-Fock wave function as a starting point toward finding an improved wave function and recovering the correlation energy. These are the so-called *post Hartree-Fock* methods. Of course, the computational demand of this alternative is notoriously higher.

3.1.2. Post Hartree-Fock Methods

Namely three types of post Hartree-Fock methods can be distinguished:

Perturbation theory

The Rayleigh-Schrödinger perturbation theory provides a scheme by which the wave function can be gradually improved by adding corrections to a given order. The idea behind this is to express the true Hamiltonian as a sum of a more tractable Hamiltonian ($\hat{H}_0$), for which the solution of the Schrödinger equation is known, and a perturbation ($\hat{V}$).

$$\hat{H} = \hat{H}_0 + \lambda\hat{V} \tag{3.8}$$

The λ parameter varies from 0 to 1 and allows expressing both the wave function and the energy as a Taylor expansion of increasing order corrections. Perturbation theory allows obtaining the expressions of the corrections to any arbitrary order.

In the Moller-Plesset (MP-n) approach, the more tractable Hamiltonian ($\widehat{\mathrm{H}}_0$) is expressed as a sum of Fock operators: $\widehat{\mathrm{H}}_0 = \sum_i^n \hat{f}_i$. The eigenvalue of this Hamiltonian is the sum of the energies of the occupied Hartree-Fock orbitals. This does not correspond to the Hartree-Fock energy, since this counts twice the electron-electron repulsion, so that the perturbation must correct this double-counting and include the electron-electron repulsion term of the true Hamiltonian. Therefore the perturbation adopts the form:

$$\widehat{\mathrm{V}} = \sum_{i}^{\mathrm{occ}} \sum_{j>i}^{\mathrm{occ}} \frac{1}{r_{ij}} - \sum_{i}^{\mathrm{occ}} \sum_{j}^{\mathrm{occ}} \left(J_{ij} - \frac{1}{2} K_{ij} \right) \tag{3.9}$$

The first-order correction (MP1) to the zeroth-order energy gives the Hartree-Fock energy. Therefore, one needs, at least, the second-order correction (MP2) to recover part of the correlation energy, which is computed as:

$$\mathrm{E}_{\mathrm{MP2}} = \mathrm{E}_{\mathrm{HF}} + \sum_{i}^{\mathrm{occ}} \sum_{j>i}^{\mathrm{occ}} \sum_{a}^{\mathrm{virt}} \sum_{b>a}^{\mathrm{virt}} \frac{\left[(\phi_i \phi_j | \phi_a \phi_b) - (\phi_i \phi_a | \phi_j \phi_b) \right]^2}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{3.10}$$

The former equation shows that the approximation to the correlation energy is made by considering many excited configurations, which ultimately requires the calculation of a huge amount of integrals. The MP-n method is not variational and thus one may obtain energies lower than the exact one.

Apart from MP2, in the present thesis we have used the spin-component-scaled MP2 (SCS-MP2) method [1] developed by Grimme which outperforms the standard MP2 in the description of the correlation energy. This is a semi-empirical modification of MP2 in which the MP2 correlation energy is partitioned into parameterized contributions from parallel and antiparallel spin components.

Configuration Interactions (CI)

The wave function is expanded with Slater determinants that represent excitations (singles, doubles, triples, etc.) over the fundamental configuration.

$$\Psi_{\mathrm{CI}} = a_0 \Psi_0 + \sum_{\mathrm{S}} a_S \Psi_{\mathrm{S}} + \sum_{\mathrm{D}} a_D \Psi_{\mathrm{D}} + \sum_{\mathrm{T}} a_T \Psi_{\mathrm{T}} + \cdots \tag{3.11}$$

When all possible excitations that can be generated from the Hartree-Fock determinant

are included in the expansion, we have a Full Configuration Interaction (FCI), which provides the best solution for a given basis set. One of the main problems of truncating the CI expansion is that is not *size consistent.* This means that the energy of a system comprised of two non-interacting fragments is not the same as the sum of the energies of the two fragments calculated separately. This property is important for describing correctly dissociation reactions.

Coupled Cluster Theory

Coupled Cluster Theory describes the full-CI wave function as:

$$\Psi_{CC} = e^{T}\Psi_0 \tag{3.12}$$

$$e^{T} = 1 + T + \frac{1}{2}T^2 + \frac{1}{6}T^3 + \cdots = \sum_{k=0}^{\infty}\frac{1}{k!}T^k \tag{3.13}$$

where T is the cluster operator and is given by:

$$T = T_1 + T_2 + T_3 + \cdots + T_N \tag{3.14}$$

The T_i operators generate all possible i^{th} excited Slater determinants. The advantage of using the exponential of T is that for a truncated T, the corresponding Taylor expansion provides all coupled excitations of order *i*. For instance, in the case of the CCSD method, where $T=T_1+T_2$ (single and double excitations), the wave function takes the following form on the basis of Eq. (3.13) and (3.14):

$$e^{T_1+T_2} = 1 + T_1 + \left(T_2 + \frac{1}{2}T_1^2\right) + \left(T_2T_1 + \frac{1}{6}T_1^3\right) + \cdots \tag{3.15}$$

From this expansion, the operator not only includes all single and double excitations, but also higher order excitations that result from coupling single and double excitations. The inclusion of higher-order excitations in this way is what makes this method size consistent. Nowadays, the CCSD(T) method [2], in which the triples contribution is estimated from perturbation theory, is considered the "gold standard of quantum chemistry" for its good compromise between computational cost and accuracy (it almost recovers all correlation energy).

3.1.3. Semi-empirical Methods

Upon increasing the size of the molecular system, with Hartree-Fock the number of integrals to compute increase hugely (as N^4) and with the aim to reduce the computational demand, the semi-empirical methods compute only a fraction of all integrals and parameterize some of them. These parameters are chosen to reproduce experimental data (thermochemical and structural), thus giving the *semi-empirical* adjective.

All semi-empirical methods make the approximation of ignoring core electrons on the basis that these will be less sensitive to changes in the chemical environment. The remaining valence orbitals are represented with a minimal basis set of Slater-type orbitals. The main differences among semi-empirical methods lie in the number of neglected integrals and the way they are parameterized.

CNDO (Complete Neglect of Differential Overlap)[3,4]

All one-electron integrals are parameterized and among all two-electron integrals $(\mu\nu|\lambda\sigma)$, only those integrals of the type $(\mu\mu\,|\lambda\lambda)$ have non-zero parameterized values. If μ and λ belong to the same atom A, the integral adopt a unique value γ_{AA}. When the atoms involved are different (A and B) the integral depends parametrically on the respective γ_{AA} and γ_{BB} values and the inter-atomic distance (r_{AB}). All three- and four-center integrals are neglected. Overall, the computational cost decreases from N^4 to less than N^2, since the number of integrals to compute is drastically reduced and the remaining ones are already parameterized and do not require explicit calculation.

INDO (Intermediate Neglect of Differential Overlap)[5]

The INDO method increases the flexibility of the CNDO method for computing one-center two-electron integrals. The integrals between different types of orbitals are distinguished and adopt different parameterized values, in contrast to CNDO.

NDDO (Neglect of Diatomic Differential Overlap)

The NDDO complements the improvement of INDO over CNDO, in describing one-center two-electron integrals, by adding flexibility to the two-center two-electron integrals. All integrals $(\mu\nu|\lambda\sigma)$ are explicitly computed provided that μ and ν belong to the same atom, and λ and σ are centered in the other atom.

Based on the NDDO formalism, Dewar and Thiel reported the MNDO (*Modified Neglect of Differential Overlap*) method [6]. They suggested modeling the two-center two-electron integrals as interactions between multipoles. By replacing the continuous charge clouds by classical multipoles the calculation is much simpler, thus reducing the

computational demand of these integrals. The nuclear repulsion energy, named as core-core term, in NDDO-based methods must be corrected, since the electron-electron terms do not compensate repulsion between nuclear charges and, at long distances, uncharged atoms or molecules experience a net repulsion. Therefore the core-core term needs to be modified and the way this is corrected underlies the difference among a variety of NDDO-based methods. In MNDO, the core-core term adopts the following form:

$$V_{NN}(\mathrm{A},\mathrm{B}) = Z_A' Z_B' (s_A s_B | s_A s_B)(1 + e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}}) \tag{3.16}$$

where Z_A' denotes that the nuclear charge has been reduced by the number of core electrons. One of the limitations of MNDO was that the repulsion was still too high and this had detrimental consequences in the description of hydrogen bonds. To alleviate this problem, the core-core term was modified by adding four Gaussian functions to each atom in AM1 (Austin Model 1) method [7] by Dewar and co-workers:

$$\mathrm{V_{NN}^{AM1}}(\mathrm{A},\mathrm{B}) = \mathrm{V_{NN}^{MNDO}}(\mathrm{A},\mathrm{B}) + \frac{Z_A' Z_B'}{R_{AB}} \left(\sum_k^4 a_{kA} e^{-b_{kA}(R_{AB}-c_{kA})^2} + \sum_k^4 a_{kB} e^{-b_{kB}(R_{AB}-c_{kB})^2} \right) \tag{3.17}$$

AM1 is one of the most broadly used semi-empirical methods in a variety of applications. However, because the parameterization process of AM1 had not been optimal, Stewart reported a reparameterization of AM1, and modified the core-core term using two Gaussian functions for each atom, in the so-called PM3 (Parameterized Model 3) [8]. RM1 (Recife Model 1) [9] is another more recent reparameterization that keeps the same expression of the core-core term of AM1 and that generally yields better results than AM1 and PM3.

d orbitals in NDDO models

To increase the flexibility of the basis set for an improved description of the wave function, d orbitals are especially necessary for hypervalent atoms such as phosphorus. Some of the methods mentioned above have been extended to use *d*-orbitals by adopting different strategies.

Thiel and Voityuk kept the original expressions and parameters of MNDO and extend it to the use of *d*-orbitals in the MNDO/d model [10]. Following the same philosophy as MNDO, new one- and two- electron integrals involving *d*-orbitals were parameterized. This model represented a significant improvement over AM1 and PM3 in the description

of hypervalent atoms. Based on the MNDO/d formalism, an extension of AM1 to d-orbitals for P, S and Cl atoms was described by Winget and co-workers in the AM1* model [11]. The only difference with standard AM1 is that the core-core term involving the newly parameterized atoms adopts a different expression with two element-pair specific parameters.

An alternative AM1 model with *d*-orbitals was developed by Nam and co-workers, called AM1/d-PhoT [12] for P, H and O atoms involved in phosphoryl transfer reactions. Given that one of the main problems of the standard AM1 was the overstabilization of hypervalent structures, in AM1/d-PhoT the original core-core term of AM1 includes a parameter (G_{scale}) that attenuates the artificially attractive interactions involving P atoms:

$$V_{NN}^{AM1}(A,B) = V_{NN}^{MNDO}(A,B) + \frac{Z_A' Z_B'}{R_{AB}} G_{scale}^{A} G_{scale}^{B} \left(\sum_{k}^{4} a_{kA} e^{-b_{kA}(R_{AB}-c_{kA})^2} + \sum_{k}^{4} a_{kB} e^{-b_{kB}(R_{AB}-c_{kB})^2} \right) \quad (3.18)$$

where G_{scale} for P is 0.3537 and 1.0 for H and O. In this model, a complete reparameterization of AM1 was done based on a set of compounds typically involved in phosphoryl transfer reactions following different mechanisms. Another specific reparameterization of AM1 for phosphoryl transfer reactions was done by Arantes and Loos [13], but this did not incorporate *d*-orbitals and was focused on C, H, O, P, and S atoms.

Following the idea of adopting modified core-core expressions with two element-pair specific parameters, Stewart made an extensive parametrization using a very large set of compounds (~9000) for 70 atomic elements in the development of PM6 (Parameterized Model 6) [14]. A specific core-core term was also designed for improving the description of hydrogen bonds.

3.1.4. Basis sets

The mathematical functions usually employed to construct the wave function are inspired in the atomic orbitals of the hydrogen atom, for which there is an analytical solution to the Schrödinger equation. As these orbitals have a physical meaning, it thus seems reasonable to use a basis set with this type of atomic orbitals (Slater-type orbitals, STOs) centered at the nuclei to construct the wave function of molecules. The radial part of STOs is given by:

$$R(r) = Nr^{n-1}e^{-\zeta r} \tag{3.19}$$

where N is a normalizing constant, n is the principal quantum number, r is the distance of the electron from the nucleus and ζ is a constant that accounts for the partial shielding of the nuclear charge by the electrons.

In practice, however, evaluation of the three- and four-center integrals is computationally inefficient with STOs and, given the large number of integrals to compute, an alternative type of function is needed. In this regard, the computationally more efficient Gaussian-type orbitals (GTOs) are usually preferred for building the basis set.

$$R(r) = Nx^i y^j z^k e^{-\alpha r^2} \tag{3.20}$$

where α determines the width of the Gaussian, x, y, z are the Cartesian coordinates and the integers i, j and k determine the type of orbital. The problem with GTOs stems from their poorer representation of their behavior at short (near the nucleus) and long distances with respect to STOs. This underlies the preference of semi-empirical methods in the use of STOs, where three- and four-center integrals are discarded. To combine the computational efficiency of GTOs with the accurate radial shape of STOs, a combination of GTOs is usually employed to represent a given STO. When a basis function is defined as a linear combination of Gaussians, called *primitives*, it is referred to as *contracted* basis function. When one contracted basis function is used for each atomic orbital, the wave function is called to be described with a *minimal* basis set. This is the case of the STO-3G basis set, where each atomic orbital is represented with three primitives. For increasing the flexibility of the basis set, *double-zeta* (DZ) and *triple-zeta* (TZ) basis sets, which contain two and three basis functions for each atomic orbital respectively, are generally used. Given that valence orbitals are those involved in chemical bonding and, thus, are more sensitive to changes in the chemical environment than core orbitals, the *Split-Valence* basis sets developed by Pople and co-workers are commonly used. Such a basis contains one contracted basis function for core orbitals and double- or triple-zeta basis sets for valence orbitals. This is the case, for instance, of the 6-31G basis set. Moreover, to increase the mathematical flexibility of the basis set to describe molecular orbitals, functions with higher angular momentum than that of the valence orbitals are also employed, which are called *polarization functions.* In addition, by using *diffuse functions* one can also add further flexibility by enabling the basis set to locate electron density far from the nucleus, which is especially needed for describing negatively charged atoms. For instance, the 6-31+G(d) basis set widely used in this thesis is a DZ basis set including diffuse and *d*-polarization functions.

Pople basis sets are characterized by using a *segmented* contraction, which means that different primitives are used for describing basis functions of the same angular momentum. Following the same philosophy, the Alhrichs basis sets are optimized to a higher extent and allow using smaller basis sets to achieve similar accuracy, thus, reducing the computational cost. As an alternative to segmented basis sets, the *correlation-consistent* split-valence basis sets (cc-pV*n*Z) developed by Dunning and co-workers use a *general* contraction, i.e. the same primitives are used for describing basis functions of the same angular momentum. The advantage of this approach is that it makes more efficient the calculation of integrals involving the same primitives, as they are required to be computed once only.

In general, for a proper use of post Hartree-Fock methods it is vitally important to use large basis sets with diffuse and polarization functions of high angular momentum to recover a high percentage of the correlation energy. For instance, this is the case of a Full-CI calculation, which would yield the exact solution with an infinite basis set. Large basis sets are also employed for carrying out single point calculations on molecular geometries that have been energy minimized with a smaller basis set. Because the geometrical parameters are less sensitive than the energy to the size of the basis set, such procedure achieves a good compromise between computational cost and accuracy in determining both energy and structure.

3.1.5. Density Functional Theory methods

Methods based on Density Functional Theory (DFT) follow an alternative route to the Schrödinger equation for describing the molecular system. In DFT, what fully determines the properties of the molecular system is the electron density, as demonstrated by the Hohenberg-Kohn theorems [15]. The first theorem establishes the existence of a one-to-one correspondence between the electron density and the wave function. It proves that the density determines the external potential, which determines the Hamiltonian, which in turn defines the wave function. In this regard, the system energy depends exclusively on the density and, as a consequence, the energy turns out to be a functional of the density. Nevertheless, this theorem only proves the existence of this functional, but does not indicate its expression. The second theorem proves that the electron density follows the variational principle, like the wave function, and therefore the better the approximation to the exact electron density the lower the associated energy. Although this theorem provides a criterion to ascertain whether a given electron density is better than other, it does not indicate how to improve it in a systematic way. With the Hohenberg-Kohn theorems in hand, DFT, in principle, provided an alternative way to the Schrödinger equation, but the ignorance of the exact functional made this theory

unpractical. It was not until Kohn and Sham [16] found a practical way to find the properties of a system directly from the density that the breakthrough in DFT-based methods started.

The fundamental idea behind the Kohn-Sham method is to consider the real system as a fictitious system of non-interacting electrons whose density is the same as that of the real system where electrons do interact. The energy functional thus adopts this form:

$$E[\rho(\mathbf{r})] = T_{ni}[\rho(\mathbf{r})] + V_{ne}[\rho(\mathbf{r})] + V_{ee}[\rho(\mathbf{r})] + \Delta T[\rho(\mathbf{r})] + \Delta V_{ee}[\rho(\mathbf{r})] \qquad (3.21)$$

where T_{ni} is the kinetic energy of non-interacting electrons, V_{ne} the nuclear-electron interaction, V_{ee} the classical electron-electron repulsion, ΔT the correction of the kinetic energy due to the inter-electronic interaction and ΔV_{ee} the quantum corrections to the electron-electron repulsion energy. The corrections to the kinetic energy and inter-electronic repulsion are gathered in the so-called Exchange-Correlation term $E_{xc}[\rho(\mathbf{r})]$. The Kohn-Sham equations that are obtained are mathematically very similar to Hartree-Fock equations:

$$h_i^{KS}\chi_i = \varepsilon_i\chi_i \qquad (3.22)$$

where h^{KS} is the Kohn-Sham monoelectronic operator:

$$h_i^{KS} = -\frac{1}{2}\nabla_i^2 - \sum_k^N \frac{Z_k}{|\mathbf{r}_i - \mathbf{R}_k|} + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}_i - \mathbf{r}'|}d\mathbf{r}' + V_{xc} \qquad (3.23)$$

V_{xc} is the one-electron operator whose expected value is E_{xc}.

There is great parallelism between the Kohn-Sham and Hartree-Fock equations, but it is important to remark an important difference. Hartree-Fock is an approximate theory, whereas the Kohn-Sham method provides the exact solution provided that the exact $E_{xc}[\rho(\mathbf{r})]$ functional is known. Because of the ignorance about the form of this functional, a wealth of approximations has been developed.

Exchange-Correlation functionals

A huge diversity of functionals has been developed under different approximations to evaluate separately the exchange and correlation contributions to E_{xc}. These functionals are built by mathematical expressions and parameters that are fitted to experimental data

and, for this reason, DFT methods could be regarded as semi-empirical methods, although their number of parameters is generally much lower than in those actually classified as semi-empiricals.

Local Density Approximation functionals (LDA)
These functionals are based on the uniform electron gas, in which the energy at a given position depends only on the density value at that point. For this model, the analytical expression of the exchange functional was derived by Slater (Eq. (3.24)) and has a simple form, in contrast to the most widely used LDA correlation functional which corresponds to the mathematical model by Vosko, Wilk y Nusair (VWN) [17].

$$E_x[\rho(\mathbf{r})] = -\frac{9\alpha}{8}\left(\frac{3}{\pi}\right)^{1/3}\int \rho^{4/3}(\mathbf{r})d\mathbf{r} \tag{3.24}$$

LDA is too inaccurate for describing molecular properties because of overbinding in chemical bonds and the underestimation of barrier heights. The application of these functionals is limited to solid-state physics.

Generalized Gradient Approximation functionals (GGA)
Because the electron density of a molecule is not uniform, it is reasonable to improve the LDA approximation by taking into account not only the local density, but also the change of density at a given position, i.e. the gradient. The methods that include a functional of the density gradient as a correction to the LDA functional are known as GGA functionals:

$$\varepsilon_{xc}^{GGA}[\rho(\mathbf{r})] = \varepsilon_{xc}^{LDA}[\rho(\mathbf{r})] + \Delta\varepsilon_{xc}\left[\frac{|\nabla\rho(\mathbf{r})|}{\rho^{4/3}(\mathbf{r})}\right] \tag{3.25}$$

where ε_{xc} is defined as the energy density, which conveniently rewrites $E_{xc}[\rho(\mathbf{r})]$ as:

$$E_{xc}[\rho(\mathbf{r})] = \int \rho(\mathbf{r})\varepsilon_{xc}[\rho(\mathbf{r})]d\mathbf{r} \tag{3.26}$$

Among all GGA corrections developed for the exchange contribution, the functional derived by Becke (B) [18] is the most popular and incorporates a single parameter (β).

$$\Delta\varepsilon_x^B[x] = -\beta\rho^{1/3}(\mathbf{r})\frac{x^2}{1+6\beta x\sinh^{-1}x} \tag{3.27}$$

$$x = \frac{|\nabla\rho(\mathbf{r})|}{\rho^{4/3}(\mathbf{r})} \tag{3.28}$$

Regarding the correlation part, the functionals developed by Lee, Yang y Parr (LYP) [19] and Perdew and Wang (PW91) [20] are the most utilized. The LYP functional was designed to compute the full correlation energy and not a correction to LDA.

Of course, the next possible improvement can be achieved by including an additional correction to a GGA functional with the second derivative of the density. Such type of functional is known as meta-GGA.

Hybrid functionals

From the Hellmann-Feynman theorem, it is established that the Exchange-correlation energy of the interacting real system can be computed from the non-interacting system according to the following expression:

$$E_{xc} = (1-a)E_{xc}^{DFT} + aE_x^{HF} \tag{3.29}$$

This expression is known as the Adiabatic Connection since it connects a non-interacting system with a system that does interact. The advantage of this is that one can approximate the E_{xc} energy by including part of the exchange energy of a non-interacting system, which we do know how to calculate it exactly. This is the exchange energy from a Hartree-Fock calculation. The idea behind this is to add part of the exact Hartree-Fock exchange energy to the GGA functionals. The optimal fraction of Hartree-Fock exchange is optimized against experimental data and varies widely among hybrid functionals. Among them, B3LYP is the most widely used functional. It uses the 3-parameter Becke functional [18] along with the LYP correlation functional.

$$E_{xc}^{B3LYP} = (1-a)E_{xc}^{LDA} + aE_x^{HF} + b\Delta E_x^B + (1-c)E_c^{LDA} + cE_c^{LYP} \tag{3.30}$$

where a, b and c were optimized to 0.20, 0.72 and 0.81 respectively. Despite the popularity of B3LYP, it has important shortcomings such as the underestimation of barrier heights or the bad description of non-covalent interactions. Indeed, a unique functional able to describe accurately, in main-group elements and transition metals, all

different types of molecular interactions and chemical reactions has not been developed so far. For this reason, and because of the parameter-dependence of standard functionals, a plethora of functionals have been designed to serve a specific purpose. For instance, MPW1K [21] was optimized for properly describing the kinetics of H-atom abstractions, whereas *m*PW1N [22] was developed for halide/haloalkane nucleophilic substitution reactions. Therefore, the choice of the XC functional must take into account the molecular properties of interest.

One of the main limitations of current functionals is their inaccurate description of long-range dispersion interactions, which are responsible for non-covalent interactions such as dative or stacking interactions. The problem stems from the incorrect description of the asymptotic $-1/r^6$ dependence of the dispersion interaction energy on the inter-atomic distance. As already mentioned, the energy in current XC functionals depends on the local density and its derivatives, which are also local, so they cannot be accurate in describing electron correlation at long distances. Different approximations for including dispersion interactions have been reported in the literature. For instance, a modified version of the exchange functional by Perdew and Wang (PW) was obtained by Adamo and Barone in the *m*PWPW91 functional [23], which is used in the present thesis for describing pentacoordinated phosphorus compounds. Nowadays, the recent M06 family of functionals by Zhao and Truhlar [24] are among the most accurate and widely used for describing non-covalent interactions and the kinetics and thermochemistry. In recent years, there have also been sound advances in the development of DFT-D methods [25] by including semi-classical corrections of dispersion interactions to standard exchange-correlation functionals. Overall, the different approaches made by many authors to describe exchange-correlation effects have contributed to increase noticeably the number of functionals currently available.

3.1.6. Wave Function Analysis

From the optimized wave function, many molecular properties can be derived besides the energy, such as atomic partial charges or chemical bonds. Because an operator for such properties does not exist, a proper criterion to partition the molecule into atoms requires to be defined. In the following, some of the widely used methods developed to analyze the wave function are described.

Atoms in Molecules

The Atoms in Molecules (AIM) theory [26] developed by Bader characterizes the chemical bond on the basis of the topology of the electron density. The density gradient

determines, by definition, the direction in which the density grows more steeply. The trajectories defined by the gradient field end at points where the local density is maximum (attractor), which corresponds to the nucleus. In the context of AIM, the set of trajectories that share the same nucleus, along with the very nucleus, define the *atom*.

The electron density displays other stationary points that are located between two attractors (called *bond critical points*) which contain information about the chemical bonds of the molecule. The two trajectories of the gradient field that link the bond critical point with the respective attractors and that are orthogonal to the inter-atomic surface define the *bond path*, which is the line of maximum density between two nuclei.

The topological features of the electron density at the bond critical point allow characterizing the type of interaction established between the atoms linked at this point. From the Hessian matrix at the bond critical point, the Laplacian of the density $\nabla^2\rho(\mathbf{r})$ is extracted. In particular, the Laplacian determines whether electron density accumulates at the bond critical point, $\nabla^2\rho(\mathbf{r}) < 0$, which characterizes covalent bonds, or decreases, $\nabla^2\rho(\mathbf{r}) > 0$, as known for non-covalent interactions such as hydrogen bonds, dative bonds or ionic bonds. Indeed, the specific requirement of a chemical bond to be classified as covalent is that the Energy Density, $\mathrm{H}(\mathbf{r})$, at the bond critical point, which is another property of the density, be negative. This density property indicates whether the charge accumulation at the critical point is stabilizing, which is known for covalent bonds.

The rigorous partition of the electron density into areas defining the constituent atoms allows computing partial atomic charges by means of the numerical integration of the electron density associated to each atom.

$$q_k = Z_k - \int_{\Omega_k} \rho(\mathbf{r})\mathrm{d}\mathbf{r} \tag{3.31}$$

where Ω_k is the volume associated to the atom with nucleus *k*.

Electrostatic Potential

When studying molecular interactions in polar molecules it is of great utility the electrostatic potential of the molecule. The electrostatic potential at a given position is a measure of the attraction or repulsion that the molecule would exert on a unit charge located at that position and is given by:

$$V_{ESP}(\mathbf{r}) = \sum_{k}^{N} \frac{Z_k}{|\mathbf{r} - \mathbf{R}_k|} - \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \quad (3.32)$$

where the first and second terms correspond to the nuclear and electronic contributions respectively. The electrostatic potential is calculated over a surface that encompasses nearly all the electron density and, by means of a color map, highlights regions of the molecule with different potential and, thus, reactivity.

Natural Bond Orbital

The Natural Bond Orbital analysis [27] developed by Weinhold and co-workers allows to describe the optimized wave function by means of *natural orbitals*. These orbitals are the best functions to construct a minimum basis set for describing the wave function. Such a description allows to interpret the wave function on the basis of the classical Lewis' concepts of localized electronic structure of molecules.

This basis set of natural atomic orbitals (NAOs) localizes the electron density at single atoms and pairs of atoms. This localization characterizes, on the one hand, orbitals from inner shells and lone pairs and, on the other, chemical bonds. It is a complex mathematical procedure structured in such a way that the NAOs concentrate the highest percentage of the electron density and thus build the most representative Lewis structure of the molecular system. The natural bond orbitals (NBOs) are those that result from the combination of two NAOs belonging to a pair of bonded atoms. The contribution of each atom to the bond can be analyzed as well as the hybridization of the corresponding NAOs being combined, which ultimately aids to determine the type of chemical bond.

The electron density is not totally recovered by the computed natural orbitals. The minimum part of the density that remains to be described by the determined Lewis structure corresponds to the partially occupied antibonding NBOs. The occupation of these orbitals arises from delocalization effects in which the antibonding NBO withdraw charge from an occupied NBO. The charge transfer that accompanies the formation of a given interaction can be evaluated by means of a second-order perturbative analysis, which characterizes hydrogen bond interactions, conjugation and hyperconjugation events.

The NBO analysis also provides a scheme to obtain partial atomic charges from the wave function that is called *Natural Population Analysis*. The occupation of NAOs within the same atom results in the total number of electrons that are assigned to that atom, which ultimately yields an atomic charge named as *natural atomic charge.*

3.2. Molecular Mechanics

The previous section clearly states that quantum-mechanical methods provide the most accurate description of the molecular electronic structure. However, a complete description of a molecular system extends beyond the knowledge of the electronic structure of a single molecular structure and, as invoked by the Born-Oppenheimer approximation, the potential energy surface requires to be explored. For large systems, such as proteins, the computational expense of quantum-mechanical methods makes this exploration unaffordable. The alternative is represented by Molecular Mechanics methods which vastly reduce the cost of the energy calculation by using force-fields constructed with parameterized mathematical expressions that only depend on the nuclear positions and ignore the electrons.

Molecular mechanics force fields express the energy of a molecular system as a summation of different contributions expressed as mathematical functions the parameters of which have been optimized according to experimental data and quantum-mechanical calculations. The most popular force-fields (AMBER [28], CHARMM [29], GROMOS [30] and OPLS/AA [31]) differ in the expressions of their functions and the strategies used for parameter optimization. Among them, the OPLS/AA force field developed by Jorgensen and co-workers has been the choice made in this thesis.

The energy of the system in any MM force field divided into bonding and non-bonding terms ($E = E_b + E_{nb}$). In the following, we describe these terms as defined by the OPLS force field.

3.2.1. Bonding interactions

The energy associated to bonding terms is computed with functions that model the energy penalties due to deviations of internal coordinates from their reference values. The expression of the bonding terms of the OPLS force field are the following:

$$E_b = \sum_{i}^{bonds} \frac{1}{2} k_i \left(\mathrm{d}_i - \mathrm{d}_{i,0}\right)^2 + \sum_{i}^{angles} \frac{1}{2} k_i \left(\Theta_i - \Theta_{i,0}\right)^2 + \sum_{i}^{torsions} \sum_{n=1}^{3} \frac{1}{2} V_n \left[1 + (-1)^{n-1} \cos\left(n\varphi_i + \varphi_{i,0,n}\right)\right] \tag{3.33}$$

The first term corresponds to the stretching between each pair of bonded atoms modeled by a harmonic potential, whose force constant reflects the bond strength. The second term is the angle bending contribution, also modeled by a harmonic potential. The description of these two contributions with a simple harmonic potential is, in principle, sufficient as non significant deviations from the equilibrium position are expected. The third term corresponds to proper torsions which model the energy changes due to bond rotations, which are responsible for the main conformational changes of the molecule. This periodic potential indicates the number of minimum energy conformations resulting from the bond rotation. Other force fields add a fourth term that models improper torsions which define out of the plane bendings.

3.2.2. Non-bonding interactions

The non-bonded terms comprise Van der Waals and electrostatic pair-wise interactions.

$$\mathrm{E_{nb}} = \sum_{i}^{n} \sum_{j \neq i}^{n} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{r_{ij}} \qquad (3.34)$$

The former are modeled with a Lennard-Jones potential, which describe the inter-atomic repulsion at very short distances and the stabilization by virtue of dispersion interactions at relatively long distances. The pair-wise ε_{ij} and σ_{ij} parameters determine the depth and the distance of the interaction energy well respectively. The simplest model for electrostatic interactions is the Coulomb's law, which defines the interaction energy between two point charges separated by a given distance. The non-bonding energy is scaled by 0.5 for those pairs of atoms separated by four consecutive bonds (1-4 interactions).

Non-bonded interactions underlie the most time-consuming part of the MM calculation, since the direct evaluation of these interactions scales as N^2, where N is the number of atoms. Spherical cutoff schemes are widely used approaches to reduce the computational cost by avoiding the evaluation for all possible pairs. Three different cutoff schemes have been developed. In the simplest scheme (truncation), only the interactions within a cutoff distance are computed. This introduces discontinuities in the distance-dependent non-bonding interaction energy, and the corresponding forces, that can lead to artifacts. Other schemes aim to gradually set to zero the distance-dependent interactions. While *shift functions* alter the interaction energy function, E(r), gradually from the very beginning so as to reach the zero value at the cutoff distance, *switch functions* smoothly alter the interaction energy within a buffer region [a,b], so that E(b)=0 and E(r) for $r \leq a$ remains

unchanged.

Given the rapid decay of the Lennard-Jones potential as $1/r^6$, cutoff schemes entail little loss of accuracy provided that sufficiently large cutoffs are used. Electrostatic interactions, on contrary, turn out to be more troubling, as they decay much more slowly (as 1/r) and, therefore, long-range interactions make non-negligible contributions to the electrostatic energy. Even for non-charged particles, dipole-dipole interactions decay more slowly (as $1/r^3$) than Van der Waals interactions. To avoid using excessively large cutoffs and minimize the loss of accuracy, alternative faster methods have been devised, as the Ewald summation method, to compute long-range interactions (see later in section).

Explicit solvation

For the simulation of proteins and other large biomolecules, the solvation of the system requires to be taken into account. The explicit solvation of the system requires the definition of a water model. Because the explicit solvation of such a large system requires a huge number of water molecules, simple MM models have been developed that describe water-mediated polar interactions. These models assume a fixed geometry for the water model and only consider non-bonding terms. They differ in the number of interaction sites. For instance, the SPC (Single Point Charge) [32] and TIP3P [33] models are 3-sites models in which a point charge is defined at the oxygen and two hydrogen atoms. 4-sites models, such as TIP4P [33], shift the negative charge of the oxygen along the bisector of the HOH angle and also consider Van der Waals interactions for the oxygen atom. TIP5P [34] model improves the representation of the overall charge distribution with point charges defining oxygen lone pairs interactions. Of course, the higher the number of interaction sites the higher the computational cost. To make a proper choice of the water model, it is important to take into account that current force fields have been parameterized in conjunction with a given water model. The use of a different water model may lead to some inconsistencies.

Periodic Boundary Conditions (PBC)

The simulation of a solute immersed in a solvent is usually done under Periodic Boundary Conditions. The modeled system is located in a unit cell that is infinitely replicated in the three spatial dimensions. When using PBC the *minimum-image* convention is followed. By minimum image, we mean that when a particle crosses the boundary of the unit cell, an image of that particle enters to replace it, thus conserving the total number of particles in the cell. Within this approximation, non-bonding interactions can use a cutoff distance of $L/2$ at most, where L is the length of the dimension of the box. Larger cutoffs will double-count interactions, since the minimum images of those particles beyond $L/2$ are already within $L/2$. The replication of the unit cell avoids, in principle, surface effects and

thus solvent molecules at the edge of the cell interact with solvent molecules as a bulk.

Depending on the shape of the system different unit cell geometries can be used to construct the lattice, being the cubic shape the most broadly used. Others are more compact for a given thickness of the water layer, e.g. rhombic dodecahedron, thus reducing the amount of solvent molecules needed in the system, being more computationally efficient.

Ewald summation method

Periodic boundary conditions are advantageously used by the *Ewald summation* method to compute the challenging long-range electrostatic interactions at a lower computational cost than that required by cutoff schemes. This technique calculates the electrostatic energy of the system with an infinite number of periodic images adopting a reciprocal-space technique, which was first developed to study the energetics of ionic crystals and that have some parallelism with crystallography.

By definition, the total electrostatic energy of the central box with the infinite array of periodic images is given by:

$$\mathrm{V} = \frac{1}{2} \sum_{|\mathbf{n}|=0}^{\infty} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{|\mathbf{r}_{ij} + \mathbf{n}|} \tag{3.35}$$

$\mathbf{n}$ is the vector ($n_x L$, $n_y L$, $n_z L$), where n_x, n_y and n_z are integers and L the size of the box. The difficulty of this sum is that it is *conditionally convergent*, which means that results from the sum of two divergent summations corresponding to the positive and negative terms respectively. Ewald devised a trick to convert this sum into the sum of two rapidly convergent series. The essential idea is to perform one summation in the *real space* and another in the *reciprocal space*. This is accomplished by surrounding each charge with a Gaussian charge distribution of opposite sign. Now the summation arising from point-charges and Gaussian charges is convergent and is carried out in the real space. The neutralizing Gaussian charge distribution is in turn neutralized by a second Gaussian charge distribution. The infinite summation over the second Gaussian charge distributions is performed in the reciprocal space by Fast Fourier Transformation. In practice, to improve the performance of the reciprocal sum, the Particle Mesh Ewald method (PME) [35], which scales as N log N, finds wide application in molecular dynamics simulations. Some linear-scaling implementations have also been done for hybrid QM/MM calculations [36].

3.3. Hybrid Quantum Mechanical / Molecular Mechanics methods

Describing the reactivity of molecular systems requires a quantum mechanical representation that, even for semi-empirical methods, involves a computational demand that is unattainable when dealing with thousands of atoms. Molecular Mechanics force fields, on the other hand, are only an efficient alternative to QM methods in the absence of bond-breaking/formation events and other electronic processes. The fact that these events tend to occur in a small part of the whole system, e.g. enzyme active sites, is advantageously exploited by hybrid approaches that describe the small reactive region with QM methods and the remaining part of the system with MM force fields, defining the so-called QM/MM technique. The advent of QM/MM approaches was pioneered by the seminal contribution by Warshel and Levitt in 1976 [37] and along the years several distinct schemes have been devised. Within the QM/MM framework the Hamiltonian is defined as:

$$\mathrm{H} = \mathrm{H_{QM}} + \mathrm{H_{MM}} + \mathrm{H_{QM/MM}} \tag{3.36}$$

where $\mathrm{H_{QM}}$ describes the interaction of all quantum mechanical particles with one other, $\mathrm{H_{MM}}$ accounts for the interaction of all particles represented by a MM force field and $\mathrm{H_{QM/MM}}$ evaluates the interaction between both QM and MM particles. Both $\mathrm{H_{QM}}$ and $\mathrm{H_{MM}}$ contributions take the same form as the standard QM and MM methods already commented and what differs among current QM/MM methods is the scheme devised to treat the QM/MM coupling term. The most widely used scheme is what is known as *electrostatic embedding*.

$$\begin{aligned}\mathrm{H_{QM/MM}} = & \sum_{i}^{\substack{solute\\electrons}} \sum_{m}^{\substack{MM\\atoms}} \frac{q_m}{r_{im}} \\ & + \sum_{k}^{\substack{solute\\nuclei}} \sum_{m}^{\substack{MM\\atoms}} \left(\frac{Z_k q_m}{r_{km}} + 4\varepsilon_{km} \left[\left(\frac{\sigma_{km}}{r_{km}} \right)^{12} - \left(\frac{\sigma_{km}}{r_{km}} \right)^{6} \right] \right)\end{aligned} \tag{3.37}$$

The first electrostatic term makes the electrons feel the partial charges of the MM atoms besides the QM nuclei field, i.e. the isolated QM region is *polarized* by the MM electrostatic field, whereas the second electrostatic term introduces the QM nuclei in the

field created by MM charges. The Lennard-Jones contribution plays a more structural role by avoiding both regions to be in excessively close contact as its effect is primarily limited to boundary atoms. Although quantum mechanics-based methods are more rigorous to describe molecular interactions, non-bonding interactions are well described by MM potential energy functions. Clearly a more realistic representation of the QM/MM coupling would incorporate the polarization of the MM environment, which ultimately would affect the way it polarizes the quantum region. Some advances in this direction have been done, but polarizable force fields are still in infancy and the applications to QM/MM schemes have been limited.

The QM/MM coupling must be carefully described when the boundary is defined across chemical bonds, which is the case for most of the situations when dealing with an enzymatic system. Even in the absence of a reaction between the substrate and the enzyme in the course of catalysis, e.g. covalently bound intermediates or proton transfers between the substrate and acid or basic amino acid residues, the inclusion, in the QM region, of some residues actively involved in the catalysis is necessary to properly account for the mutual polarization of both the substrate and the enzyme. The description of the boundary must take into consideration that the valence of any QM atom participating in a bond being cut needs to be saturated. Among the different schemes developed to accomplish this, the *Link atom* and the *Generalized Hybrid Orbital* (GHO) [38] are the two most broadly used. The former artificially binds an atom, which is usually a hydrogen, to the QM atom. Given that the boundary is usually defined as cutting C-C bonds, the replacement of the original carbon atom by a hydrogen one, which has a similar electronegativity, is not expected to alter significantly the original environment of the QM atom at the boundary. A problem with this approach is the *overpolarization* exerted by the frontier MM atom on the boundary QM atoms due to its close distance to the link atom. To alleviate this problem, one of usual procedures is to redistribute the charge of the MM atom with their bound MM atoms or using more physically realistic representations such as gaussian charge distributions centered on the MM boundary atoms. The GHO method [38] developed by Gao and co-workers, on the other hand, is an alternative boundary scheme that circumvents the *overpolarization* issue. At the MM frontier atom, GHO centers an orbital, which points toward the boundary QM atom, that is freely optimized in the SCF calculation. On balance, given the approximations needed at the boundary, to minimize artifacts it is important to carefully place the boundary as far from the reactive atoms as computationally feasible.

3.4. Potential Energy Surface: stationary points and conformational sampling

So far we have described the wide variety of methods available to evaluate the potential energy for a given nuclei configuration. Clearly this is not enough to describe a molecular system. Under the Born-Oppenheimer approximation, the nuclei move throughout a hypersurface, with 3*N*-6 internal degrees of freedom, whose topology determines the reactivity and other molecular properties of the system. According to statistical thermodynamics, it is by exploring this hypersurface that we can extract the information necessary to bridge the microscopic and macroscopic (observable) properties of the system. Altogether makes the exploration of the potential energy surface an essential endeavor. In this section, we address different methods concerning the exploration of the potential energy surface aimed at characterizing the reactivity as well as the conformational diversity of a molecular system. From the previous section, it is clear that the choice of the method to evaluate the energy of the system is inextricably linked to the extent we aim to explore the potential energy surface and, ultimately, the type of information we want to extract from the PES. For this reason, we have divided this section in two parts: (1) methods to localize stationary points and reactions paths and (2) sampling methods for deeper explorations of the PES

3. 4. 1. Location of stationary points

Statistical thermodynamics establishes that the properties of the most populated ensembles of configurations of the microscopic system are those that determine the properties of the macroscopic system. The weight of a configuration ($\mathbf{x}$) is given by the Boltzmann law:

$$\mathrm{p}(\mathbf{x}) \propto \mathrm{e}^{-\frac{\mathrm{E}(\mathbf{x})}{k_b T}} \tag{3.38}$$

The population of a given configuration decreases exponentially with the energy, so that low energy configurations turn out to be the most representative of the system. Higher energy configurations, on the other hand, can also be relevant provided that they are very numerous, i.e. high entropy. Other types of high-energy configurations are also important in the sense that they connect two stable minima. On the one hand, we are interested in locating the multiple minima that characterize different stable states of the rough energy surface and, on the other, elucidating how the system changes from one minimum to another. Indeed, any chemical reaction, complex formation or conformational change of

the system entails a transition between two different minimum energy structures. We will refer to such pair of minima as *reactants* and *products*, but this does not necessarily involves a bond breaking/formation process. The relative stability of these two minima determines the thermodynamics of the process. The transition between two minima goes through a high-energy *transition state surface,* which determines the kinetics of the process. The minimum in this surface is the *transition state structure* (TS). One way to characterize this passage from reactants to products, i.e. the reaction mechanism, is to find the *minimum energy path* (MEP), also known as the *intrinsic reaction coordinate* (IRC). Overall, extracting the most relevant information of the energy surface demands the concerted use of efficient algorithms able to locate minima, transition states as well as reaction paths.

3.4.1.1. Energy minimization methods

All minima and transition states correspond to *stationary points* of the hypersurface, which means that the first-derivative of the energy with respect to the nuclear coordinates, i.e. energy gradient, is zero. The lack, however, of an analytical expression of the potential energy surface forces the search of stationary points to be done numerically by using iterative algorithms. We can classify the optimization methods into two groups: those which only require first-derivatives of the energy with respect to the coordinates and those that also need second-derivatives. It is important to remark that these optimization methods converge to the local minimum. The search for the lowest energy structure among all minima, i.e. *global energy minimum*, is a challenging task for which there is not a single method that guarantees its finding.

First-derivatives methods

The derivative of the energy provides useful information to guide the minimum search, since the force acting on each atom, which is equal to minus the gradient, points to lower energy structures. The two most widely used minimization algorithms using only first derivatives are the *Steepest Descent* and *Conjugate Gradient*.

The steepest descent method takes a step along the direction of the force, which is the steepest direction at a given point of the PES. This method is very efficient at the first stages of a minimization process to relieve the highest energy features of the structure, but suffers from slow convergence.

The Conjugate Gradient (CG) algorithm outperforms the steepest descent method near the energy minimum by taking conjugate directions instead of perpendicular ones. The conjugate direction results from a combination of the gradient and the previous line

search.

$$\mathbf{v}_k = -\mathbf{g}_k + \gamma_k \mathbf{v}_{k-1} \tag{3.39}$$

where γ is a scalar whose definition depends on the specific CG method.

Second-derivatives methods

The second-derivative of the energy, i.e. Hessian matrix, adds information about the curvature of the function. Of course, this implies a higher computational cost that is justified for difficult minimization problems and small systems. The simplest of these methods is the *Newton-Raphson* (NR) method. On the basis of the Taylor expansion of a function to second-order:

$$\mathrm{f}(\mathbf{x}) \simeq \mathrm{f}(\mathbf{x}_0) + \mathbf{g}^{\mathrm{t}}(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^{\mathrm{t}}\,\mathbf{H}\,(\mathbf{x} - \mathbf{x}_0) \tag{3.40}$$

each step of NR is expressed as:

$$\mathbf{x}_{k+1} = \mathbf{x}_{\mathrm{k}} - \mathbf{H}^{-1}\mathbf{g} \tag{3.41}$$

this step requires the calculation of the inverse of the Hessian matrix, which is computationally demanding and problematic with near-zero eigenvalues. Moreover, the Hessian must be positive definite (all eigenvalues are positive) to ensure that the process minimizes the energy. This method performs better near the minimum where the quadratic approximation is more valid.

The computational cost of calculating and storing the Hessian at each iteration step motivated the development of methods approximating the Hessian on the basis of computed gradients. These methods are known as *Quasi-Newton* methods and, among them, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) is one of the most popular ones.

In general, once the minimization process converges it is important to check that the topology of the stationary point is correct. A calculation of the Hessian matrix reveals that the point is a true minimum when all eigenvalues are positive.

3.4.1.2. Transition state structure and reaction paths

The transition state structure is identified by a Hessian with one negative eigenvalue, which corresponds to a first-order saddle point of the potential energy surface. It corresponds to the highest-energy species along the IRC, i.e. the path that a particle would follow moving from the saddle point along the steepest descent direction with an infinitely small step down to each minimum. Those geometrical variables that change along the IRC define what is known as the *reaction coordinate*.

For locating transition states one can make a distinction between two families of methods: (1) those that optimize a starting structure reasonably close to the true transition state (local methods) and (2) those that require to know the two connected minima (reactants and products).

The first category of TS-search methods are based on the Newton-Raphson methods commented above. A good candidate structure to the true TS implies that the Hessian has an eigenvector with a negative eigenvalue pointing to the direction of the transition of interest. This Hessian guides the optimization process to minimize all degrees of freedom, except one whose energy is maximized, eventually leading to the transition state structure. The ability to propose a good candidate structure for a TS search lies in the ease by which the main features of the true reaction coordinate can be predicted. In general, this can be accomplished for simple reactions by performing systematic constraint minimizations at different points (*scanning*) along the hypothetical reaction coordinate. The highest-energy structure of the scanned coordinate is the best approximation to the transition state.

Once a given transition state structure has been reached, the most usual way to obtain the reaction path is by moving downhill to the two minima. By using a steepest descent algorithm with a finite step size, as we have shown above, the resultant path would oscillate about the true intrinsic reaction coordinate. Of the several schemes devised to circumvent this problem, the Gonzalez-Schlegel method [39] is the most widely used approach to the true IRC. Indeed, following the IRC forward and backwards from the saddle point aids to check whether the TS actually connects the reactants and products of interest. In some cases, especially when using local search methods, the optimization of the transition state can converge to a transition state connecting two other minima.

For the second category, several methods have been developed that make interpolations based on the two minima, such as the String method [40]. Others are able to find the reaction pathway by constructing a set of structures connecting both minima, such as the Nudged Elastic Band method (NEB) [41] developed by Jónsson and co-workers. This method first linearly interpolates a set of structures or *images* between reactants and products. These images are connected by harmonic springs to build an “elastic band” that

is progressively optimized to obtain the minimum energy path. Each image *i* is subjected to a force that is defined as:

$$\mathbf{F}_i = -\nabla \mathrm{V}(\mathbf{R}_i)|_{\perp} + \mathbf{F}_i^S|_{||} \tag{3.42}$$

where the first term is the perpendicular component of the force felt due to the potential energy surface V and the second term corresponds to the parallel component of the spring force on the tangent of the path. The goal in the NEB method is to optimize the images in a concerted fashion so that the force acting on each image is zero.

The spring forces aim to keep the images uniformly spaced and adopt the simple form of a harmonic potential as:

$$\boldsymbol{F}_i^S = k_{i+1}(\mathbf{R}_{i+1} - \mathbf{R}_i) + k_i(\mathbf{R}_{i-1} - \mathbf{R}_i) \tag{3.43}$$

The tangent of the path at image *i* was originally defined as the vector joining images *i*+1 and *i*-1. However, alternative definitions of the tangent have been proposed exhibiting improved performance.

3.4.2. Sampling Methods

It is worth bearing in mind that all energy-minimization methods are conceived to yield the lowest-energy structure of a given basin. In reality, however, this is merely an approach to the state defined by this basin of the PES, since temperature promotes fluctuations within the basin implying the existence of many other similar structures that contribute to describe this state. In this section, we will address the frequently used Molecular Dynamics simulation technique and other methodologies aimed at sampling the potential energy surface.

3.4.2.1. Molecular Dynamics

Molecular dynamics simulations integrate Newton's equations of motion to yield time-trajectories that trace the time evolution of atomic positions and velocities. On the one hand, these trajectories serve as a conformational search over the PES. On the other, monitoring microscopic properties over time allows the prediction of equilibrium macroscopic properties of the system. According to the ergodic hypothesis, an average of the value of a given property over time is equivalent to the average over all configurations

defining the corresponding statistical-thermodynamical ensemble. Therefore simulations must be long enough to extract statistically significant information from time-trajectories in order to predict observable properties.

Integration of equations of motion

The integration of equations of motions cannot be carried out analytically and require the use of finite-difference algorithms. As such, they subdivide the integration into small time steps Δt and require the calculation of forces acting on each particle at a given time t. These forces allow to compute the acceleration, according to the second Newton law, and the new velocities and positions at a time $t+\Delta t$. Integration algorithms assume that the time-dependent positions can be expressed with a Taylor expansion:

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + v(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 \tag{3.44}$$

By adding the former equation to the corresponding expansion for the reverse time step, i.e. $\mathbf{r}(t-\Delta t)$, one obtains the widely used *Verlet algorithm*:

$$\mathbf{r}(t+\Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \mathbf{a}(t)\Delta t^2 \tag{3.45}$$

where the acceleration is obtained directly from the force at time t. The main drawback of this algorithm is that the velocities are not included explicitly and tend to lose numerical precision. Two variations of the Verlet algorithm that circumvent both problems are the *Leap-Frog* and the *Velocity-Verlet* algorithms. For instance, the *Velocity-Verlet* algorithm provide positions and velocities at each time step:

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + v(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 \tag{3.46}$$

$$\mathbf{v}(t+\Delta t) = \mathbf{v}(t) + \frac{1}{2}\Delta t[\mathbf{a}(t) + \mathbf{a}(t+\Delta t)] \tag{3.47}$$

Time step and constraining algorithms

One important requirement of an integration algorithm is energy conservation. This aspect is closely related to the size of the time step chosen. If the time step is too large, high energy overlaps will arise causing instabilities in the integration algorithm. However, excessively short time steps will not allow to cover time scales long enough to obtain relevant chemical information of the system. Therefore, one needs to find a balance between computational expense and stability in the numerical integration. The time step

usually adopted is 1 fs, which is one order of magnitude shorter than the fastest possible motion in a molecular system, which are bond vibrations. Clearly a useful strategy to increase the time step size without causing instabilities is by freezing the bonds. To this aim, several methods including constraints in the equations of motion have been developed, being the most widely used the SHAKE [42] and LINCS [43] algorithms.

Control of temperature and pressure

To obtain macroscopic properties of the simulation of the system the proper statistical-mechanical ensemble needs to be calculated. By following the equations of motion described above the potential and kinetic energy of the system will fluctuate and exchange, so that the total energy will be conserved describing the NVE ensemble (constant number of particles, volume and energy). This ensemble, however, is not appropriate to describe molecular properties of real systems as many experimental studies are carried out at constant temperature and/or pressure. In these conditions, the thermal energy of the system is exchanged with the exterior. Therefore MD simulations require incorporating thermostats and barostats to maintain constant these variables. Note that the term constant does not mean that at each time step the variable has the same value, but that along the simulation the variable oscillates around an average value and does not drift.

Constant temperature dynamics

Because the kinetic energy of the system arises from atomic velocities, the simplest way to maintain the temperature constant is by directly scaling velocities. A widely used thermostat based on this idea is the Berendsen thermostat [44], in which an external heat bath exchanges heat with the coupled system, so that the temperature is maintained. At each step the velocities are scaled with a rate that is proportional to the difference in temperature between the bath and the system.

$$\Delta T = \frac{\Delta t}{\tau}\left(\mathrm{T}_{\mathrm{bath}} - \mathrm{T}(t)\right) \tag{3.48}$$

where τ is the coupling temperature constant, so that the lower the value the stronger the coupling. Then the velocities are scaled by a factor λ that takes the form:

$$\lambda = \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_{\mathrm{bath}}}{T(t)} - 1\right)} \tag{3.49}$$

Other thermostats are known to provide a better description of the thermal energy

distribution throughout the system than those based on scaling atomic velocities. Among them, Nosé and Hoover [45,46] developed a thermostat that produces the correct NVT ensemble by introducing the bath into the system as an additional component, and not as an external bath.

Constant pressure dynamics

The pressure of the system is maintained by changing the volume of the simulation box accordingly. Similarly to the Berendsen thermostat, the system can be coupled to a *pressure* bath (Berendsen barostat) [42], where the change of pressure is given by:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \frac{1}{\tau_\mathrm{p}}\left(P_{bath} - \mathrm{P}(t)\right) \tag{3.50}$$

where τ_p is the coupling constant, P_{bath} is the target pressure and $\mathrm{P}(t)$ is the pressure of the system at time t. The volume of the box is scaled by a factor:

$$\lambda = 1 - \kappa\frac{\Delta t}{\tau_\mathrm{p}}(P - P_\mathrm{bath}) \tag{3.51}$$

where κ is the isothermal compressibility that is related to the volume fluctuations.

Setting up a simulation

To start a simulation the coordinates and velocities of the system are required. The initial coordinates are based on experimental data, e.g. crystallographic structure, or a theoretical model. The initial assignment of velocities is usually based on the Maxwell-Boltzmann distribution at the target temperature. This gaussian distribution sets the probability that an atom of mass m has a velocity v at temperature T. By generating random numbers in the normal distribution, initial velocities are extracted from the Maxwell-Boltzmann distribution. These initial velocities are scaled so that the resulting total momentum is zero and the total energy corresponds to the temperature of interest.

Once initial velocities are assigned an equilibration phase is followed before the data-collection phase. The importance of equilibrating the system lies in ensuring that the kinetic energy (atomic velocities) is partitioned roughly equally among all degrees of freedom and oscillates around a mean value. After initializing velocities and equilibrating the system the data-production phase begins.

Running the simulation

The most computationally demanding part of each simulation step is the calculation of forces limiting the time scale accessible to the simulation. The current computational power allows MD simulations with molecular mechanics force fields to reach hundreds of nanoseconds, whereas those using a quantum-mechanical potential energy function (typically a semi-empirical) only have access to few picoseconds. Depending on the molecular properties of interest this is enough. As we will see in section 3.5 the simulation of mean-square-displacements as measured by neutron scattering experiments requires to sample conformational changes occurring in the picosecond time scale, so that simulations of few nanoseconds are sufficient to ensure convergence of the value.

Principal Component Analysis

Once we have simulated a trajectory it is not obvious the relevant information we can extract from direct visualization of the trajectory. It is possible to describe the conformational fluctuations in terms of collective variables that concentrate the most important dynamic information from the trajectory and filter out the noise from irrelevant local motions. This can be done by doing a principal component analysis (PCA) of MD trajectories, also known as *essential dynamics* [47]. First, the covariance matrix from the fluctuations of atomic positions is built as:

$$C_{ij} = \langle (\mathrm{q_i} - \langle \mathrm{q_i} \rangle)(\mathrm{q_j} - \langle \mathrm{q_j} \rangle) \rangle \tag{3.52}$$

where q_k is the k component of vector $\mathbf{q} = \{\mathrm{q}_1, \ldots, \mathrm{q}_{3N}\}$ which defines the coordinates of the system of N atoms. $\mathbf{C}$ is a symmetric $3N$ x $3N$ matrix, whose diagonal elements represent the atomic mean-square-fluctuations and the off-diagonal elements the correlation between two variables. The eigenvectors of $\mathbf{C}$ are $3N$-dimensional vectors that indicate the direction of motion of the principal components or *essential modes* and the corresponding eigenvalues correspond to the mean-square-fluctuations of the mode. The higher the eigenvalue, the higher the weight of the essential mode in the description of the dynamics of the system, since a larger fraction of the total variance of the trajectory is explained. Each eigenvector defines the direction of motion as a displacement from the average structure. The trajectory can be projected into a principal mode k ($\mathbf{v}_k$) as: $p_k(t) = \mathbf{v}_k^{\mathrm{t}} \cdot (\mathbf{q}(t) - \langle \mathbf{q} \rangle)$. Such analysis is particularly useful for characterizing conformational transitions with different amplitudes occurring in the course of the simulation.

Overcoming molecular dynamics limitations

MD simulations describe conformational fluctuations over a broad range of time scales, i.e. from picoseconds to hundreds of nanoseconds. This allow for accurate sampling of

local motions of amino acid sidechains and subdomains that take place at fast time scales. However, the main limitation of molecular dynamics is the limited amount of conformational space that can be explored. Large-amplitude conformational changes, as domain motions related to substrate-binding and allosteric events, occur at slower time scales (micro-milliseconds) that are inaccessible by standard MD techniques.

A plethora of sampling methods aimed to broaden the exploration of the conformational space has emerged in the last decade. We will not describe here the vast amount of sampling techniques currently available, but highlight some of the most relevant. Of increasing importance are Coarse-grained models, which vastly reduce the number of degrees of freedom and interaction sites by replacing sets of atoms by beads. In particular, Elastic Network Models (ENMs) use a coarse-grained representation of the protein that in combination with Normal Mode Analysis (NMA) allow to decompose the global motion of the protein into a set of modes of motion that give insight into conformational changes of functional relevance. In this thesis we have used this technique to access large-amplitude motions at reduced computational cost.

3.4.2.2. Elastic Network Models

Elastic Network Models (ENMs) use a coarse-grained representation of the protein in which the C-alpha carbons of amino acid residues define the nodes of a network, where each pair of nodes within a cutoff distance interacts via a harmonic potential. A normal mode analysis of this coarse-grained representation of the protein sheds light into the most accessible large-amplitude modes of vibration. Different types of ENMs have been developed, but all provide a consistent description of large-amplitude motions in proteins. Here we describe the ENMs by Bahar and co-workers that have been used in this thesis.

Gaussian Network Model (GNM)

The GNM defines the potential as a function of the vectorial distance between each pair of nodes and is written as:

$$\mathrm{V_{GNM}} = \frac{\gamma}{2}\left[\sum_{i,j}^{N} \mathbf{\Gamma}_{ij}\left(\mathbf{R}_{ij} - \mathbf{R}_{ij}^{0}\right)^{2}\right] \tag{3.53}$$

where N is the total number of protein residues, γ the uniform spring force constant for all residue pairs, $\mathbf{R}_{ij}$ and $\mathbf{R}_{ij}^{0}$ are the instantaneous and equilibrium distance vectors

between residues *i* and *j*, and $\mathbf{\Gamma}$ the *N* x *N* Kirchhoff matrix that defines the topology of the network and is defined as follows:

$$\Gamma_{ij} = \begin{cases} -1 & i \neq j \quad \wedge \quad \mathrm{R}_{ij} \leq r_c \\ 0 & i \neq j \quad \wedge \quad \mathrm{R}_{ij} > r_c \\ -\sum_{k \neq i}^{\mathrm{N}} \Gamma_{ik} & i = j \end{cases} \tag{3.54}$$

where r_c is the cutoff distance (typically 7 Å in GNM calculations). Its non-zero off-diagonal terms are the interacting pairs and the diagonal terms are the coordination number of the residue. The contact topology matrix $\mathbf{\Gamma}$ totally determines the dynamics, since $\mathbf{\Gamma}^{-1}$ defines the mean square fluctuations of each node (diagonal elements) and the cross-correlations of each pair of nodes (off-diagonal elements). The equilibrium positions coincide with the coordinates of the input model, e.g. crystallographic structure.

The GNM approach allows decomposing the global motion of the protein in a set of normal modes of motion. These GNM modes are obtained by eigenvalue decomposition of $\mathbf{\Gamma}$ and its contribution to the global motion is expressed as:

$$\left[\Delta \mathbf{R}_i^2\right]_k = \frac{3k_b T}{\gamma} \frac{(\mathbf{u}_k)_i^2}{\lambda_k} \tag{3.55}$$

where λ_k and $\mathbf{u}_k$ are the k^{th} eigenvalue (mode frequency) and eigenvector (shape of mode) respectively. The $(\mathbf{u}_k)_i$ term defines the mobility of residue *i* along the k^{th} mode. The low-frequency modes are those that have the highest degree of collectivity and can provide insights into cooperative motions that give rise to the biological function. These modes are those that make the largest contribution to the mean-square-fluctuations, which are directly related to crystallographic B-factors.

Overall GNM provides information on the relative size of motion of the residues, but not on the directionality of these motions as fluctuations in this model are assumed to be isotropic. The 3D characterization of the normal modes is provided by the Anisotropic Network Model.

Anisotropic Network Model (ANM)

The ANM potential is a function of the scalar distance between the interacting pair of nodes, as opposed to the vectorial distance used in GNM, and is given by:

$$V_{ANM} = \frac{\gamma}{2}\left[\sum_{i,j}^{N}\left(|\mathbf{R}_{ij}| - |\mathbf{R}_{ij}^{0}|\right)^2\right] \tag{3.56}$$

Both GNM and ANM penalize inter-residue distance changes, but GNM also accounts for the orientational deformation of the inter-residue vector. For this reason, the residue fluctuations obtained by means of ANM tend to be higher than with GNM, so that a larger cutoff is used (15 Å) to obtain better agreement with crystallographic *B*-factors.

A normal mode analysis can be carried out from the 3*N* x 3*N* Hessian matrix **H** of the ANM potential. The modes of motion and its frequencies are extracted thus from diagonalisation of **H** yielding 3*N*-6 non-zero modes, as opposed to *N*-1 modes in GNM. A given ANM mode contains the *x*-, *y*- and *z*- components of the motion of each residue providing directionality to the normal modes. The directionality of the modes provided by ANM can be used to generate deformed structures.

Other ENMs following different approximations have also been developed. Hinsen [48] introduced a variation in ANM by using a force constant that, instead of being uniform among all pairs of interacting nodes, is a parameterized function k(r) that decays with the inter-residue distance. More recently, the ed-ENM model [49], which was trained against a database of molecular dynamics trajectories, includes also a definition of the force constant that depends both on the Cartesian and sequence distance of residue pairs. Other methods are devised to analyze very large structures, such as the Rotations-Translation of Blocks (RTB) [50] and the Block Normal Mode (BNM) [51] models, which are based on partitioning the protein into a set of blocks, which define the degree of coarse-graining.

Analytical tools with normal modes

With the normal modes of motion in hand, one can extract useful information with a set of algebraic operations. For instance, to ascertain which normal modes are more relevant to protein function, one can determine the overlap between each mode and an experimentally observed conformational change, e.g. deformation between apo and holo protein states. Moreover, the degree of similarity between the conformational spaces accessible by two systems can be quantified by the overlap between the subspaces described by a given subset of modes (subspace overlap). To identify rigid-body motions of parts of the system, the use of distance variation maps is particularly helpful. In addition, to understand how the dynamics of a protein region is affected by the rest of the system, the use of an effective hessian for the subsystem of interest in the normal mode analysis is very useful in a variety of applications, e.g. oligomerization effects in the

dynamics of subunits or comparing the dynamics of proteins with different sequences but similar in structure. These are some examples of the analysis performed in the course of this thesis and that are summarized in the articles.

Network models of protein communication

Chennubhotla and Bahar developed a Markov-based model of network communication [52,53] to analyze signal transduction events among protein residues. A Markov process is a stochastic process in which the probability of occurring an event depends merely on the immediately previous event.

This Markov-based model was inspired in the aforementioned Gaussian Network Model. In this model, the interaction between pairs of residues is defined with an affinity matrix, **A**, whose A_{ij} elements are calculated as:

$$A_{ij} = \frac{N_{ij}}{\sqrt{N_i N_j}} \tag{3.57}$$

where N_{ij} is the number of atom-atom contacts between residues *i* and *j* within a cutoff distance of 4 Å and N_i, N_j are the number of heavy atoms of both residues. From the affinity matrix it is defined the degree diagonal matrix $\mathbf{D} = \{d_i\}$ where $d_i = \sum_{j=1}^{N} A_{ij}$ and reflects the packing density at each residue.

A Markov process of communication across the network requires the definition of the Markov transition matrix **M**, where $M_{ij} = A_{ij}/d_j$ is the conditional probability of transmitting a signal from residue *j* to residue *i* in one time step. By defining $-\log(M_{ij})$ as the distance between two residues, in terms of communication, it can be built the maximum-likelihood path that communicates either residue pair using Dijkstra's shortest path algorithm [54]. This communication pathway provides information about which residues are involved in the transmission of signals between residues.

From the affinity and degree matrices, the Kirchhoff matrix associated to the topology of the network is defined as: $\mathbf{\Gamma} = \mathbf{D} - \mathbf{A}$. Indeed, the Kirchhoff matrix in GNM is built in the same manner, but simplifies the pair-wise interactions by setting $A_{ij} = 1$ for those residue pairs within the cutoff distance. This Markovian description of signal transmission allows determining two basic communication properties: *hitting time* and *commute time* [53]. Hitting time H_{ji} is the number of steps it takes to send information from residue *i* to residue *j*, whereas commute time C_{ij} is the sum: $H_{ji} + H_{ij}$, so that it refers to the number of steps it takes to close a cycle of communication between a residue pair. Both properties are directly related to the inverse of the Kirchhoff matrix and the

local density at each residue:

$$H_{ji} = \sum_{k=1}^{N} \left(\Gamma_{ki}^{-1} - \Gamma_{ji}^{-1} - \Gamma_{kj}^{-1} + \Gamma_{jj}^{-1} \right) d_k \tag{3.58}$$

$$C_{ij} = \left(\Gamma_{ii}^{-1} + \Gamma_{jj}^{-1} - 2\Gamma_{ij}^{-1} \right) \sum_{k=1}^{N} d_k \tag{3.59}$$

Likewise GNM, Γ^{-1} contains the information on the dynamics of the system, so that both expressions establish a crucial link between intrinsic protein dynamics and its inherent ability to transmit signals across the structure [53]. See the original publication for details on how Eqs. (3.58) and (3.59) are obtained.

3.4.3.2. Brownian Dynamics

We have already seen that standard molecular dynamics have the limitation of describing protein dynamics up to the nanosecond time scale, which implies that the information on large-amplitude motions is inaccessible and thus alternative techniques are required to achieve enhanced sampling. For computational reasons, another dynamic process describing protein behavior that molecular dynamics cannot describe is the diffusion in concentrated protein solutions, where the size of the system to consider is extremely large and the time scales of interest much longer (microseconds) [55]. There are MD studies describing the association process of two proteins in solution, but what is limiting is the description of diffusional properties in crowded solutions where the interactions with many other protein molecules affect the translational and rotational diffusive behavior. For this purpose, Brownian dynamics simulations are well suited to describe diffusional properties taking into account interactions among hundreds of molecules at time scales ranging from nanoseconds to milliseconds [56,57]. Several assumptions are made in this kind of simulations and it is important to be aware of their impact in the results.

Theoretical framework

The Brownian motion was first identified by Robert Brown who observed the irregular motion of fine particles immersed in a fluid due to collisions with the much smaller solvent molecules. It was Einstein who described the physics behind this kind of motion. The basic assumption in Brownian dynamics (BD) simulations lies in setting a parallelism between the motion of protein molecules in solution, which are much larger than solvent molecules, and this kind of motion (Brownian). This implies adopting an implicit

description of the solvent to take into account electrostatic screening effects. The second assumption is the treatment of protein molecules as rigid bodies, so that the only type of motions considered are overall translations and rotations. Of course, the absence of explicit water molecules along with a rigid-body description of the protein molecule drastically reduce the number of degrees of freedom of the system providing access to much longer time scales than with MD.

The algorithm to integrate the Brownian motion was described by Ermak and McCammon [58]:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \frac{D\mathbf{F}(t)}{k_b T} + \mathbf{R}(\mathrm{t}) \tag{3.60}$$

where $\mathbf{r}$ is the position of the center of mass of the molecule, D is the translational diffusion coefficient at infinite dilution, $\mathbf{F}$ is the force acting on the molecule, $\mathbf{R}$ is a random displacement vector and Δt the time step. An analogous expression to Eq. (3.60) is described for rotational diffusion. The algorithm only considers the positions at each time step and no information on velocities is required. Since the internal dynamics of the protein is neglected, larger time steps than in MD can be adopted, typically being 1-2 ps. The force acting on the protein is generally attributed to electrostatic interactions and steric effects due to the excluded volume originated from the presence of other molecules in the system. In what follows, we describe the energy model adopted by Elcock and co-workers for the simulation of crowded protein solutions [56,57].

Electrostatic interactions

For the calculation of the electrostatic interaction between two protein molecules in solution, the protein is treated as a body with a low dielectric constant and the solvent (water) is modeled as a continuum with high dielectric constant. In such a case, where the dielectric constant in the system is not uniform, the Coulomb law cannot be applied. Indeed, the Coulomb law is a particular case of the more general *Poisson* equation describing the relationship between the charge density ($\rho(\mathbf{r})$), a non-uniform dielectric constant ($\varepsilon(\mathbf{r})$) and the potential ($\phi(\mathbf{r})$).

$$\nabla\varepsilon(\mathbf{r}) \cdot \nabla\phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{3.61}$$

When electrolytes are present in the medium, the Poisson equation includes an additional term resulting in the *Poisson-Boltzmann* equation:

$$\nabla\varepsilon(\mathbf{r})\cdot\nabla\phi(\mathbf{r})-\varepsilon(\mathbf{r})\lambda(\mathbf{r})\kappa^2\frac{k_bT}{q}\sinh\left[\frac{q\phi(\mathbf{r})}{k_bT}\right]=-4\pi\rho(\mathbf{r}) \qquad (3.62)$$

where q is the charge of the electrolytes, $\lambda(\mathbf{r})$ is a switching function that is zero in regions inaccessible to the ions and one otherwise, and κ is the inverse of the Debye length which depends on the ionic strength. This is a non-linear differential equation that is particularly difficult to solve. At low ionic strength, it can be simplified by using a truncated expansion of the hyperbolic sine, giving the *linearized Poisson-Boltzmann* equation:

$$\nabla\varepsilon(\mathbf{r})\cdot\nabla\phi(\mathbf{r})-\varepsilon(\mathbf{r})\lambda(\mathbf{r})\kappa^2\phi(\mathbf{r})=-4\pi\rho(\mathbf{r}) \qquad (3.63)$$

For the solution of both linear and non-linear Poisson-Boltzmann equations, several numerical methods have been developed, namely based on finite-differences [59]. The numerical solution provides the electrostatic potential of the molecule in a grid. The size of the grid must be large enough so that at the extremes the potential is zero. It is also important to compute the grid with high resolution to capture more fine details of the electrostatic potential. The atomic charges from which the electrostatic potential is calculated are taken from a force field considering the more likely protonation states at a given pH.

In the course of the simulation, the electrostatic interaction between the atomic charges of one protein and the electrostatic potential of a second protein should be calculated by numerical solution of the Poisson-Boltzmann equation at each time step. However, the computational expense of such calculation makes it unpractical for a simulation with thousands of steps. To alleviate this problem, Gabdoulline and Wade developed a method to derive *effective charges* [60] in a uniform dielectric environment that permit to speed up the calculation of the electrostatic interactions. The idea is to replace the atomic charges of the protein by a smaller number of effective charges that, in a uniform solvent dielectric, reproduce the aforementioned electrostatic potential grid already computed with the Poisson-Boltzmann equation. The effective charges are obtained then by minimizing the following functional:

$$\mathrm{J}=\int \mathrm{d}^3\mathbf{r}\left|\Phi^{(0)}(\mathbf{r})-\sum q_j^{\mathrm{eff}}\frac{\mathrm{e}^{-\mathrm{k_s}|\mathbf{r}-\mathbf{r}_j|}}{\varepsilon_\mathrm{s}|\mathbf{r}-\mathbf{r}_j|}\right| \qquad (3.64)$$

where $\Phi^{(0)}(\mathbf{r})$ is the electrostatic potential calculated by solution of the Poisson-Boltzmann equation, q^{eff} the effective charges, k_s the Debye-Hückel inverse length of the solvent with dissolved ions and ε_s the dielectric constant of the solvent. Note in the expression that the electrostatic potential exerted by the effective charges is calculated according to the Debye-Hückel model, which describes a charge immersed in a solvent with a given ionic strength and at a given temperature. The idea to speed up the BD simulation is to use the same pre-computed effective charges and electrostatic potential to calculate electrostatic interactions over all simulation steps. This obviously neglects polarization effects due to encounters between protein molecules, but it is found to provide reasonable results in comparison with more rigorous simulations recalculating the potential [55].

Van der Waals interactions

The other important interactions to consider are the Van der Waals interactions, which mediate steric effects as well as stabilizing hydrophobic and dispersion interactions. These interactions are modeled with a Lennard-Jones potential in which the depth of the energy well, ε, must be adequately parameterized. This is the only adjustable parameter of the energy model, which is usually calibrated by fitting experimental data such as the second-virial coefficient or the self-diffusion coefficient. In this regard, care must be taken when giving a physical meaning to the parameterized value ε, since the calibration will implicitly incorporate a correction for some deficiencies of the energy model in describing properly the physical property used in the calibration.

Running the simulation

In the same way as in MD, the system is simulated under periodic boundary conditions. The simulation starts with an equilibration phase in which the potential energy of the system must reach a stable value. Once the energy is stabilized the production phase can start. With the Ermak-McCammon algorithm the simulation can be easily restarted by simply knowing the center-of-mass positions of all molecules in the system and their orientations.

3.5. Calculation of Elastic Incoherent Neutron Scattering properties

Neutron scattering experiments obtain information of the atomic motions in a sample by measuring the exchange of momentum (Q) and energy (ω) of the incident neutrons in a scattering process with the nuclei [61]. As neutrons do not have electric charge they can penetrate matter deeper than electric particles and interact with the atomic nucleus via nuclear forces, which are short-ranged. What these experiments measure is the dynamic structure factor $S(Q,\omega)$. It arises from the constructive interference of neutron waves scattered from the nuclei of the sample. This function contains two types of information on the dynamics of the system: (1) coherent which results from the interference of neutrons scattered by different nuclei at different times and describe cross-correlations of atomic motions; (2) incoherent, which instead describe single-atomic motions, as arise from the interference of the neutron wave scattered by one nucleus with the wave scattered by the same nucleus after an elapsed time. The incoherent part is the dominant contribution to the scattering of the sample when hydrogen atoms are present, so herein we will focus on incoherent neutron scattering.

The dynamic structure factor gives information on the dynamics of the sample because is a Fourier transform of a time autocorrelation function of atomic positions, which is called the intermediate scattering function $F_{inc}(Q,t)$:

$$S_{inc}(Q,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F_{inc}(Q,t)e^{-i\omega t}dt \tag{3.65}$$

$$F_{inc}(Q,t) = \frac{1}{N}\sum_{\alpha} b_{\alpha,inc}^{2}\langle e^{i\mathbf{q}(\mathbf{R}(t)-\mathbf{R}(0))}\rangle \tag{3.66}$$

where $\mathbf{R}_\alpha(t)$ and $\mathbf{R}_\alpha(0)$ are position vectors of atom α, $\mathbf{q}$ the momentum transfer vector, Q the modulus of the $\mathbf{q}$ vector, $b_{\alpha,inc}$ the incoherent scattering length and N the total number of atoms. The scattering length of non-hydrogen atoms is very low compared to hydrogen atoms, so that incoherent scattering only probes the motion of hydrogens. This gives an average measure of the dynamics of the sample as hydrogen atoms are homogeneously distributed throughout a protein structure.

In the experiment, the energies of incident neutrons are distributed around an average value. This energy distribution, or *energy resolution function* (R(ω)), determines the smallest energy exchange that can be measured and depends on the neutron spectrometer. Typically R(ω) takes the form of a Gaussian, a Lorentzian or triangle function depending

on the spectrometer. Thus, the measured structure factor should be deconvoluted by the energy resolution to get the theoretical structure factor. Put it in another way, the convolution of the theoretical $S(Q,\omega)$ with the energy resolution function R(ω) is necessary in order to compare with the experiment.

$$S_{meas}(Q,\omega) = S_{inc}(Q,\omega) \otimes R(\omega) = \int_{-\infty}^{\infty} S_{inc}(Q,\omega') \cdot R(\omega-\omega')\, d\omega' \quad (3.67)$$

The width of R(ω) determines the time scale of motions accessible by the instrument, with narrower widths corresponding to longer time scales. The momentum transfer, on the other hand, defines the spatial scale of motions accessible by the instrument. Therefore, $S_{meas}(Q,\omega)$ gives information on the dynamics of the sample within a well-defined space-time window of observation. This window usually probes motions of few angstroms in time scales ranging from picoseconds to few nanoseconds.

In neutron scattering experiments, the elastic peak corresponds to the structure factor measured without exchange of energy, $S(Q,\omega=0)$. Ideally, neutrons with perfectly defined energy would give an elastic peak like a Dirac delta function. Because of the finite energy resolution, the elastic peak has the width of the energy resolution function ($\Delta\omega$) and should be expressed as $S(Q,0\pm\Delta\omega)$. When doing elastic incoherent neutron scattering experiments, the Gaussian approximation is usually employed for extracting the mean-square-displacement of the sample. The elastic intensity is Q-dependent and for confined motions takes the following form:

$$S(Q,\omega=0) = e^{-\frac{1}{6}\langle u^2\rangle Q^2} \quad (3.68)$$

where $\langle u^2\rangle$ is the mean-square-displacement (MSD) averaged over the atoms in the protein. Linearization of Eq. (3.69) by taking the slope of a natural log plot of $S(Q,\omega=0)$ *vs* Q^2 (Guinier plot) allows obtaining $\langle u^2\rangle$. The Gaussian approximation makes two important assumptions [62]. Firstly, it expands the average autocorrelation function and takes the first term:

$$\langle e^{i\mathbf{q}(\mathbf{R}(t)-\mathbf{R}(0))}\rangle = e^{-\frac{1}{2}\langle\{\mathbf{q}(\mathbf{R}(t)-\mathbf{R}(0))\}^2\rangle \pm \ldots} \quad (3.69)$$

The validity of this truncation is limited to low values of the exponent, which implies low Q-values and/or small amplitude motions, e.g. short-time dynamics, confined or harmonic motions. Secondly, the mean-square-displacement of all the atoms in the

sample is assumed to be the same, thus, overseeing that there is a distribution of mean-square-displacements in the protein (motional heterogeneity).

Comparison between simulations and experiment

The analytical relationship between the time-dependence of atomic positions and the measured structure factor makes molecular simulations and ideal tool to reproduce neutron scattering experiments. Moreover, the motions probed by neutron scattering occur in fast time scales that are easily covered by molecular dynamics.

The energy resolution function and the range of Q values used in the experiment are needed to properly analyse the simulated trajectory. Several steps require to be followed in the analysis [62]:

1. Calculation of the intermediate scattering function
2. Fourier-transformation to get the dynamic structure factor
3. Convolution of S(Q,ω) with the energy resolution function
4. Apply Gaussian approximation

The simulated trajectory must be long enough to converge the description of the atomic motions occurring at the time scale probed by the energy resolution function. Moreover, the time separation between consecutive snapshots of the trajectory must be short to describe with high resolution the atomic motions at the experimental time scale. For instance, the IN13 instrument probes motions at ~100 ps, so that a trajectory of 2 ns length saved every 1ps provides a good description of the atomic motions of interest.

For each experimental Q-value, $\mathrm{F_{inc}}(Q,t)$ is computed with Eq. (3.66) and averaged over a number of $\mathbf{q}$-vectors with random orientations and the same modulus $Q = |\mathbf{q}|$. Therefore $\mathrm{F_{inc}}(Q,t)$ must be averaged over a number of $\mathbf{q}$-vectors to guarantee an isotropic distribution of $\mathbf{q}$-vectors. At this stage there are indeed two possible ways to get the convoluted structure factor. First, as shown above, one can make a Fourier transformation of $\mathrm{F_{inc}}(Q,t)$ and subsequently a convolution of the resultant $\mathrm{S}(Q,\omega)$ with the energy resolution function. A second possibility is to take advantage of the property of the Fourier transformation that the convolution in frequency domain equals the product in time domain:

$$\mathcal{F}\{f\} \otimes \mathcal{F}\{g\} = \mathcal{F}\{f \cdot g\} \tag{3.70}$$

where $\mathcal{F}$ stands for Fourier transform, and f and g are functions in the time domain. Thus, one can make the Fourier transform of the product $\mathrm{F_{inc}}(Q,t) \cdot \mathrm{R}(t)$, where $\mathrm{R}(t)$ is

the Fourier transform of the energy resolution function $\mathrm{R}(\omega)$, to obtain the convoluted $\mathrm{S}(Q,\omega)$. Computationally this second way of calculating $\mathrm{S}(Q,\omega)$ is more efficient, as it avoids evaluating the convolution explicitly, and thus is the usual method of choice. Finally, by taking the value of $\mathrm{S}(Q,0)$ in the studied Q-range the Gaussian approximation can be used to obtain the mean-square-displacement $\langle u^2 \rangle$. It is important to underscore the difference between the $\langle u^2 \rangle$ obtained in this way, by calculation of neutron scattering properties, and the theoretical definition of the mean-square-displacement, which is written as:

$$\mathrm{MSD}(\Delta t) = \frac{1}{N(T-\Delta t+1)} \sum_{i}^{N} \sum_{t=0}^{T-\Delta t} [R_i(t+\Delta t) - R_i(t)]^2 \tag{3.71}$$

where MSD is averaged over N atoms and the number of frames separated by Δt.

Bibliography

Cramer CJ (2002) Essentials of Computational Chemistry: Theories and Models. West Sussex: Wiley

Cui Q, Bahar I, editors (2006) Normal Mode Analysis: Theory and applications to biological and chemical systems. Boca Raton: Chapman & Hall, CRC Press, Taylor & Francis Group.

Field MJ (2007) A Practical Introduction to the Simulation of Molecular Systems. New York: Cambridge University Press.

Jensen F (2001) Introduction to Computational Chemistry. West Sussex: Wiley

Leach AR (1996) Molecular Modeling: Principles and Applications. Essex: Addison Wesley Longman.

Schlick T (2010) Molecular Modeling and Simulation: An Interdisciplinary Guide. New York: Springer.

Serdyuk IN, Zaccai NR, Zaccai G (2007) Methods in Molecular Biophysics: Structure, Dynamics and Function. New York: Cambridge University Press.

References

1. Grimme S (2003) Improved second-order Moller-Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies. *J Chem Phys* 118: 9095-9102.
2. Raghavachari K, Trucks GW, Pople JA, Head-Gordon M (1989) A 5th-order perturbation comparison of electron correlation theories. *Chem Phys Lett* 157: 479-483.
3. Pople JA, Santry DP, Segal GA (1965) Approximate self-consistent molecular orbital theory. I. Invariant Procedures. *J Chem Phys* 43: S129-S135.
4. Pople JA, Segal GA (1965) Approximate self-consistent molecular orbital theory. 2. Calculations with complete neglect of differential overlap. *J Chem Phys* 43: S136-S151.
5. Pople JA, Beveridge.D. l, Dobosh PA (1967) Approximate self-consistent molecular-orbital theory. 5. Intermediate neglect of differential overlap. *J Chem Phys* 47: 2026-2033.

6. Dewar MJS, Thiel W (1977) Ground states of molecules. 38. MNDO method - approximations and parameters. *J Am Chem Soc* 99: 4899-4907.
7. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) The development and use of quantum-mechanical molecular models. 76. AM1-a new general-purpose quantum-mechanical molecular model. *J Am Chem Soc* 107: 3902-3909.
8. Stewart JJP (1989) Optimization of parameters for semiempirical methods. 1. Method. *J Comput Chem* 10: 209-220.
9. Rocha GB, Freire RO, Simas AM, Stewart JJP (2006) RM1: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J Comput Chem* 27: 1101-1111.
10. Thiel W, Voityuk AA (1996) Extension of MNDO to d orbitals: Parameters and results for the second-row elements and for the zinc group. *J Phys Chem* 100: 616-626.
11. Winget P, Horn AHC, Selcuki C, Martin B, Clark T (2003) AM1*parameters for phosphorus, sulfur and chlorine. *J Mol Model* 9: 408-414.
12. Nam K, Cui Q, Gao JL, York DM (2007) Specific reaction parametrization of the AM1/d Hamiltonian for phosphoryl transfer reactions: H, O, and P atoms. *J Chem Theory Comput* 3: 486-504.
13. Arantes GM, Loos M (2006) Specific parametrisation of a hybrid potential to simulate reactions in phosphatases. *Phys Chem Chem Phys* 8: 347-353.
14. Stewart JJP (2007) Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J Mol Model* 13: 1173-1213.
15. Hohenberg P, Kohn W (1964) Inhomogeneous electron gas. *Phys Rev B* 136: B864-B871.
16. Kohn W, Sham LJ (1965) Self-consistent equations including exchange and correlation effects. *Phys Rev* 140: A1133-A1138.
17. Vosko SH, Wilk L, Nusair M (1980) Accurate spin-dependent electron liquid correlation energies for local spin-density calculations - a critical analysis. *Can J Phys* 58: 1200-1211.
18. Becke AD (1988) Density functional exchange energy approximation with correct asymptotic behavior. *Phys Rev A* 38: 3098-3100.
19. Lee CT, Yang WT, Parr RG (1988) Development of the Colle-Salvetti correlation energy formula into a functional of the electron density. *Phys Rev B* 37: 785-789.
20. Perdew JP, Wang Y (1992) Accurate and simple analytic representation of the electron gas correlation energy. *Phys Rev B* 45: 13244-13249.
21. Lynch BJ, Fast PL, Harris M, Truhlar DG (2000) Adiabatic connection for kinetics. *J Phys Chem A* 104: 4811-4815.

22. Kormos BL, Cramer CJ (2002) Adiabatic connection method for X-+RX nucleophilic substitution reactions (X = F, Cl). *J Phys Org Chem* 15: 712-720.
23. Adamo C, Barone V (1998) Exchange functionals with improved long-range behavior and adiabatic connection methods without adjustable parameters: The mPW and mPW1PW models. *J Chem Phys* 108: 664-675.
24. Zhao Y, Truhlar DG (2008) Density functionals with broad applicability in chemistry. *Acc Chem Res* 41: 157-167.
25. Grimme S (2006) Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J Comput Chem* 27: 1787-1799.
26. Bader RFW (1991) A quantum-theory of molecular structure and its applications. *Chem Rev* 91: 893-928.
27. Reed AE, Curtiss LA, Weinhold F (1988) Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chem Rev* 88: 899-926.
28. Weiner SJ, Kollman PA, Nguyen DT, Case DA (1986) An all-atomic force-field for simulations of proteins and nucleic-acids. *J Comput Chem* 7: 230-252.
29. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM - A program for macromolecular energy, minimization and dynamics calculations. *J Comput Chem* 4: 187-217.
30. Scott WRP, Hunenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, van Gunsteren WF (1999) The GROMOS biomolecular simulation program package. *J Phys Chem A* 103: 3596-3607.
31. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* 118: 11225-11236.
32. Berendsen HJC, Postma JPM, Van Gunsteren WF, Hermans J (1981) Interaction models for water in relation to protein hydration. In: Pullman B, editor. Intermolecular Forces. Reidel, Dordrecht. pp. 331-342.
33. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* 79: 926-935.
34. Mahoney MW, Jorgensen WL (2000) A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J Chem Phys* 112: 8910-8922.
35. York DM, Darden TA, Pedersen LG (1993) The effect of long-range electrostatic interactions in simulations of macromolecular crystals - a comparison of the Ewald and truncated list methods. *J Chem Phys* 99: 8345-8348.
36. Nam K, Gao JL, York DM (2005) An efficient linear-scaling Ewald method for long-range electrostatic interactions in combined QM/MM calculations. *J Chem Theory Comput* 1: 2-13.

37. Warshel A, Levitt M (1976) Theoretical studies of enzymic reactions - dielectric, electrostatic and steric stabilization of carbonium-ion in reaction of lysozyme. *J Mol Biol* 103: 227-249.
38. Gao JL, Amara P, Alhambra C, Field MJ (1998) A generalized hybrid orbital (GHO) method for the treatment of boundary atoms in combined QM/MM calculations. *J Phys Chem A* 102: 4714-4721.
39. Gonzalez C, Schlegel HB (1991) Improved algorithms for reaction-path following - higher-order implicit algorithms. *J Chem Phys* 95: 5853-5860.
40. Weinan E, Ren WQ, Vanden-Eijnden E (2002) String method for the study of rare events. *Phys Rev B* 66: 052301.
41. Jónsson H, Mills G, Jacobsen KW (1998) Nudged elastic band method for finding minimum energy paths of transitions; Berne BJ, Ciccotti G, Coker DF, editors. Singapore: World Scientific. 385-404 p.
42. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of cartesian equations of motion of a system with constraints - molecular dynamics of n-alkanes. *J Comput Phys* 23: 327-341.
43. Hess B, Bekker H, Berendsen HJC, Fraaije J (1997) LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* 18: 1463-1472.
44. Berendsen HJC, Postma JPM, Van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81: 3684-3690.
45. Nose S (1984) A molecular dynamics method for simulations in the canonical ensemble. *Mol Phys* 52: 255-268.
46. Hoover WG (1985) Canonical dynamics - equilibrium phase-space distributions. *Phys Rev A* 31: 1695-1697.
47. Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. *Proteins: Struct, Funct, Genet* 17: 412-425.
48. Hinsen K, Petrescu AJ, Dellerue S, Bellissent-Funel MC, Kneller GR (2000) Harmonicity in slow protein dynamics. *Chem Phys* 261: 25-37.
49. Orellana L, Rueda M, Ferrer-Costa C, Lopez-Blanco JR, Chacon P, Orozco M (2010) Approaching Elastic Network Models to Molecular Dynamics Flexibility. *J Chem Theory Comput* 6: 2910-2923.
50. Tama F, Gadea FX, Marques O, Sanejouand YH (2000) Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Struct, Funct, Genet* 41: 1-7.
51. Li GH, Cui Q (2002) A coarse-grained normal mode approach for macromolecules: An efficient implementation and application to Ca2+-ATPase. *Biophys J* 83: 2457-2474.
52. Chennubhotla C, Bahar I (2006) Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. *Mol Syst Biol* 2: 36.

53. Chennubhotla C, Bahar I (2007) Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS Comput Biol* 3: e172.
54. Dijkstra EW (1959) A note on two problems in connexion with graphs. *Num Math* 1: 269-271.
55. Elcock AH (2004) Molecular simulations of diffusion and association in multimacromolecular systems. Numerical Computer Methods, Pt D. pp. 166-198.
56. McGuffee SR, Elcock AH (2006) Atomically detailed simulations of concentrated protein solutions: The effects of salt, pH, point mutations, and protein concentration in simulations of 1000-molecule systems. *J Am Chem Soc* 128: 12098-12110.
57. McGuffee SR, Elcock AH (2010) Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Comput Biol* 6: e1000694.
58. Ermak DL, McCammon JA (1978) Brownian dynamics within hydrodynamic interactions. *J Chem Phys* 69: 1352-1360.
59. Baker NA (2004) Poisson-Boltzmann methods for biomolecular electrostatics. *Numerical Computer Methods, Pt D* 383: 94-118.
60. Gabdoulline RR, Wade RC (1996) Effective charges for macromolecules in solvent. *J Phys Chem* 100: 3868-3878.
61. Gabel F, Bicout D, Lehnert U, Tehei M, Weik M, Zaccai G (2002) Protein dynamics studied by neutron scattering. *Q Rev Biophys* 35: 327-367.
62. Hayward JA, Smith JC (2002) Temperature dependence of protein dynamics: Computer simulation analysis of neutron scattering properties. *Biophys J* 82: 1216-1225.

CHAPTER 4

RESULTS

The results presented in this thesis are divided into three parts that are devoted to different aspects of the enzymatic function. The first part focuses on reactivity in the context of phosphoryl transfer reactions. The second part concentrates on dynamical properties associated to substrate binding and allosteric regulation of the activity. The third, and last part, is concerned on the thermal stability of enzymes and, in particular, on the dynamical properties of two enzymes with different thermal behavior. A brief summary of each study is presented and followed by the corresponding publication.

4.1. Phosphoryl transfer reactions

Rationalizing and predicting the reaction mechanism of phosphoryl transfer reactions in enzymes requires to understand the factors that determine the stability of the pentacoordinated phosphorus species that can be formed in the course of the reaction. In this section, we present our results on phosphorane model systems, the methodology required for their description and the application of this knowledge to characterize the controversial mechanism of the β-phosphoglucomutase enzyme.

4.1.1. Pentacoordinated Phosphorus: structure, reactivity and biological implications

By systematically studying small phosphorane compounds (Figure 1), we first explored how the electrodonor character of both equatorial and apical groups determines the mechanism of a phophoryl transfer reaction as well as the geometry of pentacoordinated intermediates. We have made special emphasis on the apical bond throughout this study as it determines the direction of the reaction, involving the nucleophilic and leaving groups. Therefore we have characterized the apical bond by wave function analysis methods.

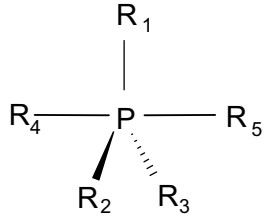


Figure 1. Structure of a pentacoordinated phosphorus. R_1, R_2 and R_3 are termed equatorial groups, and R_4 and R_5 correspond to the apical groups (nucleophile and leaving group).

From this systematic study, we observed two trends: (1) the more electrodonor the character of the equatorial group the longer the apical bond distance ; (2) the more electrodonor character of the apical group the shorter the apical bond distance. The observed variations in the apical bond distances are within 0.2 Å, which illustrates the sensitivity of the apical bond to inductive effects. Interestingly, when the two apical groups are different, the group with stronger electrodonor character shortens the bond distance at the expense of lengthening the bond distance of the other apical group with weaker electrodonor character. In general, we found that important differences in the

character of the two apical groups tend to destabilize the pentacoordinated phosphorus species.

All of the above already shows that the apical bond is highly polarizable, but it was the study of the compound shown in Figure 1A that revealed the extent to which the apical bond can be polarized.

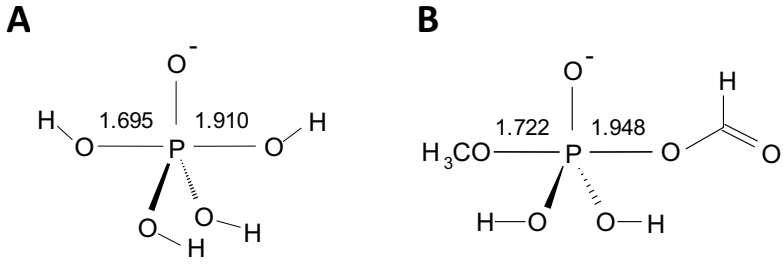


Figure 2. Structure of phosphorane models with strong polarization. (A) the source of polarization arises from the orientation of equatorial groups, whereas in (B) the polarization is induced by the different character of the apical groups

The asymmetrical orientation of the equatorial OH groups polarizes the apical bond lengthening the bond distance in 0.15 Å, with respect to the symmetrical compound. This emphasizes the ionic character of the apical bond. The natural implication of this effect is that anisotropic electrostatic effects from the environment can easily polarize the molecule and thus alter its geometry and stability. To explore this idea, we studied the effect of electric fields applied in the apical direction of a set of pentacoordinated phosphorus compounds. Interestingly, electric fields applied in the apical direction notoriously change the apical bond distance. Depending on the magnitude and direction of this electric field those compounds with low stability can dissociate or, on the contrary, become more stable. Such sensitivity of putative pentacoordinated intermediates in a phosphoryl transfer implies that the reaction mechanism can be strongly influenced by external electric fields. As an example, we have studied how the reaction mechanism yielding the compound shown in Figure 2B changes with the electric field (Figure 3). Without electric field, the reaction proceeds through a step-wise mechanism. The pentacoordinated intermediate is kinetically very unstable as is markedly asymmetric, being the dissociation energy barrier of the weaker apical bond 16 kcal/mol lower than that of the stronger one. By applying an electric field in one direction, the intermediate is not stable anymore, so that the lower barrier disappears turning the mechanism into a concerted process. An electric field in the opposite direction, instead, strengthens the weaker bond to the extent that the dissociation barrier increases 9 kcal/mol, with respect to the intermediate species, at the expense of reducing the dissociation barrier of the

stronger bond in 8 kcal/mol. On balance, the kinetic stability of the intermediate, in this case, is enhanced by the electric field.

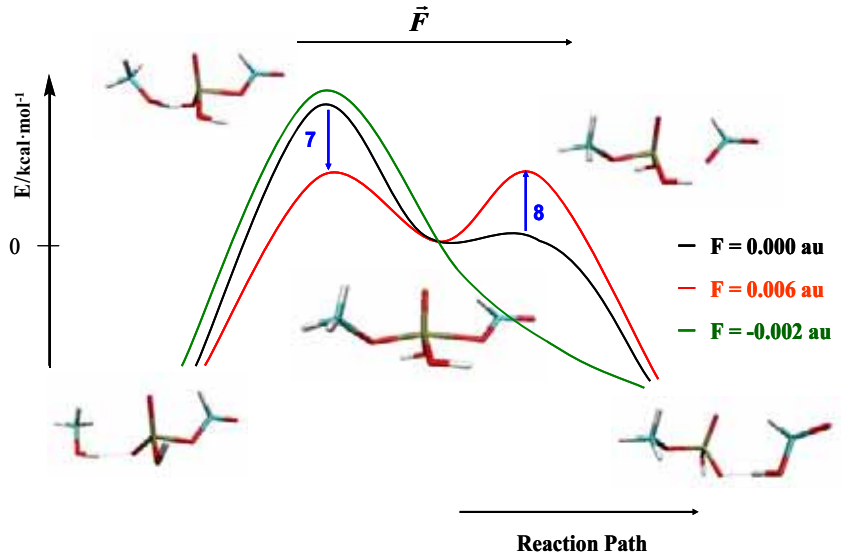


Figure 3. Scheme of the reaction profile calculated in the presence of electric fields of different magnitude applied to the apical direction.

This study illustrates the polarizable character of the apical bond in pentacoordinated phosphorus compounds. Furthermore, the notable polarizability of this bond is manifested in the capacity of external electric fields in modifying the geometry and reactivity of pentacoordinated phosphorus. The sensitivity of these species to electric fields can have important consequences in reactions catalyzed by enzymes where the charge distribution of the active site is very heterogeneous leading to electrostatic fields important for catalysis.

A detailed presentation of the results and methodologies used in this study can be found in the article: *Inductive and External Electric Field Effects in Pentacoordinated Phosphorus Compounds*. (2008) J. Chem. Theory Comput. 4:49-63

J. Chem. Theory Comput. **2008,** *4,* 49−63 **49**

JCTC Journal of Chemical Theory and Computation

Inductive and External Electric Field Effects in Pentacoordinated Phosphorus Compounds

Enrique Marcos, Ramon Crehuet,* and Josep M. Anglada*

Grup de Química Teòrica i Computacional, Departament de Química Orgànica Biològica, Institut d'Investigacions Químiques i Ambientals de Barcelona, IIQAB − CSIC, c/ Jordi Girona 18, E-08034 Barcelona, Spain

Received August 28, 2007

Abstract: Pentacoordination at phosphorus is associated with a nucleophilic displacement reaction at tetracoordinated phosphorus compounds and shows a great variability in what respects their geometrical and energetic features. By means of a systematic theoretical study on a series of elementary model compounds, we have analyzed the bonding features. The pentacoordinated phosphorus compounds are held together by dative bonds, and the geometry and stability depends on the inductive effects originated by different substitutes at phosphorus. We show also that an external electric field can modify the geometrical features and the reactivity of the nucleophilic substitution reactions. This issue may have great interest in biological reactions involving pentacoordinated phosphorus where the electric field originated by the folded protein could influence the catalytic process. We report also additional calculations on the geometry and NMR spectra on three triphenyl phosphonium ylide derivatives, and our results compare well with the experimental data.

Introduction

A detailed knowledge on the electronic nature of pentacoordination at phosphorus is of great interest in chemistry and biochemistry.[1−12] Pentacoordination at phosphorus is mainly associated with a nucleophilic displacement reaction at tetracoordinated phosphorus compounds, which is associated with cell signaling and energetics and many aspects of biosynthesis. These nucleophilic reactions occur in the so-called associative processes, which can follow a concerted pathway, with a trigonal bipyramid transition state, or an addition−elimination pathway, involving a pentacoordinated phosphorane intermediate.[7−9,11,13,14] These processes are important in chiral reactions. Those following a concerted pathway take place with inversion of configuration, but in pathways involving pentacoordinate intermediates, a Berry pseudorotation may occur, which could involve retention of configuration.[5,11] Pentacoordinated phosphorus intermediates are found, for instance, in the Wittig reaction,[15] in human α-thrombin inhibitors,[1] and as intermediates in the hydrolysis of phospholipids catalyzed by phospholipase D.[6] It may exist in phosphoryl transfer in GTP hydrolysis by RAS proteins[10] and as an intermediate in the phosphoryl transfer reaction catalyzed by a β-phosphoglucomutase,[2,16] although some controversy exists in the literature regarding the true nature of this intermediate.[3,4]

Pentacoordination at phosphorus occurs mainly in trigonal bipyramid structures, and it has been observed that the apical bond lengths show a great variability, which depends on several factors as the nature of the substitutes at P, the influence of hydrogen bonding or the charge around phosphorus.[1,10,11,17,18] It appears therefore that such variability would affect not only the stability of these compounds but also the transition states involving pentacoordination at phosphorus and consequently the reactivity. The factors affecting this variability are crucial for a complete understanding of nucleophilic displacement at phosphorus. They are still not well rationalized and are the main goal of this study.

Extensive theoretical studies have also been reported in the literature, which have provided valuable information regarding different aspects of the reaction mechanisms of

* Corresponding author e-mail: anglada@iiqab.csic.es (J.M.A.), resqtc@iiqab.csic.es (R.C.).

10.1021/ct700220z CCC: $40.75
Published on Web 12/05/2007

phosphate reactions, the importance and possible existence of pentacoordinated intermediates depending on the reaction conditions, and the effects of the solvent in the reactivity.[19−47] In this study we have focused our attention on the inductive effects affecting pentacoordination at phosphorus and its bonding features. To this end, we have considered, in the first stage, a series of model systems for which we have investigated the effect of different substitutes at phosphorus as well as the effect of polarization and the effect of an external electric field. In the second stage, we have also investigated a series of triphenylphosphonium ylide derivatives for which experimental data exist in the literature.

Computational Details

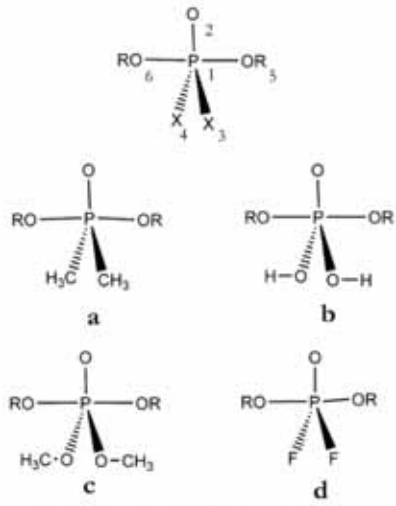
All geometry optimizations carried out in this work have been performed with the density functional *m*PW1PW91[48] employing the 6-31+G(d) basis set.[49] At this level of theory we have also calculated the harmonic vibrational frequencies to verify the nature of the corresponding stationary point (minima or transition state) and to provide the zero point vibrational energy (ZPE). The *m*PW1PW91 functional has been found to be adequate to describe systems with long-range interactions, especially with dative bonds.[50] Moreover, the reliability of this functional, with respect to the geometrical parameters, has also been checked by performing, for some test models, comprehensive test calculations employing the MP2[51−53] ab initio approach and using the 6-31+G(d), 6-311+G(d), and 6-311+G(3df, 3pd) basis sets. The results obtained compare quite well and are collected in the Supporting Information. Moreover, for the cases where reactivity has been considered, we have performed, for each transition state, intrinsic reaction coordinate calculations (IRC)[54−56] in order to ensure that the transition states connect the desired reactants and products. In the second step, the relative energies of the stationary points were corrected by performing single point energy calculations using the *m*PW1PW91 functional with the 6-311+G(3df,2p) basis set.[57] In addition, we have also checked the reliability of the activation and reaction energies by performing, for all stationary points of a given reaction (reaction **1b**, see below), additional single point energy calculations at the higher level of theory CCSD(T)/IB.[58−62] The results obtained at the *m*PW1PW91 and CCSD(T) level compare very well and are contained in the Supporting Information.

For the three triphenylphosphonium ylides considered, we have also computed the NMR spectra by performing B3LYP single-point calculations[63] at the optimized geometries, using the GIO method[64,65] and employing the 6-311+G(2d,p) basis set.[57]

The quantum chemical calculations carried out in this work were performed by using the Gaussian[66] program package, and the Molden program[67] was employed to visualize the geometric and electronic features.

The bonding features of the different systems considered were analyzed by employing the natural bond orbital (NBO) partition scheme by Weinhold and co-workers[68] and the atoms in molecules (AIM) theory by Bader.[69] The topological properties of wave functions were computed using the AIMPAC program package.[70]

Scheme 1 [a]



[a] RO = HO (**1**); CH_3O (**2**); HCOO (**3**); CF_3O (**4**).

Results and Discussion

The Model Systems $POX_2(RO)_2$ Pentacoordinated Compounds. One important point regarding the chemistry of pentacoordinated phosphorus compounds refers to the variability of the apical bond distances.[5,11] In order to analyze and rationalize this issue we have carried out a series of calculations on the $POX_2(RO)_2$ model systems. These pentacoordinated model systems have been depicted in Scheme 1 and possess a trigonal bipyramid structure. Here X are equatorial substitutes (X = CH_3 (**a**); HO (**b**); CH_3O (**c**); and F (**d**)) and RO are apical substitutes (RO = HO (**1**); CH_3O (**2**); HCOO (**3**); and CF_3O (**4**)). Along this work, the different models are labeled by a number as a prefix, according to the apical substitutes, followed by a letter as a suffix according to the equatorial substitutes. Thus, compound **1a** corresponds to $PO(CH_3)_2(OH)_2$, whereas compound **3c** corresponds to $PO(CH_3O)_2(HCOO)_2$ (see Scheme 1).

Please note also that in these model systems the charge of the system is −1, the two apical substitutes are identical, and that in **b** and **c** the two equatorial substitutes are oppositely oriented. Moreover, it is also worth reminding the reader that the donor character of the apical substitutes is HO > CH_3O > HCOO > CF_3O and the donor character of the equatorial substitutes is CH_3 > HO > CH_3O > F so that these series of model systems allow us to analyze combinations of electron withdrawing groups and electron donor groups on the phosphorus coordination. The most significant geometrical parameters of the optimized structures are displayed in Table 1, which also includes the tetracoordinated phosphoric acid H_3PO_4 for comparison. Figure 1 shows the dependence of the apical bond lengths with respect to the apical and equatorial substitutes at P. The Cartesian coordinates of each pentacoordinated model system are reported in the Supporting Information.

For H_3PO_4, Table 1 shows that our calculations predict the P···O bond length to be 1.476 Å and the three P···OH

Table 1. Optimized Bond Lengths (in Å) to Phosphorus in H_3PO_4 and the Pentacoordinated **1a**−**4d** Model Compounds

	substitutes				
compd	apical	equatorial	r(P−OR_{apical})	r(P−$X_{equatorial}$)	r(P−O)
H_3PO_4				1.602	1.476
1a	OH	CH_3	1.814	1.849	1.545
1b	OH	OH	1.768	1.660	1.527
1c	OH	OCH_3	1.742	1.673	1.539
2a	OCH_3	CH_3	1.841	1.841	1.521
2b	OCH_3	OH	1.778	1.660	1.510
2c	OCH_3	CH_3	1.749	1.677	1.514
3a	OC(H)O	CH_3	1.908	1.832	1.507
3b	OC(H)O	OH	1.835	1.635	1.495
3c	OC(H)O	OCH_3	1.818	1.639	1.499
3d	OC(H)O	F	1.780	1.609	1.496
4a	OCF_3	CH_3	1.980	1.826	1.488
4b	OCF_3	OH	1.878	1.625	1.480
4c	OCF_3	OCH_3	1.882	1.616	1.484
4d	OCF_3	F	1.808	1.591	1.476

bond lengths to be 1.602 Å. The nucleophilic addition of the HO anion to H_3PO_4 leads to the pentacoordinated compound **1b** and produces a lengthening of 0.051 Å in the P···O bond and of 0.058 Å in the equatorial P···OH bonds, compared with the P···O and the P···OH bond lengths in phosphoric acid, but the two apical P···OH bond distances (1.768 Å) are predicted to be much longer (see Table 1).

Regarding the remaining pentacoordinated model systems, the results of Table 1 and Figure 1 show a great variability of the P···OR_{apical} bond length, which depends on the character of both X and RO. Thus, the apical bond distance changes as much as 0.238 Å, from 1.742 Å in **1c** to 1.980 Å in **4a**, whereas the changes in the equatorial bond lengths (P···$X_{equatorial}$ and P···$O_{equatorial}$) are smaller than 0.070 Å for all the model compounds. Table 1 and Figure 1 also show that, for the same equatorial substitute, the P···OR_{apical} bond length is shorter as the donor character of OR increases, while the donor character of the equatorial substitute X results in an increase of the P···OR_{apical} bond length. Thus, for instance, for X = CH_3, the P···O_{apical} bond distance changes from 1.814 Å in **1a** (apical substitute = OH) to 1.980 Å in **4a** (apical substitute = OCF_3), while for X = CH_3O, the P···O_{apical} bond distance changes from 1.742 Å in **1c** (apical substitute = OH) to 1.883 Å in **4c** (apical substitute = OCF_3) (see Table 1 and Figure 1). In the case whether the equatorial substitute X is F, we have only found pentacoordinated compounds with apical substitutes HCOO (**3d**) and CF_3O (**4d**).

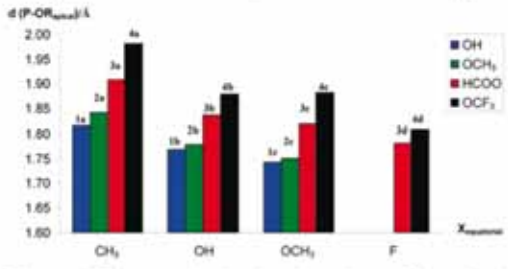


Figure 1. Diagram showing the dependence of the apical bond lengths in the pentacoordinated $POX_2(RO)_2$ model compounds on the nature of the apical (RO) and equatorial (X) substitutes on P.

Table 2. Natural Occupation at Phosphorus, Stabilization Energies ($\Delta E(2)$ in kcal·mol^{-1}) Associated with the Most Important Donor−Acceptor Interaction Involving the Apical Bonds and the Natural Charges at Phosphorus (Q in e)[d]

	P natural occupation			stabilization energies			
compd	s	p	d	$\sigma_{P1O5} \rightarrow \sigma^*_{P1O6}$ [a]	$\sigma_{P1X} \rightarrow \sigma^*_{P1O6}$ [b]	$\sigma_{P1O2} \rightarrow \sigma^*_{P1O6}$ [c]	Q_{NBO}(P)
1a	0.91	1.85	0.09	33.51	30.67	28.88	2.12
1b	0.77	1.59	0.11	29.80	22.66	21.59	2.51
1c	0.77	1.56	0.11	28.86	20.02	23.54	2.54
2a	0.91	1.81	0.08	31.86	39.14	27.90	2.17
2b	0.77	1.55	0.10	25.50	30.71	20.28	2.56
2c	0.76	1.51	0.10	24.27	24.41	16.98	2.59
3a	0.94	1.82	0.08	35.52	35.23	31.58	2.13
3b	0.77	1.57	0.10	29.04	24.04	23.34	2.53
3c	0.77	1.53	0.10	28.80	24.08	22.12	2.58
3d	0.76	1.49	0.11	24.88	20.59	24.90	2.62
4a	0.95	1.82	0.07	35.94	38.97	34.06	2.12
4b	0.77	1.57	0.10	29.43	24.43	26.59	2.53
4c	0.77	1.53	0.10	32.39	25.32	27.89	2.58
4d	0.75	1.50	0.11	28.48	22.90	28.43	2.62

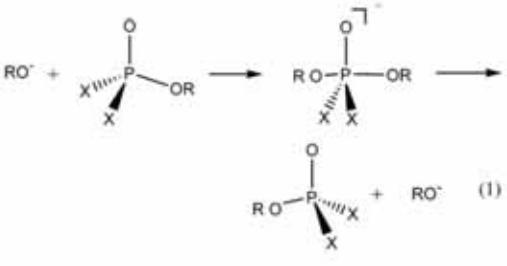
[a] $\sigma_{P1O5} \rightarrow \sigma^*_{P1O6}$ has the same value as $\sigma_{P1O6} \rightarrow \sigma^*_{P1O5}$. [b] The same interaction occurs from each σ_{PX} equatorial to each σ^*_{PO} apical bond. [c] $\sigma_{P1O2} \rightarrow \sigma^*_{P1O6}$ has the same value as $\sigma_{P1O2} \rightarrow \sigma^*_{P1O5}$. [d] Bond numbering is according Scheme 1.

These results indicate two important points, namely a different nature of the apical and equatorial bonds and a great importance of inductive effects on pentacoordinated phosphorus compounds. In order to get a deeper knowledge of the features of bonding at phosphorus, we have carried out a study of the bond properties according to the Atoms in Molecules (AIM) theory by Bader and the Natural Bond Orbital (NBO) theory by Weinhold. A detailed discussion of the AIM analysis is given in the Supporting Information along with the computed topological parameters at the bcp of the P···O_{apical}, the P···$X_{equatorial}$, and the P···O bonds collected in Table S4. In general, and regarding the apical bonds, the range of values for ρ_b and $\nabla^2\rho_b$ are characteristic of "closed shell" interactions, so that *the two apical bonds in these model systems can be classified as dative*. The large variability of the P···O apical bonds pointed out above, depending on the nature of the substitutes, is also typical of dative interactions.[71,72]

Additional information is provided by the NBO analysis. In Table 2 we have displayed the natural occupation at the P atom, the natural charge on P, and the stabilization energies ($\Delta E(2)$) associated with the charge-transfer interactions of the relevant donor−acceptor orbitals involving the apical bonds, that is, the bonding σ(P−O_{apical}), σ(P−$X_{equatorial}$), and σ(P−O) NBOs with the antibonding acceptor σ^*(P−O_{apical}) NBO. This stabilization energy has been computed with the second-order perturbation theory with the Fock matrix in the NBO analysis and the natural charges on phosphorus. The NBO analysis indicates that the natural occupation in the d

shell is always less than or equal to 0.11 and consequently *excludes the participation of the d-orbital in the hybridization picture*. Thus, there is a formal sp^2 hybridization at P in all pentacoordinated model compounds. The d orbitals act as polarization functions in a similar way as pointed out by Reed and co-workers in a study on chemical bonding in hypervalent molecules[73] and in pentacoordinated silicon compounds bonded also by dative bonds.[72] This formal hybridization scheme is also compatible with the simple MO diagram based on a three-center four-electron (3c4e) model.[11,74] Another important point to be mentioned here refers to the topological features of the NBO orbitals linked to phosphorus. Those NBO orbitals designed as bonding orbitals of the type P···O or P···F in Table 2 are highly polarized toward the O or F atom, whereas those designed as antibonding orbitals have an almost exclusive contribution of phosphorus. Thus, the donor−acceptor interaction between these NBOs displayed in Table 2 represents quite well charge-transfer interactions. Moreover, this topological picture agrees very well with the dative description of the P···O bonds provided by the AIM analysis and discussed above. By the same way, the P···C bonds in compounds **1a**, **2a**, **3a**, and **4a** (with CH_3 as equatorial substitutes) have an almost equal contribution of phosphorus and carbon, according to the covalent character predicted by the AIM analysis (see above). The most important perturbative donor−acceptor interactions involving the equatorial substitutes ($\sigma_{PX-equatorial} \rightarrow \sigma^*_{PO-apical}$) are those having X = CH_3 (compounds **1a**, **2a**, **3a**, and **4a**) according to the well-known donor character of the methyl substitute and decreases according to the donor character of the equatorial substitutes (see above). Also very interesting are the perturbative donor−acceptor interactions between the two apical bonds ($\sigma_{P1O5} \rightarrow \sigma^*_{P1O6}$) and ($\sigma_{P1O6} \rightarrow \sigma^*_{P1O5}$) that involve charge transfer between the two apical bonds. Here it is also worth pointing out that these apical donor−acceptor interactions are symmetrical because the two apical groups are the same (see footnote b of Table 2). However, as will be shown below, when the two apical groups are different, the two apical donor−acceptor interactions are different, pointing out the competition of these two groups to form a dative bond to phosphorus and therefore having a direct influence on the corresponding P···O bonds.

Nucleophilic Substitution on the Model $POX_2(OR)_2$ Pentacoordinated Compounds. A very important point concerning the pentacoordinated $POX_2(OR)_2$ compounds discussed above refers to their relative stability. This has been studied in connection with the formation via a S_N2 reaction according to eq 1.



(1)

Please note that in this section, all reactions considered are symmetric as the entrance and leaving groups are identical. Each reaction described by eq 1 has been named according to the substitutes in the same way as has been done in the previous section to characterize the pentacoordinated phosphorus model compounds as displayed in Scheme 1. Thus, for instance, reaction **1a** means RO^- = HO^- and X = CH_3, or reaction **4c** means RO = CF_3O^- and X = OCH_3; that is, each reaction has the same name that identifies the pentacoordinated intermediate. A schematic representation of the corresponding potential energy surfaces has been drawn in Figures 2 and 3, whereas the geometric parameters of the corresponding stationary points are collected in the Supporting Information. The energetic of these processes is contained in Table 3.

Figure 2a shows a schematic potential energy profile of reactions **1a**−**4a**, having the CH_3 group as equatorial substitute X. Each reaction begins with the formation of a prereactive hydrogen-bonded complex which occurs previous to the transition state and the formation of the pentacoordinate intermediate. Every prereactive complex has two hydrogen bonds, which occur between the oxygen of the anion (RO^-) and one of the hydrogen atoms of each equatorial methyl substitute. For reaction **3a** (red line, having HCOO as apical substitutes), the two hydrogen bonds in the prereactive complex are formed between each one of the oxygen atoms of the HCOO anion and one of the hydrogen atoms of each equatorial methyl substitute. The stability of these hydrogen-bonded complexes at 0 K is computed to vary among 24.3 and 16.6 kcal·mol^{-1} (for reactions **1a**−**4a**, see Table 3), and these energy values in gas phase are typical of hydrogen bond interactions involving an anion. After surmounting an energy barrier of the order of 5−6 kcal·mol^{-1}, the corresponding pentacoordinate intermediate is formed, and its stability, at 0 K, is computed to be among 33.1 and 15.9 kcal·mol^{-1}, relative to RO^- plus $POX_2(OR)$. The stability in these intermediates depends on the donor character of the apical substitutes. There is a large difference in the relative stability of the **1a** (blue line, apical substitute HO) and the stability of **4a** (black line, with apical substitute CF_3O), which amounts 16 kcal·mol^{-1}, so that *the compounds having the apical substitute with higher donor character are more stable.* This higher stability is associated with shorter apical bond lengths as discussed in the previous section for compounds **1a**, **2a**, **3a**, and **4a**.

In the case of the two equatorial (Figure 2b) substitutes X is the F atom, and we have only considered the reaction with RO^- = $HCOO^-$ (**3d**) and RO^- = CF_3O^- (**4d**), since these are the only ones in which the F substitutes remain in the equatorial position as pointed out in the previous section. Both reactions occur by direct formation of a pentacoordinated phosphorus compound, whose stability at 0 K has been computed to be 32.1 and 20.7 kcal·mol^{-1}, for **3d** (red line) and **4d** (black line), respectively, according also to the higher donor character of the apical substitute in **3d** (see also Table 3). The processes are similar to those described recently by van Bochove and co-workers[24] in a recent study on nucleophilic substitution at phosphorus having fluorine atoms as equatorial substitute.

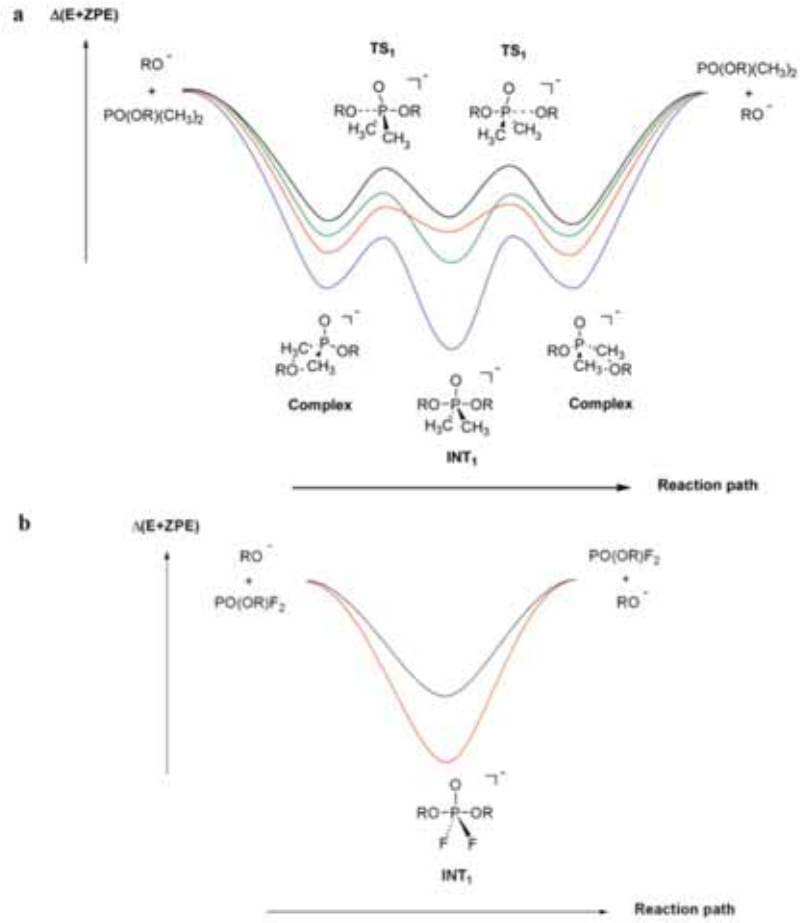


Figure 2. Schematic potential energy diagram for the nucleophilic substitution reactions: In (a), $RO^- + PO(OR)(CH_3)_2 \rightarrow PO(OR)(CH_3)_2 + RO^-$; (RO = HO, blue line; CH_3O, green line; HCOO, red line; and CF_3O, black line). In (b), $RO^- + PO(OR)(F)_2 \rightarrow PO(OR)(F)_2 + RO^-$; (RO = HCOO, red line; CF_3O, black line). The relative energies are computed at the *m*PW1PW91/6-311+G(3df,2p)//*m*PW1PW91/6-311+G(d) level of theory.

For the equatorial substitute X = OH, namely reactions **1b**−**4b**, the precursors of the corresponding pentacoordinated phosphorus compound are not RO^- and $PO(OH)_2OR$ as described in eq 1, but its respective conjugate acid (ROH) and basis ($PO(OH)(O)OR^-$), which occurs because $PO(OH)_2OR$ is a stronger acid than H_2O, CH_3OH, HCOOH, and CF_3OH, respectively (among 13.3 and 64.6 kcal·mol^{-1}, see reactions **1b**−**4b** in Table 3). Therefore, these model reactions involve a proton transfer linked to the formation of a pentacoordinated phosphorus compound, in a similar way as many reactions of biological interest. The schematic reaction profiles are depicted in Figure 3a, which shows that the reaction begins with the formation of a hydrogen-bonded complex which occurs previous to the formation of the pentacoordinated intermediate. This is a concerted process where the proton transfer from ROH to $PO(OH)(O)OR^-$ takes place simultaneously to the addition of the RO group to phosphorus. The results displayed in Table 3 and Figure 3a show that the computed stability of the prereactive hydrogen-bonded complexes ranges among 12.7 and 27.2 kcal·mol^{-1} and that the barrier that has to be overcome to form the pentacoordinated intermediate ranges among 37.1 and 27.5 kcal·mol^{-1} for **1b**−**4b**, respectively. Table 3 and Figure 3a show that all pentacoordinated intermediates lie energetically above the reactants (among 17 and −0.4 kcal·mol^{-1}). This reaction mechanism and the corresponding energetic profile is comparable to that of the dimethylphosphate hydrolysis and the ethylene phosphate hydrolysis reported recently.[26,28]

The last model reactions we have considered are those having X = CH_3O as equatorial substitutes and correspond to reactions **1c**−**4c**. A look at the schematic energy profile in Figure 3b shows that, for **1c**, **2c**, and **3c** (blue, green, and red lines respectively), the reaction has a 5-fold well. As

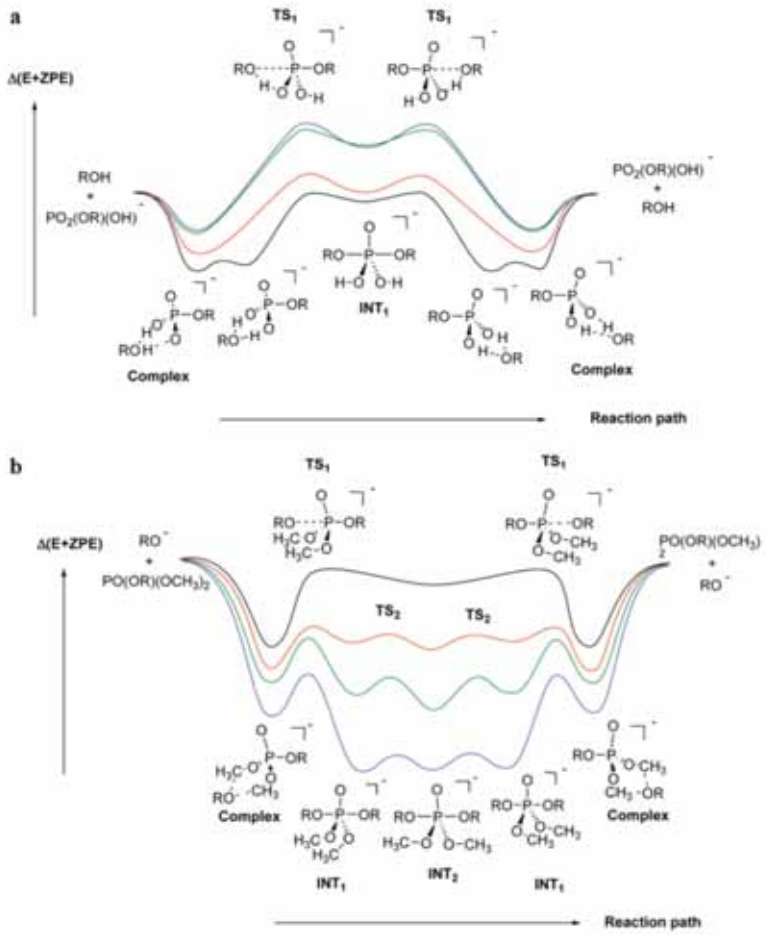


Figure 3. Schematic potential energy profiles for the nucleophilic substitution reactions: $RO^- + PO(OR)(X)_2 \rightarrow PO(OR)(X)_2 + RO^-$: (RO = HO, blue line; CH_3O, green line; HCOO, red line; and CF_3O, black line) in (a) X = (HO) and in (b) X = (CH_3O). The relative energies are computed at the *m*PW1PW91/6-311+G(3df,2p)//*m*PW1PW91/6-31+G(d) level of theory.

before, the reactions begin with the formation of a pentacoordinated hydrogen-bonded complex (first minimum), while the second, third, and fourth minima correspond to the pentacoordinated intermediates with the OCH_3 equatorial substitutes having different orientations, namely parallel to the side of the reactants (**INT1**), opposite (**INT2**), and parallel to the side of the products (**INT1**). The occurrence of similar multiple transition states separating the pentacoordinated species from the precursor complexes has been reported recently by van Bochove and co-workers,[24] who addressed this phenomenon to the increased steric bulk. For **4c** (black line) only one pentacoordinated intermediate has been found (**INT2**), being that the CH_3O equatorial substitutes are oppositely oriented. A more detailed discussion on these different conformers of the pentacoordinated phosphorus compounds will be given in the next section, and, for the aim of this section, it is only worth remarking here that the stability of the two pentacoordinated conformers differ only at most by 3 kcal·mol^{-1} (see Table 3). The results displayed in Table 3 reveal that the prereactive hydrogen-bonded complexes are computed to be among 26 and 15 kcal·mol^{-1} more stable than the reactants, and the formation of the pentacoordinated intermediate requires to surmount an energy barrier of among 7 and 12 kcal·mol^{-1}. Moreover, the stability of the pentacoordinated phosphorus intermediates follows the same trends as discussed above, namely that the intermediates having apical substitutes with higher donor character are more stable. That is 35.7 kcal·mol^{-1} for **1c** (apical substitute HO); 25.4 kcal·mol^{-1} for **2c** (apical substitute CH_3O); 15.2 kcal·mol^{-1} for **3c** (apical substitute HCOO); and 3.5 kcal·mol^{-1} for **4c** (apical substitute CF_3O). In addition, it is also worth mentioning that, as shown in Table 3, in the case of **1c** and **2c** (having apical substitutes with a large donor character) the pentacoordinated phosphorus intermediates are considerably more stable than the

Table 3. Relative Energies ($\Delta(E+ZPE)$ in kcal·mol^{-1}) Computed at the *m*PW1PW91/6-311+G(3df,2p)// *m*PW1PW91/6-31+G(d) Level of Theory for the Nucleophilic Substitution Reactions **1a**−**4d**

reaction[a]	$RO^- + POX_2(OR)$	$ROH + POX_2(OR)^-$	complex	TS1	INT1	TS2	INT2
1a	0.0		−24.3	−19.4	−33.1		
2a	0.0		−18.1	−13.3	−22.0		
3a	0.0		−20.4	−14.0	−17.9		
4a	0.0		−16.6	−9.9	−15.9		
1b	64.6	0.0	−12.7	24.4	17.2		
2b	52.4	0.0	−13.0	22.3	16.5		
3b	28.7	0.0	−18.8	5.7	1.3		
4b	13.3	0.0	−27.2	0.3	−0.4		
1c	0.0		−25.8	−18.9	−34.9	−32.7	−35.7
2c	0.0		−20.4	−13.2	−22.4	−19.9	−25.4
3c	0.0		−18.3	−11.5	−14.2	−12.7	−15.2
4c	0.0		−15.1	−2.2	−3.5		
3d	0.0				−32.1		
4d	0.0				−20.7		

[a] The following acronyms stand for the corresponding reactions: **1a** = $HO^- + OP(CH_3)_2(HO)$; **2a** = $CH_3O^- + OP(CH_3)_2(CH_3O)$; **3a** = $HCOO^- + OP(CH_3)_2(HCOO)$; **4a** = $CF_3O^- + OP(CH_3)_2(CF_3O)$; **1b** = $HO^- + OP(OH)_2(HO)$; **2b** = $CH_3O^- + OP(OH)_2(CH_3O)$; **3b** = $HCOO^- + OP(OH)_2(HCOO)$; **4b** = $CF_3O^- + OP(OH)_2(CF_3O)$; **1c** = $HO^- + OP(OCH_3)_2(HO)$; **2c** = $CH_3O^- + OP(OCH_3)_2(CH_3O)$; **3c** = $HCOO^- + OP(OCH_3)_2(HCOO)$; **4c** = $CF_3O^- + OP(OCH_3)_2(CF_3O)$; **3d** = $HCOO^- + OPF_2(HCOO)$; **4d** = $CF_3O^- + OPF_2(CF_3O)$.

prereactive hydrogen-bonded complexes, as opposed to what occurs for **3c** and **4c**.

Finally, it is also worth pointing out that the main reaction features described for these nucleophilic substitutions at phosphorus occur also in nucleophilic substitution reactions at silicon as reported by Bento and co-workers.[75]

Conformational Change in Equatorial Substitutes. The Polarization Effects. In the previous section we have pointed out that reactions **1c**−**3c** occur in several steps involving conformational changes in the orientation of the CH_3O equatorial substitutes. The corresponding energy barriers are smaller than 3.0 kcal·mol^{-1}, whereas the two conformers differ in energy at most by 3 kcal·mol^{-1} (see Table 3). Despite these small energetic differences in the two conformers, an analysis of its structures reveals significant differences with respect to the geometrical parameters concerning the apical substitutes. Therefore we have investigated the effect of the conformational changes (opposite and parallel orientation) on the equatorial substitutes in the model systems having HO and CH_3O as equatorial substitutes. In the case of the HO equatorial substitutes, only the model having HO as apical substitutes has both conformers stable (**1b** and **1b′**), whereas for the CH_3O equatorial substitutes the models with the HO, CH_3O, and HCOO apical substitutes have the two conformers stable (**1c** and **1c′**; **2c** and **2c′**; and **3c** and **3c′**; respectively). In Figure 4 we have displayed the most significant geometrical parameters of these conformers.

As pointed out in a previous section, the **1b** model has the two apical P···O(H) bond lengths equal to 1.768 Å (see Table 1). However a conformational change in the equatorial substitute leading to a parallel orientation (model **1b′**) produces an important change in the two apical P···O(H) bond lengths (1.695 and 1.910 Å, respectively); that is, the P···O apical bond length opposite to the orientation of the two equatorial OH substitutes is reduced by 0.073 Å and the other P···O apical bond length is enlarged by 0.142 Å. On the other hand, the changes in the equatorial bond lengths are very small (see Figure 4). From an energetic point of view, both conformers are separated by only 0.87 kcal·mol^{-1} ($\Delta(E + ZPE)$ value), being that **1b′** is more stable than **1b**. Looking for the origin of these differences we have first considered the possible existence of intramolecular hydrogen bond interactions that could stabilize one of these two conformers, but the AIM analysis ruled out this fact. Moreover, the NBO analysis indicates that the parallel orientation of the equatorial substitutes (structure **1b′**) induces a differential polarization effect on P, which results in a change of the phosphorus ability to bear an electronic charge and affecting therefore the axial bond length. In other words, the polarization on P produces a greater or less repulsion with the axial group (depending on the side) originating a change on the corresponding equilibrium bond distance. This polarization effect is not produced in those compounds with opposite oriented equatorial substitutes (structure **1b**) because of a cancellation effect due to the opposite orientation. For **1b**, the NBO analysis has already been reported in a previous section (see Table 2), where it has been pointed out that charge transfer occurs symmetrically. The perturbative donor−acceptor interactions involving the equatorial substitutes ($\sigma_{PO-equatorial} \rightarrow \sigma^*_{PO-apical}$) are equal to 22.6 kcal·mol^{-1} (from each of the two $\sigma_{PO-equatorial}$ to each of the two $\sigma^*_{PO-apical}$), whereas perturbative donor−acceptor interactions between the two apical bonds ($\sigma_{P1O5} \rightarrow \sigma^*_{P1O6}$ and $\sigma_{P1O6} \rightarrow \sigma^*_{P1O5}$) are both equal to 29.8 kcal·mol^{-1}. In the case of the conformer **1b′**, the situation changes radically, and the perturbative donor−acceptor interactions are not symmetrical anymore. The inductive effects, reflected in the perturbative donor−acceptor interactions involving the equatorial substitutes ($\sigma_{P1O3} \rightarrow \sigma^*_{P1O6}$ and $\sigma_{P1O4} \rightarrow \sigma^*_{P1O6}$), are 25.2 kcal·mol^{-1}, whereas ($\sigma_{P1O3} \rightarrow \sigma^*_{P1O5}$ and $\sigma_{P1O4} \rightarrow \sigma^*_{P1O5}$) are 20.5 kcal·mol^{-1}. That is, there is a greater charge transfer to the P1O6 side (σ^*_{P1O6} orbital) that affects the perturbative donor−acceptor interactions between the two apical bonds

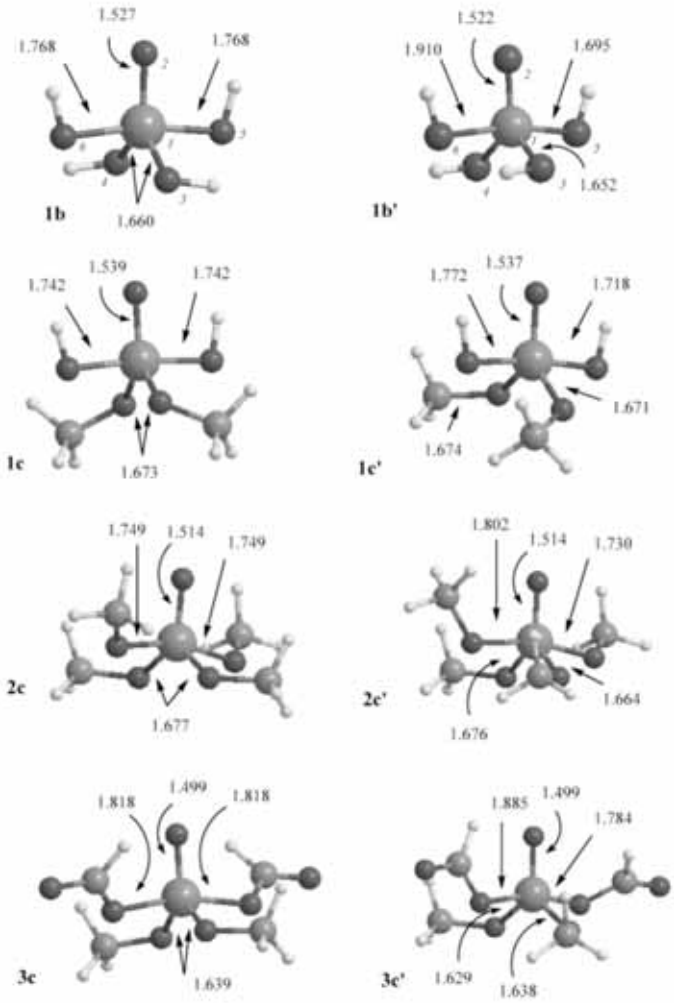


Figure 4. Selected geometrical parameters (in Å) for the optimized structures **1b**, **1b'**, **1c**, **1c'**, **2c**, **2c'**, **3c**, and **3c'**.

($\sigma_{P1O5} \rightarrow \sigma^*_{P1O6}$ = 32.1 kcal·mol^{-1} and $\sigma_{P1O6} \rightarrow \sigma^*_{P1O5}$ = 27.7 kcal·mol^{-1}), and, consequently, we can conclude that *the differential polarization effect originated by the conformational change induces a competition between the two equal apical substitutes in the pentacoordinated phosphorus compound.*

In order to visualize this induced polarization effect on P, we have considered the phosphoryl moiety derived from the two conformers **1b** and **1b'**, that is, we have deleted in both conformers the two apical substitutes. In the two resulting $PO(OH)_2$ moieties (one with the two HO opposite oriented and the other with the two HO parallel oriented) we have computed the molecular electrostatic potential (MEP), and the corresponding results are plotted in Figure 5. The phosphoryl having the two HO substitutes oppositely oriented, that derived from **1b**, (Figure 5a) has a symmetric

Figure 5. Molecular electrostatic potential representation of the $PO(OH)_2$ phosphoryl moiety of **1b** and **1b'** in a plane containing 98% of the electronic density: (a) one of the two symmetrical planes of **1b**; (b) opposite side of the equatorial hydrogens in **1b'**; and (c) side having the equatorial hydrogens in **1b'**.

distribution of the MEP in both sides of the equatorial plane, but, the phosphoryl group having the two HO substitutes parallel oriented, that derived from **1b'**, does not. Figure 5b,c

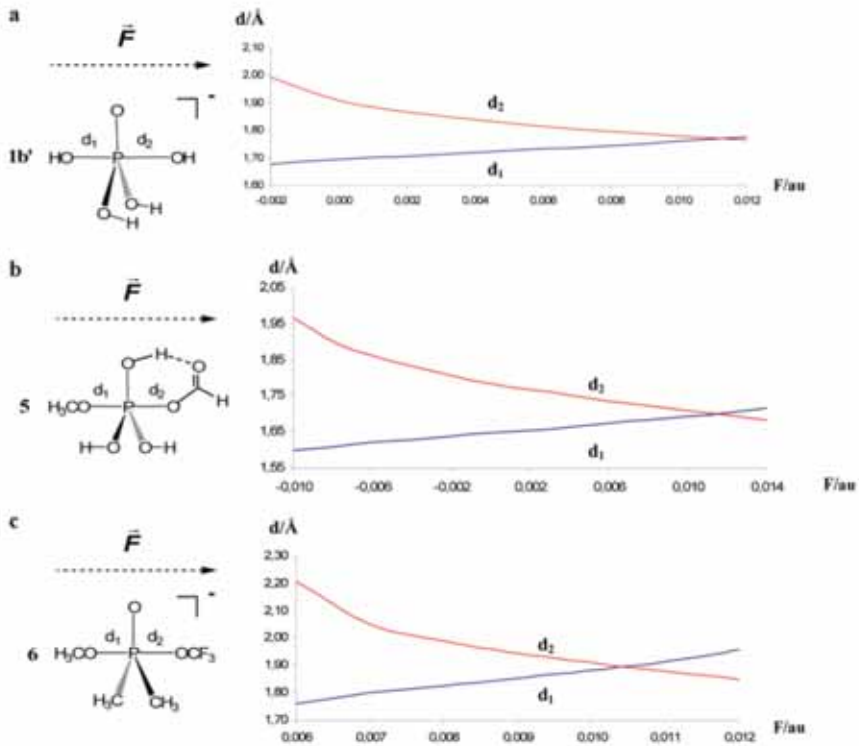


Figure 6. Dependence of the apical bond lengths on the intensity of an external electric field (*F*) for the pentacoordinated phosphorus compounds **1b′** (a), **5** (b), and **6** (c).

shows that it has a more positive charge density in the opposite side of the equatorial hydrogens. These results agree very much with the above discussion on **1b′**, where we have pointed out that a parallel orientation of the equatorial HO substitutes originates a large charge transfer to the side of the equatorial hydrogens (P1O6 bond in Figure 4).

A similar situation occurs with the compounds with equatorial substitutes CH_3O, structures **1c**−**3c**, and their corresponding conformers **1c′**−**3c′**. In a similar way as discussed above for **1b** and **1b′**, and as pointed out in the previous section, each pair of conformers differs at most by 3 kcal·mol^{-1}, being that the conformers **c** are more stable than the conformers **c′** (see Table 3). Figure 4 shows that compounds **1c**−**3c** have the CH_3O equatorial substitutes oppositely oriented and the two apical bond lengths equal (see also Table 1 and above), but a conformational change leading to the two equatorial substitutes parallel oriented (compounds **1c′**−**3c′**) produces, as just discussed for **1b** and **1b′**, a polarization effect on phosphorus that results in a significant change in the apical bond lengths. This is not so dramatic as for **1b** and **1b′**, because of the different electronegative character of the CH_3O equatorial substitutes, and the bond length changes induced amounts among 0.024 and 0.067 Å, depending on the apical substitutes (see Figure 4).

Effect of an External Electric Field. The high sensitivity to the polarization effects on the apical bonds, analyzed in the previous section, suggested to us to investigate the influence that an external electric field will produce on these kinds of bonds. To this end, we have performed a series of calculations on three pentacoordinated model systems and in two model reactions in order to analyze the effects of an external electric field on the geometries of the stationary point (minima) and on the reactivity. We have considered the effect of the external electric field in two different orientations, namely along a line in the plane defined by the phosphorus and the three equatorial substitutes and along the axis defined by the phosphorus and the apical substitutes. In the first case no substantial influence of the external electric field on the structures of the pentacoordinated models has been observed, but in the second case relevant effects have been found. Therefore, the results presented in this section correspond to the external electric field having the direction of the apical axis only. Putting the apical axis in the X direction and the origin of the coordinates at phosphorus, the external electric field follows the positive values of the X axis, while negative values means that the field direction was reversed. The results are displayed in Figures 6 and 7.

Regarding the influence of the electric field in the bonding and structural features, the first example we have considered

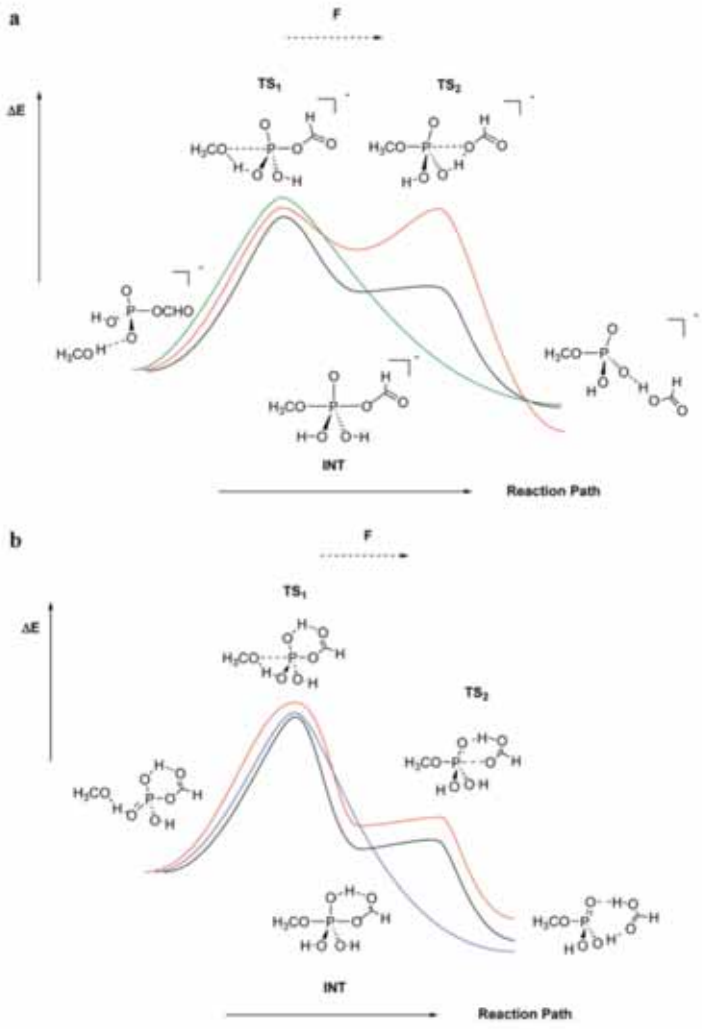


Figure 7. Schematic potential energy profiles computed under an external electric field at different intensities. (a) Corresponds to reaction 2 with $F = 0.0000$ au (black line); $F = 0.0060$ au (red line); and $F = -0.0020$ au (green line). (b) Corresponds to reaction 3 with $F = 0.0000$ au (black line); $F = 0.0060$ au (red line); and $F = -0.0060$ au (blue line).

is the model **1b′** ($PO(OH)_2(OH)_2$) discussed in the previous section and having the two equatorial OH substitutes parallel oriented (see Figure 4). We have pointed out that, in absence of external electric field, both apical P···O(H) bond lengths are different, 1.695 and 1.910 Å, respectively, for d_1 and d_2, but an external electric field produces important changes in these apical bond lengths. Figure 6a shows these changes as a function of the intensity of the external electric field. As the strength of F increases, d_1 is enlarged and d_2 is shortened so that applying an electric field of $F = 0.0111$ au both apical distances are equal, with a value of 1.769 Å. Figure 6a also shows that upon reversing the direction of the field, the opposite effect is observed, that is the d_2 is enlarged while d_1 is shortened, and with electric field with $F < -0.0030$ au, this pentacoordinated model is not stable anymore and dissociates in a process that involves a proton-transfer producing $H_2O + H_2PO_4^-$, as occurs in the process **1b** discussed in a previous section.

The second model we have considered under the effects of the electric field is $P(CH_3O)(HCOO)(HO)_3$ (compound **5**, see Figure 6b). This model is neutral, having three HO substitutes in an equatorial position, while the apical substitutes are CH_3O and HCOO. In the absence of an external electric field the two PO bond lengths are different (1.646 Å for P···OCH_3 and 1.785 Å for P···OCHO) as expected because, as pointed out above, the CH_3O apical substitute has a higher donor character. However, Figure 6b shows that the electric field produces a shortening of

the P···O(CHO) bond length and a lengthening of the P···O(CH_3) bond distance so that with an external electric field of $F = 0.0107$ au the two apical P···O bond distances become equal to 1.699 Å. Figure 6b also shows that with fields with $F > 0.0107$ au P···OCH_3 becomes larger than P···OCHO, which means that the electric field changes the relative strength of the two apical bonds.

The third model we have considered is $PO(CH_3O)(OCF_3)(CH_3)_2$ having the two CH_3 as equatorial substitutes and OCH_3 and OCF_3 as apical substitutes (compound **6**, Figure 6c). This model has been chosen because in the absence of an external electric field, the pentacoordinated phosphorus compound is not stable and dissociates into $PO(CH_3O)(CH_3)_2$ and CF_3O^-. However, under a field of $F > 0.0060$ au, this pentacoordinated model is stable, being that the P···O(CH_3) bond length is shorter than the P···O(CF_3) until $F = 0.0105$ au, where both apical P···O bond distances become equal to 1.898 Å (see Figure 6c). When F increases beyond 0.0105 the P···O(CH_3) bond distance becomes larger than the P···O(CF_3) bond length, inverting thus the relative strength of the two apical bonds.

These three examples point out *a net influence of an external electric field on the bonding competence of the two apical dative bonds on phosphorus*.

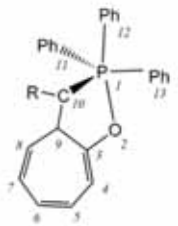
With regard to the study of the effect of F on the reactivity we have considered the two following nucleophilic substitutions:

$$CH_3OH + POO(OH)(OCHO)^- \rightarrow POO(OH)(CH_3O)^- + HCOOH \quad (2)$$

$$CH_3OH + PO(OH)_2(OCHO)^- \rightarrow PO(OH)_2(CH_3O) + HCOOH \quad (3)$$

These two reactions differ in the fact that in the second one we have added a proton in order to have a neutral reaction. The results are displayed in Figure 7. In both cases the reaction begins with the formation of a prereactive hydrogen-bonded complex, whereas in the exit channel a hydrogen-bonded complex is also formed before the release of the products. As we are mainly interested in what concerns the pentacoordination at phosphorus, we will consider these hydrogen-bonded complexes as reactive products of reactions 2 and 3. Moreover, as for reactions **1b**−**4b** discussed above, these reactions involve, in the entry and exit channels, a proton transfer which is linked to the formation (breaking) of the pentacoordination at phosphorus. For reaction 2, Figure 7a shows that in the absence of an external electric field (black line), the pentacoordinated phosphorus intermediate **7** is computed to be 17.2 kcal·mol^{-1} higher in energy than the prereactive complex. Its formation (via **TS1**) requires the surmounting of an energy barrier of 34.3 kcal·mol^{-1}, whereas the energy barrier for the exit channel (**TS2**) is only 1.3 kcal·mol^{-1}, that is, **TS1** is clearly the limiting step of the reaction. Figure 7a shows also that the reaction profile is significantly altered under an external electric field. Thus, with a $F = 0.0060$ au (red line), the pentacoordinated intermediate **7** is destabilized by about 9 kcal·mol^{-1}, and, more interestingly, the computed energy barrier for the exit channel (**TS2**) is the same as that of the back reaction (**TS1**) to the reactants. On the other side, with an external field of $F = -0.0020$ au (green line), the pentacoordinated phosphorus intermediate is not stable anymore, and the reaction occurs in a single step. A similar behavior is observed for the neutral reaction 3 (Figure 7b). In the absence of an external electric field, the reaction occurs through the pentacoordinated intermediate **6** (black line) and is slightly destabilized when a $F = 0.0060$ au is applied (red line). However, with an $F = -0.0060$ au (blue line) the pentacoordinated intermediate is not stable anymore, and the reaction occurs in a single step. These two examples point out that *the external electric field affects the stability of pentacoordinated phosphorus compounds and it may also affect the reactivity of nucleophilic substitution at phosphorus*.

Scheme 2 [a]



[a] **8a**: R = CH_3; **8b**: R = H; **8c**: R = CN.

These results may be of relevance in biological reactions involving pentacoordinated phosphorus, where the electric field originated by the folded protein could influence the catalytic process. In fact, it has been pointed out very recently the role of the electric field in the active site of the aldose reductase[76] and how the electric field may control the selectivity in heme enzymes.[77]

Triphenylphosphonium Ylide Derivatives. In an attempt to get a deeper insight in the hypervalence at phosphorus we have extended our investigation to the study on the bonding features of three neutral triphenylphosphonium ylide derivatives (**8**, Scheme 2) having pentacoordination at phosphorus and for which crystallographic X-ray data are available.[78,79]

These compounds have a trigonal bipyramid structure and are interesting for the purposes of this investigation because a change in the substitute R (R = CH_3 (**8a**), H (**8b**), and CN (**8c**)), which is not directly bonded to phosphorus, results in large changes in the P···O bond distance (2.00 Å for **8a**; 2.21 Å for **8b**; and 2.36 Å for **8c** (X-ray data)). The X-ray data first suggested that **8a** and **8b** form the PO bond but **8c** does not. Further analysis of the crystallographic data, together with results from ^{31}P and ^{13}C NMR spectra, had lead **8a**, **8b**, and **8c** to be viewed as resonance hybrids of structures A, B, and C (Scheme 3).[79,80] The δ_P and δ_C NMR spectra have been also collected in Table 4. **8a** shows a large δ_P (among −16.0 and −22.1 ppm, see also Table 4) that suggested a large contribution of the P−O bonding of the resonance structure A (Scheme 3). On the other hand, **8c** has a δ_P among −2.8 and −9.0 and has been related to the resonance structures B and C.

Scheme 3

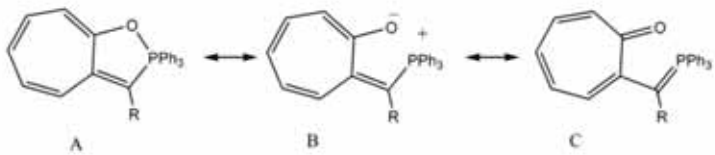


Table 4. Experimental and Computed ^{31}P and ^{13}C NMR Spectra (δ in ppm) for Compounds **8a**, **8b**, and **8c**[a]

	experimental values[69,70]					this work (gas phase)[b]		
	δ_P ($CDCl_3$)		δ_P (solid)	δ_C ($CDCl_3$)			δ_C	
compd	rt	−60 °C	rt	C3	C10	δ_P	C3	C10
8a	−17.9	−16.0	−22.1	170.1	87.0	−21.8	179.7	84.2
8b	−3.6	−1.8	−11.1	173.6	75.3	−16.2	180.2	77.8
8c	+7.8	+9.0	+2.8	177.5	49.4	−6.35	181.2	61.0

[a] Atom numbering is according to Scheme 2. [b] δ values are relative to H_3PO_4 for P and to TMS for C.

In the present work we have fully optimized and characterized as true minima the structures **8a**, **8b**, and **8c**, and their most significant geometrical parameters are displayed in Figure 8. It is gratifying to observe that the computed P···O bond distances compare well with the X-ray values for **8a** and **8b** (2.067 Å and 2.213 Å, respectively), whereas for **8c** our computed P···O bond length (2.239 Å) is 0.121 Å shorter than the X-ray value. Moreover, the remaining geometrical parameters compare also quite well with the experimental results. At this point, it should be taken into account that the calculated values should be compared with gas-phase values, while the X-ray data from the literature include packing effects that are shown to have an important role.[5,71] Besides the absolute values, the computed geometrical parameters follow the same trends with respect to the P···O bond lengths (**8a** < **8b** < **8c**). The bonding features have been analyzed, as above, employing the AIM and NBO methods, and the most significant results are displayed in the Supporting Information (Table S5). For each of the three triphenylphisphonium ylide considered (**8a**, **8b**, and **8c**), we have found a bcp between the phosphorus and oxygen atoms having the same topological features as those described in the previous sections for the P−O_{apical} bonds in the model systems, that is, the values of the density and the Laplacian of the density are small and positive, indicating that there is a *PO bond, that can be classified as dative*. Moreover, and as above, the NBO analysis indicates that the phosphorus has a formal sp^2 hybridization scheme. On the other hand, the large differences in the P···O bond distances observed for the three compounds, and originated by the different substitutes R, can be mainly associated with the different ability to delocalize the π system through the seven-member ring. Thus, **8c** with R = CN has a certain amount of π character between C and N, which prevents, in part, the delocalization of the π system through the C9−C10 bond. This results in a shorter CO bond distance with less ability to transfer charge to phosphorus, and consequently the P···O bond distance is larger. The opposite case **8a**, with R = CH_3, implies a different delocalization of the π system through the seven-member ring resulting in a larger CO bond distance with more ability to transfer charge from oxygen to phosphorus and consequently with a smaller P···O bond length. In any case, these results show that small changes in the electronic features produce large changes in the P···O bonding.

For the sake of completeness we have also computed the ^{31}P and ^{13}C NMR spectra of **8a**, **8b**, and **8c**, and the results have been collected in Table 4 along with the experimental data. The computed NMR spectra correspond to gas-phase

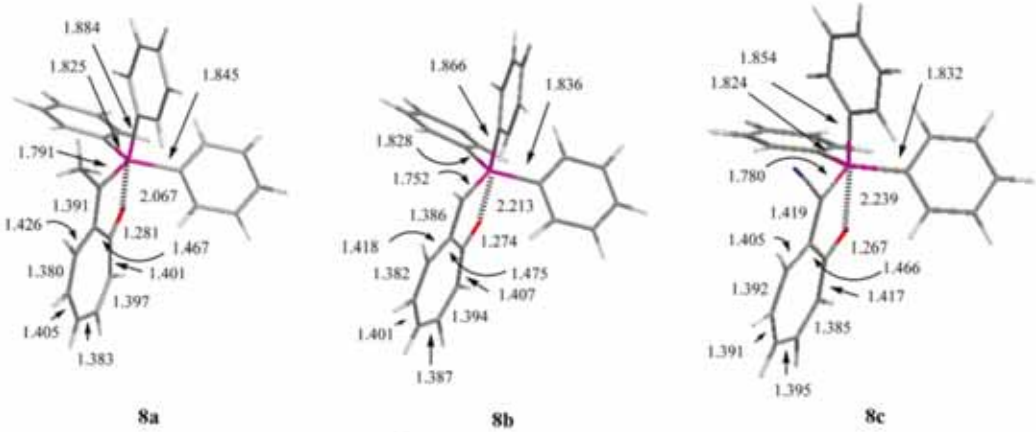


Figure 8. Selected geometrical parameters (in Å) for the optimized structures **8a**, **8b**, and **8c**. Atom numbering is according to Scheme 2.

optimized structures and show the same tendencies as the experimental values, that is larger δ_P for **8a** than for **8b** and for **8c** (−21.8, −16.2, and −6.35 ppm, respectively). Moreover, these results agree with the electronic features of the P←O dative bond. The stronger the P←O bond is, the higher the charge transfer associated with the dative bond and the shorter the corresponding bond length, which results in a higher shielding on P as reflected in the NMR spectra. It appears thus that the ^{31}P NMR spectra is a direct measure of the strength of dative bonding at phosphorus, and changes of its value in different media (as for instance in solid phase or in $CDCl_3$ for **8a**, **8b**, and **8c**, see refs 79 and 80 and also Table 4) would reflect differences in the bond length and strength. This also agrees with the linear correlation observed between δ_P and the X-ray P←O bond length as reported by Naya and Nitta.[79]

Conclusions

The results of the present investigation lead us to emphasize the following points: (1) All the pentacoordinated phosphorus compounds considered in this work have a trigonal bipyramid structure where the apical bonds show great variability. The topological and NBO analysis of the corresponding wave function indicates that these apical bonds can be classified as dative. These compounds are charge-transfer complexes, where the phosphorus has a formal sp^2 hybridization, which is compatible with the diagram based on a three-center four-electron (3c4e) model. (2) The features of the apical bonds depend strongly on the nature of the apical and equatorial substitutes. Compounds having apical substitutes with higher donor character are more stable and possess shorter apical bonds. On the other hand, the higher the donor character of the equatorial substitutes, the larger the apical bond length and the destabilization effect in pentacoordinated phosphorus compounds. (3) Polarization and electric field effects play an important role in the dative bonds of pentacoordinated phosphorus compounds, with consequences in both the geometry and the stability. These effects may change the competition between different apical substitutes, and they can even alter the reactivity of nucleophilic substitution at phosphorus. These effects may be of great relevance in enzymatic reactions, where the electric field originated by the folded protein could influence the catalytic process. (4) With regard to the three triphenylphosphonium ylide compounds considered (**8a**, **8b**, and **8c**), our results predict quite well the experimental (X-ray) geometrical data from the literature and show that in all cases there is a dative bond between the phosphorus and oxygen atoms, whose strength is correlated to the NMR displacement at P.

Acknowledgment. This research has been supported by the Generalitat de Catalunya (Grant 2005SGR00111). The calculations described in this work were carried out at the Centre de Supercomputació de Catalunya (CESCA) and at the Centro de Supercomputación de Galicia (CESGA), whose services are gratefully acknowledged. R.C. thanks also the Spanish Ramón y Cajal program.

Supporting Information Available: Cartesian coordinates of all structures reported in this paper and tables containing apical bond lengths for several model systems, optimized with different methods, activation and reaction energies for reaction **1b** obtained at different levels of theory, and AIM topological parameters for compounds **8**. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Skordalakes, E.; Dodson, G. G.; Green, D. S.; Goodwin, C. A.; Scully, M. F.; Hudson, H. R.; Kakkar, V. V.; Deadman, J. J. *J. Mol. Biol.* **2001**, *311*, 549−555.

(2) Lahiri, S. D.; Zhang, G. F.; Dunaway-Mariano, D.; Allen, K. N. *Science* **2003**, *299*, 2067−2071.

(3) Blackburn, G. M.; Williams, N. H.; Gamblin, S. J.; Smerdon, S. J. *Science* **2003**, *301*, 1184.

(4) Allen, K. N.; Dunaway-Mariano, D. *Science* **2003**, *301*.

(5) Holmes, R. R. *Acc. Chem. Res.* **2004**, *37*, 746−753.

(6) Leiros, I.; McSweeney, S.; Hough, E. *J. Mol. Biol.* **2004**, *339*, 805−820.

(7) Williams, N. H. *Biochim. Biophys. Acta* **2004**, *1697*, 279−287.

(8) Hengge, A. C.; Onyido, I. *Curr. Org. Chem.* **2005**, *9*, 61−74.

(9) Cleland, W. W.; Hengge, A. C. *Chem. Rev.* **2006**, *106*, 3252−3278.

(10) Wittinghofer, A. *Trends. Biochem. Sci.* **2006**, *31*, 20−23.

(11) Swamy, K. C.; Kumar, N. S. *Acc. Chem. Res.* **2006**, *39*, 324−333.

(12) Catrina, I.; O'Brien, P. J.; Purcell, J.; Nikolic-Hughes, I.; Zalatan, J. G.; Hengge, A. C.; Herschlag, D. *J. Am. Chem. Soc.* **2007**, *129*, 5760−5765.

(13) Mildvan, A. S. *Proteins* **1997**, *29*, 401−416.

(14) Allen, K. N.; Dunaway-Mariano, D. *Trends. Biochem. Sci.* **2004**, *29*, 495−503.

(15) Vedejs, E.; Marth, C. F. *J. Am. Chem. Soc.* **1988**, *110*, 3948−3958.

(16) Tremblay, L. W.; Zhang, G. F.; Dai, J. Y.; Dunaway-Mariano, D.; Allen, K. N. *J. Am. Chem. Soc.* **2005**, *127*, 5298−5299.

(17) Godfrey, S. M.; McAuliffe, C. A.; Pritchard, R. G.; Sheffield, J. M. *Chem. Commun.* **1998**, 921−922.

(18) Chandrasekaran, A.; Timosheva, N. V.; Day, R. O.; Holmes, R. R. *Inorg. Chem.* **2003**, *42*, 3285−3292.

(19) Hu, C. H.; Brinck, T. *J. Phys. Chem. A* **1999**, *103*, 5379−5386.

(20) Bianciotto, M.; Barthelat, J. C.; Vigroux, A. *J. Phys. Chem. A* **2002**, *106*, 6521−6526.

(21) Berente, I.; Beke, T.; Náray-Szabó, G. *Theor. Chem. Acc.* **2007**, *118*, 129−134.

(22) Wang, Y. N.; Topol, I. A.; Collins, J. R.; Burt, S. K. *J. Am. Chem. Soc.* **2003**, *125*, 13265−13273.

(23) Pepi, F.; Ricci, A.; Rosi, M.; Di Stefano, M. *Chem.-Eur. J.* **2004**, *10*, 5706−5716.

(24) Van Bochove, M. A.; Swart, M.; Bickelhaupt, F. M. *J. Am. Chem. Soc.* **2006**, *128*, 10738−10744.

(25) Lopez, X.; Schaefer, M.; Dejaegere, A.; Karplus, M. *J. Am. Chem. Soc.* **2002**, *124*, 5010−5018.

(26) Lopez, X.; York, D. M.; Dejaegere, A.; Karplus, M. *Int. J. Quantum Chem.* **2002**, *86*, 10−26.

(27) Lopez, X.; Dejaegere, A.; Leclerc, F.; York, D. M.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 11525−11539.

(28) Imhof, P.; Fischer, S.; Kramer, R.; Smith, J. C. *J. Mol. Struct. (THEOCHEM)* **2005**, *713*, 1−5.

(29) Grzyska, P. K.; Czyryca, P. G.; Golightly, J.; Small, K.; Larsen, P.; Hoff, R. H.; Hengge, A. C. *J. Org. Chem.* **2002**, *67*, 1214−1220.

(30) Chen, S. L.; Fang, W. H.; Himo, F. *J. Phys. Chem. B* **2007**, *111*, 1253−1255.

(31) Klahn, M.; Rosta, E.; Warshel, A. *J. Am. Chem. Soc.* **2006**, *128*, 15310−15323.

(32) Iche-Tarrat, N.; Ruiz-Lopez, M.; Barthelat, J. C.; Vigroux, A. *Chem.-Eur. J.* **2007**, *13*, 3617−3629.

(33) Cramer, C. J.; Gustafson, S. M. *J. Am. Chem. Soc.* **1993**, *115*, 9315−9316.

(34) Seekute, J.; Menke, J. L.; Emnett, R. J.; Patterson, E. V.; Cramer, C. J. *J. Org. Chem.* **2005**, *70*, 8649−8660.

(35) Uchimaru, T.; Tanabe, K.; Nishikawa, S.; Taira, K. *J. Am. Chem. Soc.* **1991**, *113*, 4351−4353.

(36) Yliniemela, A.; Uchimaru, T.; Tanabe, K.; Taira, K. *J. Am. Chem. Soc.* **1993**, *115*, 3032−3033.

(37) Tole, P.; Lim, C. M. *J. Phys. Chem.* **1993**, *97*, 6212−6219.

(38) Lim, C.; Tole, P. *J. Phys. Chem.* **1992**, *96*, 5217−5219.

(39) Lim, C.; Tole, P. *J. Am. Chem. Soc.* **1992**, *114*, 7245−7252.

(40) Chang, N. Y.; Lim, C. *J. Phys. Chem. A* **1997**, *101*, 8706−8713.

(41) Chang, N. Y.; Lim, C. *J. Am. Chem. Soc.* **1998**, *120*, 2156−2167.

(42) Dudev, T.; Lim, C. *J. Am. Chem. Soc.* **1998**, *120*, 4450−4458.

(43) Zhou, D. M.; Taira, K. *Chem. Rev.* **1998**, *98*, 991−1026.

(44) Taira, K.; Uchimaru, T.; Storer, J. W.; Yliniemela, A.; Uebayasi, M.; Tanabe, K. *J. Org. Chem.* **1993**, *58*, 3009−3017.

(45) Uchimaru, T.; Tsuzuki, S.; Storer, J. W.; Tanabe, K.; Taira, K. *J. Org. Chem.* **1994**, *59*, 1835−1843.

(46) Uchimaru, T.; Stec, W. J.; Tsuzuki, S.; Hirose, T.; Tanabe, K.; Taira, K. *Chem. Phys. Lett.* **1996**, *263*, 691−696.

(47) Range, K.; McGrath, M. J.; Lopez, X.; York, D. M. *J. Am. Chem. Soc.* **2004**, *126*, 1654−1665.

(48) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664−675.

(49) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

(50) Gilbert, T. M. *J. Phys. Chem. A* **2004**, *108*, 2550−2554.

(51) Moeller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(52) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. *Chem. Phys. Lett.* **1990**, *166*, 281.

(53) Head-Gordon, M.; Head-Gordon, T. *Chem. Phys. Lett.* **1994**, *220*, 122.

(54) Ishida, K.; Morokuma, K.; Kormornicki, A. *J. Chem. Phys.* **1977**, *66*, 2153.

(55) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154.

(56) Gonzalez, C.; Schlegel, H. B. *J. Phys. Chem.* **1990**, *94*, 5523.

(57) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

(58) Cizek, J. *Adv. Chem. Phys.* **1969**, *14*, 35.

(59) Barlett, R. J. *J. Phys. Chem.* **1989**, *93*, 1963.

(60) Pople, J. A.; Krishnan, R.; Schlegel, H. B.; Binkley, J. S. *Int. J. Quantum Chem. XIV* **1978**, 545−560.

(61) Truhlar, D. G. *Chem. Phys. Lett* **1998**, *294*, 45−48.

(62) Fast, P. L.; Sanchez, M. L.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *111*, 2921−2926.

(63) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(64) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. F. *C. J. Chem. Phys.* **1996**, *104*, 5497−5509.

(65) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251−8260.

(66) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R. J. A.; Montgomery, J.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.01*; Gaussian, Inc.: Wallingford, CT, 2004.

(67) Shaftenaar, G.; Noordik, J. H. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 123−134.

(68) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899−926.

(69) Bader, R. F. W. *Atoms in Molecules. A Quantum theory*; Clarendon Press: Oxford, 1995; Vol. 22, pp 1−458

(70) Bader, R. F. W. *AIMPAC*. http://www.chemistry.mcmaster.ca/aimpac (accessed May 2002)

(71) Leopold, K. R.; Canagaratna, M.; Phillips, J. A. *Acc. Chem. Res.* **1997**, *30*, 57−64.

(72) Anglada, J. M.; Bo, C.; Bofill, J. M.; Crehuet, R.; Poblet, J. M. *Organometallics* **1999**, *18*, 5584−5593.

(73) Reed, A. E.; Schleyer, P. V. *J. Am. Chem. Soc.* **1990**, *112*, 1434−1445.

(74) Massey, A. G. In *Main Group Chemistry*; John Wiley and Sons: Chichester, England, 2000.

(75) Bento, A. P.; Bickelhaupt, F. M. *J. Org. Chem.* **2007**, *72*, 2201−2207.

(76) Suydam, I. T.; Snow, C. D.; Pande, V. S.; Boxer, S. G. *Science* **2006**, *313*, 200−204.

(77) Shaik, S.; de Visser, S. P.; Kumar, D. *J. Am. Chem. Soc.* **2004**, *126*, 11746−11749.

(78) Kawamoto, I.; Hata, T.; Kishida, Y.; Tamura, C. *Tetrahedron Lett.* **1971**, 2417-&.

(79) Naya, S.; Nitta, M. *J. Chem. Soc., Perkin Trans.* **2002**, *2*, 1017−1023.

(80) Nitta, M.; Naya, S. *J. Chem. Res.-S* **1998**, 522−523.

CT700220Z

4.1.2. Pentacoordinated Phosphorus: methodologies to describe polarization effects

After having observed, in the previous study, the relevance of electric field effects in phosphoryl transfer reactions, it is clear that, from the methodological point of view, modeling this kind of reactions requires the use of computational methods able to describe how electric fields affect the geometry and reactivity. Given the implications this might have in enzyme active sites and the large size of such systems, it is necessary to find computationally efficient methods providing a reasonable description of these effects.

We first aimed to find the basis set with best compromise between computational cost and accuracy for use in conjunction with the *m*PWPW91 functional, which has a good performance in describing pentacoordinated phosphorus species and was used in the previous study. We chose two pentacoordinated phosphorus models to test the accuracy of different basis sets in describing the electric field effects on geometry and reactivity. In particular, we focused on the electric field dependence of the apical bond distances and the dissociation energy barrier of the two apical bonds. As a reference we used the 6-311+G(3df,3pd) basis set and found that the 6-31+G(d) was the best choice.

Taking into account the importance of polarization and diffuse functions in modeling these effects, we questioned whether semi-empirical methods, which use minimal basis sets and are widely used to study enzymatic reaction mechanisms with QM/MM approaches, are able to describe the polarization of pentacoordinated phosphorus. To this aim, we made a systematic study on the performance of a wide range of semi-empiricals (with and without *d* orbitals). Among them, the AM1/d-PhoT developed by Nam and co-workers for phosphoryl transfer reactions was the best choice for reproducing electric field effects, in comparison with DFT. It is worth pointing out that the most popular AM1 and PM3 methods exhibit bad performance. Therefore, the recommendation when modeling enzymatic phosphoryl transfer reactions is the use of the semi-empirical AM1/d-PhoT.

We previously hypothesized that these electric field effects could have some relevance in enzyme active sites, since the spatial distribution of polar amino acid residues generates a net electric field. We wondered whether enzyme electric fields are strong enough to produce similar effects to those observed in the tested models. To this aim, we calculated the electric field in three different enzymes involved in a phosphoryl transfer reaction: β-phosphoglucomutase, fructose-1,6-bisphosphatase and N-Acetyl-Glutamate Kinase. We focused on the projection of the net electric field in the apical direction of the phosphorylation reactions. The values we obtained for the X-ray structures of these enzymes were in the range of 0.002-0.005 au, which falls in the range of intensities

considered in the models. Therefore, it is plausible that electric fields present in enzyme active sites can affect the reaction profile in enzymes.

A detailed presentation of the results and methodologies used in this study can be found in the article: *Description of pentacoordinated phosphorus under an external electric field: which basis sets and semi-empirical methods are needed*? (2008) Phys. Chem. Chem. Phys., 10, 2442-2450.

PAPER www.rsc.org/pccp | Physical Chemistry Chemical Physics

Description of pentacoordinated phosphorus under an external electric field: which basis sets and semi-empirical methods are needed?†

Enrique Marcos, Josep M. Anglada and Ramon Crehuet*

Received 2nd January 2008, Accepted 14th February 2008
First published as an Advance Article on the web 7th March 2008
DOI: 10.1039/b719792f

Phosphate transfer reactions are ubiquitous in nature and play fundamental roles in ATP hydrolysis and protein phosphorylation processes. The mechanisms of these reactions involve a pentacoordinated phosphorus atom that can be an intermediate or a transition state. These structures are very sensitive to both internal and external electrostatic effects and their description with quantum mechanical methods is challenging. We have investigated the variations of geometry an energetics under an external electric field for two different molecules and their transition states of formation. The DFT method, with the *m*PW1PW91 functional employing several basis sets, and different semi-empirical methods have been tested. Compared to zero-field cases, one needs more extended basis sets to achieve the same precision. A good compromise for large systems is the 6-31 + G(d). Many semi-empirical methods are unable to describe polarisation effects in pentacoordinated structures. The best methods to describe geometries are PM6 and AM1/d-PhoT and for energetics AM1/d-PhoT. Methods without d orbitals have poorer performances but the best among those is the AM1 parametrization of Arantes *et al* (*Phys. Chem. Chem. Phys.*, 2006, **8**, 347).

1. Introduction

Phosphate transfer reactions are ubiquitous in nature and play fundamental biological roles.[1,2] Phosphorylation of proteins is the main signalling process to control metabolic pathways and ATP hydrolysis, *i.e.* phosphoryl transfer to water, represents the main source of chemical energy used by cells. The interest in the understanding of these reactions is thus unquestionable, and especially the role of enzymes in accelerating a chemical step that otherwise would not take place. Indeed, inositol phosphatase has the greatest catalytic effect ever found, accelerating a reaction by 21 orders of magnitude.[3]

Because of the hypervalency of phosphorus, phosphoryl transfer reactions are more complex than their carbon analogues, but two main mechanisms can be devised: *dissociative* leading to a metaphosphate intermediate or *associative* giving rise to a pentacoordinated phosphorus, which can represent both a relatively stable intermediate (phosphorane), or a transition state.[4] Those processes occurring through a long-lived intermediate are important in chiral reactions, since such a pentacoordinated structure can lead to a retention of configuration as a result of a pseudorotation process.

A deep knowledge of these processes requires the concurrence of experimental and theoretical studies. The theoretical approach employed in such studies is usually based on the hybrid QM/MM formalism, where a small part of the enzyme (the reactive site) is treated by a quantum mechanical method and the environment is treated employing molecular mechanics.[5–7] The large size of these systems generates a rugged potential energy landscape[8] which calls for the use of extensive sampling methods, namely molecular dynamics or (less frequently) Monte Carlo techniques. A consequence of this computational demand is that semi-empirical methods are frequently used in the QM part of the QM/MM approach. In some cases, local optimisations, single-point calculations or corrections based on a few reaction coordinates[9] can be done using a more powerful (and costly) DFT or *ab initio* method. Therefore, it is extremely important to be confident that the theoretical approach describing phosphoryl transfer reactions can account for the main features of the process to be studied.

In a recent work,[10] we have investigated the bonding features on a series of pentacoordinated phophorus intermediates which present a trigonal bipyramid structure. The apical bonds have dative character with formal sp^2 hybridization at P and with the d orbitals acting as polarization functions. We have also pointed out that their energy and stability depends largely on the inductive effects originated by the nature of the equatorial substitutes at phosphorus. Moreover, the intramolecular polarization produced by the ability of both apical groups to transfer charge to the phosphoryl moiety plays an important role, so that when the two apical groups are different there is a competition of these two groups to form a dative bond to P (a competitive effect). In addition, and closely linked to such polarization effects, an external electric field can modify not only the geometrical features, but also the stability and even the reactivity of nucleophilic substitution reactions at phosphorus.

Theoretical and Computational Chemistry Group, Departament de Química Orgánica Biológica, Institut d'Investigacions Químiques i Ambientals de Barcelona, IIQAB - CSIC, c/Jordi Girona 18, E-08034 Barcelona, Spain. E-mail: rcsqtc@iiqab.csic.es; Fax: +34 93 204 59 04; Tel: +34 93 400 61 11

† Electronic supplementary information (ESI) available: Tables containing the Cartesian coordinates of the optimized geometries and Figures depicting the variation of geometry and energy with respect to the field. See DOI: 10.1039/b719792f

In this work, and in prospect to a further study on more complex systems, we aim to extend the previous investigation by studying the effect of an external electric field on two associative reactions. To carry out this study we have considered two kinds of theoretical approaches: namely a DFT method, with different basis sets, and a set of semi-empirical methods. The goal is to compare the reliability of different theoretical approaches to describe these polarization effects. Our study has been focused on the model reactions 1 and 2 (see Fig. 1), where we considered the pentacoordinated intermediate and the transition state involved in the weakest apical bond cleavage. The choice of both reactions tries to be a complete set of representative situations of the competition between the apical substitutes. We have chosen an intermediate with equal substitutes (**1**), where the polarisation in the P arises from the conformation of the hydroxyls in the phosphoryl moiety. In the second intermediate (**2**) the asymmetry arises from the different electrodonor ability of the axial substitutes.

The source of polarisation effects and their effect on the type of bond have been studied in our previous work[10] and here we focus on the correct description of these effects under different theoretical approaches. We have not considered total exothermicities as the generation of separated species in an electric field results in energies that mainly depend on the reorientation of the molecules.

Finally, it is worth pointing out here that the range of intensities of the electric field considered has been chosen by calculating the electric field at P of three different enzymes that transfer a phosphoryl group and comparing these results with others in the literature (see below).

2. Methodology

2.1 Quantum mechanical methods

The DFT calculations have been carried out by using the *m*PW1PW91 functional, which was designed by Adamo and co-workers[11] in order to improve the description of non-covalent interactions such as dative bonds, as occurs in pentacoordination at phosphorus. In a previous work, we have already checked the goodness of this functional by performing some test calculations comparing results obtained with this functional with *ab initio* methods.[10] In the present work, we have employed in a first stage the large 6-311+G(3df,3pd) basis set, as it is considered flexible enough to account for polarization effects. In a second stage, we have also employed the 6-31+G(d), 6-31G(d), 6-31+G, 6-31G, 6-311G and 6-311G(d) basis sets in order to check the dependence of the results on the basis set. These relatively small basis sets have been chosen as they are a usual choice in the study of large systems such as enzymes. In all cases, all stationary poins have been optimized and characterized by calculating the harmonic vibrational frequencies to verify the nature of the corresponding stationary point (minima or transition state), and to provide the zero point vibrational energy (ZPE) and enthalpic corrections. The DFT calculations have been performed using the Gaussian03 program package.[12]

There is a plethora of semi-empirical approaches to treat phosphorus compounds. In the current work, we have chosen the classical, but still widely used, AM1[13] and PM3.[14] We have also included a new parametrization of AM1 optimized for C, H, O, P and S compounds by Arantes *et al.*[15] (hereafter called AM1/Arantes) and the recent RM1.[16] Among the methods that include d orbitals, we have opted for the general MNDO/d[17] and PM6[18] and the specific parametrization of AM1/d-PhoT for P, H, and O by Nam *et al.*[19] and the AM1* by Winget *et al.*[20] All of them, except MNDO/d, include a modified form of the core–core term so that not only the parameters are different, but also the expression of the Hamiltonian. We refer the reader to the original publications for the rationale behind each different semi-empirical flavour. Semi-empirical calculations were performed with MOPAC2007.[21] AM1, PM3, PM6, MNDO/d and RM1 are included in the normal distribution, AM1/Arantes requires only a new set of parameters and in order to include AM1* and AM1/Pho-T we have modified the source code by implementing the different core–core interactions.

All calculations have been performed in vacuum, because continuum solvation models are usually not used in QM/MM calculations and because there is not a universal way to model

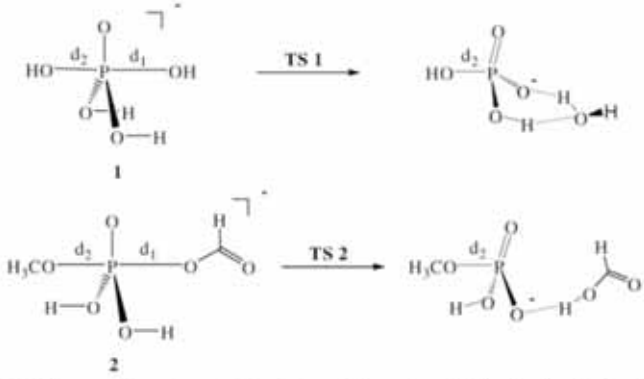


Fig. 1 Scheme of the reactions of dissociation of the weakest apical bonds of models **1** and **2**.

the electrostatic properties of an enzyme active site with continuum models.[22]

2.2 Application and calculation of an electric field

In models **1** and **2** a uniform electric field was applied in the direction of the shorter P–O apical bond. Even though the molecules have net charge, translation (which would naturally occur in a uniform electric field) is removed by the codes by carrying out the calculation in *z*-matrix form.

The calculation of the electric field at the active site of the enzymes was performed from the classical description given by molecular mechanics and only the residues of the protein, at their crystallographic position, were considered. The OPLS charges were used and standard protonation states were chosen for all residues. Crystallographic waters were retained and all hydrogen atom positions have been optimized. A dielectric constant of 4 was chosen as usually done in electrostatic calculations in proteins.[22,23] Although the protonation state of a residue or the presence of a metal close to the phosphoryl can change the field, the aim of these calculations is to show the range of magnitudes of electric fields in active sites, and not a detailed evaluation of each system. For one of the enzymes, the electric field at the active site has been also calculated by performing several hybrid QM/MM calculations. In this way, we have taken into account the influence of polarisation. For these calculations we have used the AM1 semi-empirical hamiltonian. The protonation of the residues and the evaluation of the electric field have been done by means of the Pymol program[24] and the Dynamo Library,[25] respectively.

3. Results and discussion

3.1 Electric field at the enzyme active site

We have evaluated the electric field in three different enzymes that transfer a phosphoryl group: β-phosphoglucomutase (β-FGM), fructose-2,6-bisphosphatase (FBP) and *N*-acetyl-L-glutamate kinase (NAGK) whose PDB codes are 1O08,[26] 1NUX[27] and 1OH9,[28] respectively. The crystallographic structures of β-FGM and FBP contain the intermediate state of the phosphoryl transfer (a phosphorane in the former and a metaphosphate in the latter). With respect to NAGK, we have considered the X-ray structure of the enzyme complexed with an inhibitor (AlF_4^-) analogous to the transition state. The electric field has been computed at the P position in the first two cases and at the Al position in the third one. It is necessary to point out that we have focused on the projection of the electric field in the apical direction in order to compare directly with the range of magnitudes that we have considered in models **1** and **2**. Our results are 0.004, 0.005 and 0.002 au for β-FGM, FBP and NAGK, respectively. Remark that these are the fields generated by the protein residues, and that the P atom also 'feels' the field generated by the substrates. In an enzymatic QM/MM calculation the latter would be included in the QM region. For NAGK, the evaluation of the electric field has been also calculated by including a sphere of residues around the active site in a hybrid QM/MM calculation. Inclusion of this polarisation changes the field at most by a 10% (see Table S1 in the ESI†) but the magnitude of the field remains the same. The use of a dielectric constant of 4 is also questionable, especially for QM/MM calculations, where electrons are explicitly included. A detailed consideration of this issue will be necessary when performing the actual calculations of the enzyme mechanism. Some authors[29,30] would justify lower dielectric constants which would increase the calculated field. That would reinforce the aim of this paper on the importance of a correct field-dependence description.

The external electric field that we use mimics the protein electrostatic environment and adds a polarisation effect to the effects generated by the ligands. In our models, the QM calculation implicitly accounts for this source of polarisation. The field values that we obtain are in agreement with those mentioned by Boxer and co-workers.[31] In fact, these values are within the range of magnitudes we have contemplated for the models (see below), so that the effects of an external electric field that we have observed in the models may well play a similar role in the course of an enzymatic reaction.

The field created by an enzyme, as opposed to many surface catalysis, is not constant,[32] because enzymes are flexible structures.[33,34] This can be used by the enzyme to improve catalysis. Detailed description of the field effect during a turnover cycle should take enzyme's floppiness into account, but this is beyond the scope of this work.

3.2 Geometries

3.2.1 DFT methods. We have only considered the effect of the electric field in the apical direction since we have found previously that the effect in other directions is not so relevant.[10]

Fig. 2 displays the dependence of the two apical bond lengths on the intensity of the electric field. In the range of distances before bond cleavage, it is found a linear relationship between the apical bond distance (d) and the intensity of the field (F) which is given by the equation $d = aF + b$, where a corresponds to the field dependence and b to the distance without field, and the results obtained with the different basis sets are depicted in Table 1. Taking as a reference the results obtained with the larger 6-311 + G(3df,3pd) basis set, a quick look at Table 1 shows that at least a 6-31 + G(d) basis set is needed to correctly describe the field dependence.

For model **1**, the role of polarisation and diffuse function is different and complementary. The former are needed for a good description of the zero-field geometries (correct b and a too low) and the latter are needed for the field dependence (correct a but b too high). This simple description cannot be generalized to model **2**, probably because of the different origin of the polarisation of the P atom. In model **1** the polarisation arises from the conformation of the phosphoryl moiety itself, and not from the apical substitutes, which are the same. So, even though there is a poor description of the apical bond, this wrong description is the same for both bonds and the errors cancel out, having a lesser influence on the field dependence, giving an independent effect of diffuse and polarisation functions. On the other hand, in model **2** the source of polarisation is caused by the different apical groups. It is thus expectable that a wrong description of these bonds leads to a

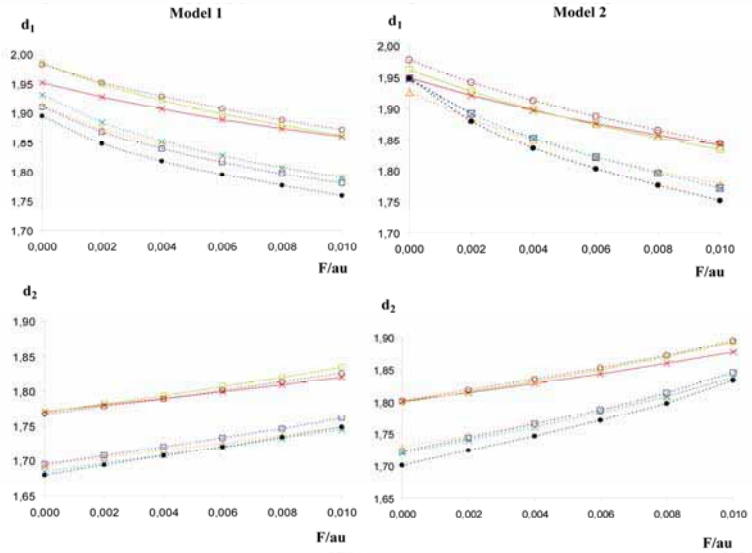


Fig. 2 Representation of the dependence of both apical bond distances (in Å) on the electric field (F) for models **1** and **2**, where d_1 corresponds to the weakest apical bond and d_2 the strongest one, for different basis sets (6-31 + (d): dashed line with squares, 6-31G(d): dashed line with triangles, 6-31 + G: solid line with squares, 6-31G: solid line with crosses, 6-311G: dashed line with empty circles, 6-311G(d): dashed line with crosses, 6-311 + G(3df,3pd): dashed line with filled circles).

wrong description of the field polarisation effects. Therefore, because d functions are necessary to describe the polarisation effects,[10] diffuse functions alone cannot reproduce the field effects, and the errors they give are much higher relative to model **1**. The different origin in the polarisation can also be seen in the different field dependence, which is almost twice as strong in model **2**.

The apparent lack of influence of diffuse functions on zero-field geometries is also misleading. In fact, diffuse functions are unnecessary on the P atom, as its absence leads practically to the same field dependence (a = 6.50) obtained for the 6-31 + G(d) basis set. Diffuse functions are necessary on oxygen atoms with compensating effects. On equatorial oxygens, they allow them to bear a larger negative charge, increasing the positive charge on the phosphorus and, thus, decreasing apical bond distances. On apical atoms, diffuse functions increase the size of the oxygen atom, which leads to a lengthening of the apical bonds. Both effects cancel out in our models, but in biological systems where equatorial atoms bear different charges and can coordinate metals this will surely not be the case. This is a good example of how a basis set study can provide us valuable insights on the nature of molecular interactions.

Table 1 Parameters of the linear regression of the d_2 bond length dependence on the external electric field for several split-valence basis sets

	Model **1**		Model **2**	
Basis set	*a*	*b*	*a*	*b*
6-31G	4.83	1.771	6.97	1.800
6-31 + G	6.39	1.769	8.40	1.798
6-311G	5.81	1.766	8.52	1.800
6-31G(d)	5.44	1.693	9.59	1.727
6-31 + G(d)	6.57	1.694	10.80	1.722
6-311G(d)	5.99	1.684	10.49	1.719
6-311G(3df,3pd)	6.86	1.679	11.65	1.700

The flexibility to describe the correct field dependence can also be obtained using a triple zeta basis set, so that the results of 6-311G and 6-31 + G are strikingly similar, albeit the 6-311G has a higher cost in terms of the number of basis functions (98 *vs.* 90 in model **1**). The same is true for the similarity of 6-311G(d) and 6-31 + G(d). This agreement occurs in both models. Thus 6-31 + G(d) seems to be a fair choice.

As we can see in Fig. 2, the apical bonds d_1 and d_2 of the intermediate undergo significant changes (0.130 and 0.066 Å, respectively) in the range of field magnitudes we have considered. However, considering the TS apical distances in model **1** (Table S2), it is remarkable the stronger variation of d_1 compared to d_2 (0.155 *vs.* 0.025 Å). In fact, this is a consequence of the more covalent character of d_2, which is far less polarizable than the dative d_1, in the transition state.

3.2.2 Performance of semi-empirical methods. Some methods cannot describe the competition between the ligands and the polarisation effects they induce, and this fact leads to wrong equilibrium geometries and energies. AM1, PM3, and MNDO/d show a poor performance, and the best ones are the more recent AM1/d-PhoT and PM6 as depicted in Fig. 3.

In particular, the widely used AM1 tends to strongly overbind both axial ligands, thus neglecting the competitive effect, because of its generally overstabilising core–core interaction in

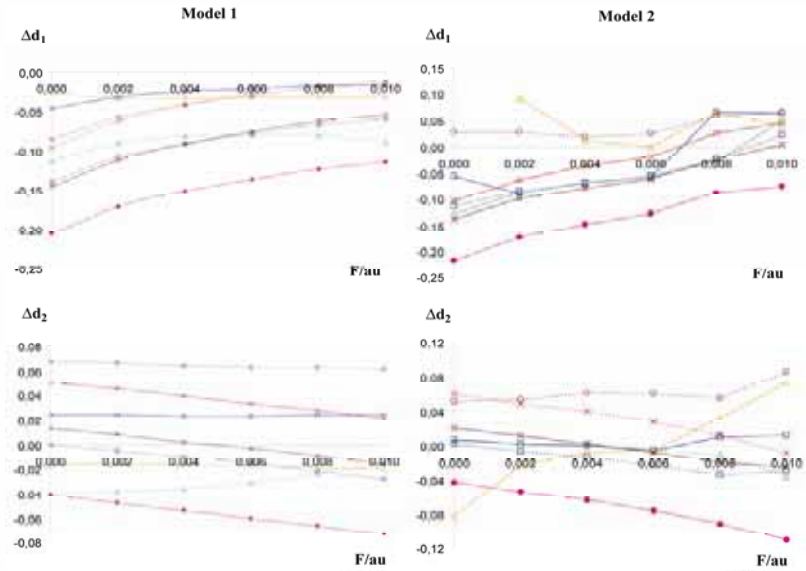


Fig. 3 Representation of the error (difference with respect to DFT) in the semi-empirical apical bond distances (in Å) along the electric field (F) for models **1** and **2**, where d_1 corresponds to the weakest apical bond and d_2 the strongest one (AM1: solid line with dots, PM3: dashed line with crosses, RM1: dashed line with triangles, AM1/Arantes: dashed line with squares, MNDO/d: dashed line with empty circles, AM1*: solid line with crosses, AM1/d-PhoT: solid line with triangles, PM6: solid line with squares). DFT results (dashed line with filled circles) are obtained by means of the *m*PW1PW91 functional and the 6-31 + G(d) basis set.

hypervalent molecules.[19] Newer parametrizations using the same Hamiltonian (AM1/Arantes, RM1) considerably improve AM1. Among the recent methods that include d orbitals, AM1* does not lead to significant improvements, whereas PM6 and AM1/d-PhoT approach surprisingly well the DFT results.

In Fig. 3 we can observe that semi-empirical methods have more difficulty in describing the longest apical bond (d_1). This is reflected in the greater variability obtained for the d_1 bond length with respect to d_2.

With regard to model **1**, if we now consider the field dependence of d_2 (Fig. S2), we see that in general d functions are needed to describe the slope of the distance variation and that methods without it tend to underestimate this field dependence (RM1 being the only clear exception). Some methods, such as AM1/d-PhoT and PM6 give remarkably good results. Surprisingly, while DFT methods need the incorporation of diffuse or triple zeta functions to describe a good field dependence, these semi-empirical methods attain good results with only a polarisation function.

The situation is more complex for model **2**. As before, AM1 tends to considerably overbind. This is improved in AM1/Arantes and RM1, which give a correct d_2 description but a d_1 which is still too short. In PM3 the competitive effect is not well described because both distances are far too similar. These methods without d orbitals give a fair description of the field variation of d_2, but are less able to describe d_1. Unfortunately the inclusion of d orbitals in this case does not considerably improve the situation. MNDO/d gives a good description on d_1, but it is probably because of its tendency to underbind, as it is seen in the d_2 description. AM1* results are similar to AM1/Arantes and RM1. The good description of PM6 at zero-field is decreased for finite fields, and gives a d_1 dependence which is not better than AM1/Arantes and RM1. AM1/d-PhoT is clearly the best method for large fields, but fails to locate the pentacoordinated structure at zero-field. In fact, as we will see in the next section, the DFT enthalpy of **2** at zero-field is only 0.24 kcal mol^{-1} lower than the enthalpy of **TS2**. From a field of 0.002 au on, AM1/d-PhoT locates the pentacoordinated structure. Finally, it is necessary to point out that the strange behaviour at a field between 0.008 and 0.010 au for almost all methods is due to a rotation of the formic moiety, which does not take place employing the DFT approach.

Taking into account the results obtained for these phosphoranes geometries, we can conclude that AM1/d-PhoT and PM6 are the methods that fit better the DFT field dependence of the geometry.

All these results pinpoint the importance to include electrostatic embedding for the QM part when describing enzyme active sites.[5,7,35–37] Formalisms such as IMOMM[38] should be used carefully in relation to the choice of the semi-empirical method responsible for the description of the interaction between the subsystems. The same limitation applies to ONIOM[39] type methods if the environmental effects are treated with a limited semi-empirical method that cannot respond properly to the polarisation. Thus, results such as those obtained by Webster,[40] who studied the nature of the

pentacoordinated phosphorus in β-FGM26 with the ONIOM method, should be object of discussion.

3.3 Energies

3.3.1 DFT methods. Fig. 4 depicts the field dependence of the height of the bond-breaking barrier for models **1** and **2**. The magnitude of the slope of this variation caused by the field is impressive, since it actually reflects a great sensitivity of the reactivity to the strength of the external field.

As expected, larger geometry changes for a given field in model **2** lead to larger energy changes. In fact, the weakest apical bond is strengthened by the interaction with the electric field and, as a consequence, the molecule is less unstable with respect to the product complex, whose electronic structure is far more insensitive to an external field. Therefore, according to the Hammond postulate, the electric field shifts the geometry of the TS towards the geometry of the product complex leading to larger distances and consequently higher energies, since this geometry changes enhance the charge separation against the external field. This is an important reason why the reaction mechanisms of phosphoryl transfer reactions depend a lot on the environment, and is also an aspect that enzymes can exploit to stabilise the transition state.

The almost linear dependence of the barrier on the electric field is also surprising. Remark that we have considered the first order expansion of the dipole moment $\boldsymbol{\mu}$ with respect to the electric field (eqn (1)):

$$\boldsymbol{\mu} = \boldsymbol{\mu}_0 + \boldsymbol{\alpha}\mathbf{F} \quad (1)$$

where $\boldsymbol{\mu}_0$ is the dipole moment at zero field, $\boldsymbol{\alpha}$ the polarizability and $\mathbf{F}$ the electric field. This gives a quadratic dependence of the energy (eqn (2)):

$$\mathbf{E} = \mathbf{E}_0 - \boldsymbol{\mu}_0\mathbf{F} - \tfrac{1}{2}\mathbf{F}^{\mathrm{T}}\boldsymbol{\alpha}\mathbf{F} \quad (2)$$

The previous equation is exact for a fixed geometry. When a geometry is optimised the energy retains the quadratic character of the field-dependence (see Fig. S1 in the ESI†). However, it is worth pointing out that the linear and quadratic terms do not directly correspond to the previous analytical result (eqn (2)). Indeed, in an optimized structure there are contributions to the energy not only from the interaction with the electric field, but also from changes in the geometry produced by the field.

In principle, on the basis of this quadratic dependence of the energy of an optimized geometry, the field-dependence of the energy barrier should be quadratic (see Fig. 4). However, the *zz*-component of the polarizability tensor is very similar for the TS and the intermediate in model **1** (57.3 and 53.5 au, respectively) and model **2** (100.3 and 97.0 au, respectively). Therefore, the quadratic term tends to cancel out, giving rise to a practically linear energy barrier. The quadratic effects would only show up at higher field values.

Concerning the basis sets, we can point out two results. First, the error given by the basis sets depends on the electric field strength, and, except for the poorer basis sets, it tends to increase for stronger fields. Second, the well-understood behaviour of basis set functions smears out in strong fields, *i.e.*, at low fields, we can understand the importance of polarisation, multiple-zeta and diffuse functions in the completeness of the basis set, with the major contribution arising form polarisation functions; however, at higher fields, the effects seem to be less systematic. Therefore when confronting an enzymatic study, we suggest not to check the adequacy of a basis set *only* for a set of small model reactions, but to include the electrostatic environment on the tests. If this is not possible, it would be appropiate to take a basis set larger than the one that gives satisfactory results for model systems. As a rule of thumb, for all cases the 6-31 + G(d) basis set gives results that are below 1 kcal mol^{-1} of error with respect to the reference 6-311 + G(3df,3pd).

3.3.2 Performance of semi-empirical methods

Reaction 1. The energetic results for semi-empirical methods are shown in Fig. 5. Only three methods give reasonable results, with an error less than 15 kcal mol^{-1}: PM6, AM1/d-PhoT and AM1/Arantes. The first two are remarkably good, and AM1/d-PhoT gives a correct field dependence of the energy barrier. The rest of the methods give enthalpies at zero-field that have an error larger than 25 kcal mol^{-1}, so that a more detailed analysis of their field dependence is

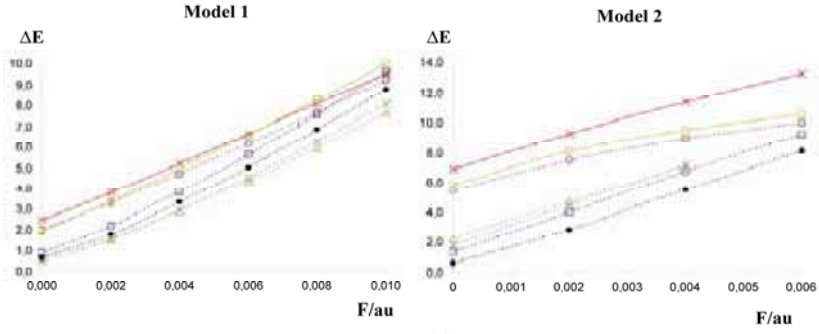


Fig. 4 Representation of the dependence of the energy barrier (in kcal mol^{-1}) of the weakest apical bond cleavage on the electric field for models **1** and **2**, for different basis sets (6-31 + G(d): dashed line with squares, 6-31G(d): dashed line with triangles, 6-31 + G: solid line with squares, 6-31G: solid line with crosses, 6-311G: dashed line with empty circles, 6-311G(d): dashed line with crosses, 6-311 + G(3df,3pd): dashed line with filled circles). The TS could not be located for the 6-31G(d) and the 6-311G(d) at a field of 0.006 au in model **2**.

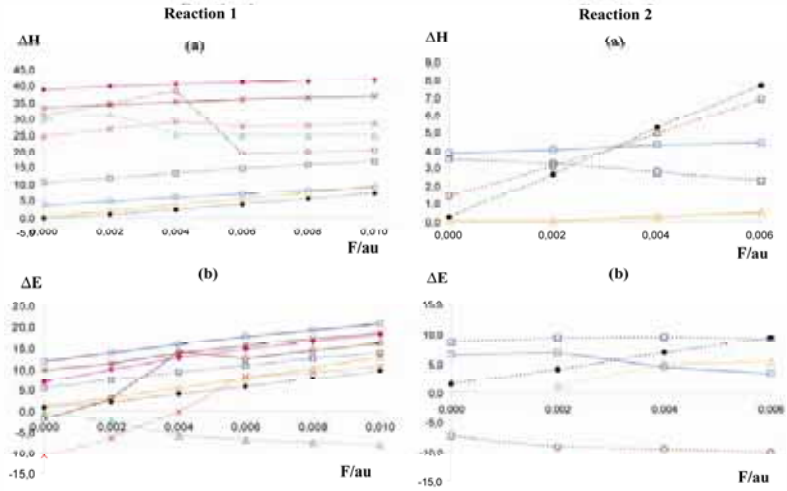


Fig. 5 Representation of the energy barrier (in kcal mol^{-1}) for reactions 1 and 2 calculated with semi-empirical methods (AM1: solid line with dots, PM3: crosses, RM1: dashed line with triangles, AM1/Arantes: dashed line with squares, MNDO/d: dashed line with empty circles, AM1*: solid line with crosses, AM1/d-PhoT: solid line with triangles, PM6: solid line with squares) (a) and with DFT single-point calculations at semi-empirical geometries (b). DFT results (dashed line with filled circles) are obtained by means of the *m*PW1PW91 functional and the 6-31 + G(d) basis set.

meaningless. Besides, RM1, MNDO/d and PM3 show the undesirable artefact of changing the nature of the transition state at relatively low fields, which causes the bump in the energy profile. It is also worth pointing out that the DFT zero-field enthalpy of the TS is lower than that of **1**, which strictly speaking means that the correct result for semi-empirical methods would be not to locate a stable species **1** at zero-field. They all fail in that, but AM1/d-PhoT and PM6 give a small energy barrier, below 3.5 kcal mol^{-1}.

In a biochemical application the usual procedure would be to do single point calculations with a higher level method to validate the semi-empirical results. In such a situation, the semi-empirical geometry is more relevant than the energy. The single point DFT energies are plotted in Fig. 5. As expected, RM1, MNDO/d and PM3 give incorrect trends but the rest of the methods have a reasonable field dependence, and the errors are below the ones given by their enthalpy results. AM1/d-PhoT continues to be the more reliable method, which means that it produces very good geometries. AM1/Arantes is also pretty good, whereas it is surprising to see that PM6 gives the worst results, so that it is the only case where PM6 energies are much better than DFT ones based on a semi-empirical geometry. This is even more unexpected since PM6 gives very good geometries for the field dependence of **1** (Fig. 2) which means that the source of error must be the TS geometry. In fact, a reasonably good description of the TS geometry is challenging for semi-empirical methods in general, since its long-range interactions require bigger basis sets than the minimal one and an adequate core-core repulsion function.

Reaction 2. This reaction turned out to be more challenging than the previous one. The reason is that some of the methods where unable to locate their TSs. We made several unsuccessful attempts with the final DFT geometry and other starting guesses. AM1, PM3 and RM1 always give a transition state where both OH point at the same direction. This TS connects with a phosphorane where, again, both OH point at the same direction (Fig. 6). This is a severe limitation. Not only do they fail in the description of the correct mechanism, but also they give a mechanism that goes through a pentacoordinated species that is unstable at the DFT level, so that further comparison of their performance in this reaction is pointless.

As we have already discussed, AM1/d-PhoT does not locate the pentacoordinated intermediate at zero field in contrast to DFT and the rest of semi-empirical methods, which give small energy barriers (with the exception of AM1*). However, the interaction with the electric field gives rise to a transition state for the bond breaking process, as expected. In this model, the dependence of the barrier on the field is rather poor for all methods (Fig. 5a), since they are practically insensitive to the field, with the surprising exception of MNDO/d taking into account its bad performance in the previous model. Moreover, AM1* changes the mechanism at low fields, so that further comparison with DFT is unnecessary.

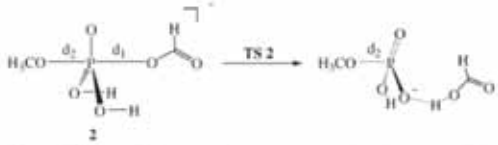


Fig. 6 Scheme of the anomalous pentacoordinated intermediate obtained by AM1, PM3 and RM1 in reaction 2.

The results of DFT single point calculations (Fig. 5b) are as disappointing as the previous ones with regard to the sensitivity of the methods to the field. In addition, the performance of MNDO/d got worse drastically in contrast to the outstanding trend showed before. Nevertheless, AM1/d-PhoT is the unique semi-empirical method still able to approach the tendency of DFT although with less success than in reaction 1.

4. Conclusions

There are three main sources of the P polarisation. First, an internal polarisation due to the orientation of the substituents of the phosphoryl moiety. Second, the electrodonor ability of the axial groups, and last but not least, the electrostatic environment. The fields created by enzymes lead to the remarkable fact that the three effects in phosphorane structures have similar strengths and effects, both in energy barriers and geometries. Therefore the correct description of all these effects is essential to understand the source of enzyme catalysis and differenciate the electrostatic effects in the active site with those in water.

The basis sets used for traditional calculations give lower quality results in the presence of an electric field, and, thus need to be extended. A minimally reasonable choice for large systems is the 6-31 + G(d) basis set.

Semi-empirical methods have a wide range of performances. Considering the overall performance for geometries, PM6 and AM1/d-PhoT give very good results. AM1/Arantes and RM1 are also reasonably good and are clear improvements over AM1, PM3 or MNDO/d. The behaviour of AM1* was not as good as expected. For the transition states, the same trend is valid, with the exception of PM6, which in one case gave wrong geometries for the transition state. Thus, single point calculations with higher level methods are a valid procedure to improve semi-empirical results, except for PM6. The particular misbehaviour of PM6 for one reaction cannot be generalized but needs further investigations. In general, AM1/d-PhoT stands as the best method to describe pentacoordinated species and their reactivity.

Electric field effects should become valuable data in future semi-empirical parametrizations for phosphorus in order to study enzymatic phosphoryl transfer reactions with more accuracy. It represents a challenging but simple analogue of what the chosen method should be able to reproduce in an enzyme active site.

Acknowledgements

We would like to thank James Stewart for providing the source code of MOPAC2007 and discussing implementation issues. We acknowledge financial support from the MEC (Grant CTQ2006-01345/BQU and BQU2005-07790) and the Generalitat de Catalunya (Grant 2005SGR00111). RC thanks the Ramón y Cajal programme. EM thanks the JAE programme from the CSIC. This research has been partly performed using the CESCA resources.

References

1 P. Cohen, *Eur. J. Biochem.*, 2001, **268**, 5001–5010.
2 F. H. Westheimer, *Science*, 1987, **235**, 1173–1178.
3 C. Lad, N. H. Williams and R. Wolfenden, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 5607–5610.
4 K. N. Allen and D. Dunaway-Mariano, *Trends Biochem. Sci.*, 2004, **29**, 495–503.
5 H. M. Senn and W. Thiel, in *Atomistic Approaches in Modern Biology*, *(Topics in Current Chemistry)*, ed. R. M. Berlin, Springer, 2007, vol. 268, pp. 173–290.
6 M. J. Field, *J. Comput. Chem.*, 2002, **23**, 48–58.
7 H. Lin and D. G. Truhlar, *Theor. Chem. Acc.*, 2007, **117**, 185–199.
8 D. J. Wales and T. V. Bogdan, *J. Phys. Chem. B*, 2006, **110**, 20765–20776.
9 J. J. Ruiz-Pernia, E. Silla, I. Tunon, S. Marti and V. Moliner, *J. Phys. Chem. B*, 2004, **108**, 8427–8433.
10 E. Marcos, R. Crehuet and J. M. Anglada, *J. Chem. Theory Comput.*, 2008, **4**, 49–63.
11 C. Adamo and V. Barone, *J. Chem. Phys.*, 1998, **108**, 664–675.
12 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. G. Johnson, W. Chen, M. W. Wong, C. Gonzalez and J. A. Pople, *GAUSSIAN 03 (Revision C.02)*, Gaussian, Inc., Wallingford, CT, 2004.
13 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902–3909.
14 J. J. P. Stewart, *J. Comput. Chem.*, 1989, **10**, 209–220.
15 G. M. Arantes and M. Loos, *Phys. Chem. Chem. Phys.*, 2006, **8**, 347–353.
16 G. B. Rocha, R. O. Freire, A. M. Simas and J. J. P. Stewart, *J. Comput. Chem.*, 2006, **27**, 1101–1111.
17 W. Thiel and A. A. Voityuk, *J. Phys. Chem.*, 1996, **100**, 616–626.
18 J. J. P. Stewart, *J. Mol. Model.*, 2007, **13**, 1173–1213.
19 K. Nam, Q. Cui, J. L. Gao and D. M. York, *J. Chem. Theory Comput.*, 2007, **3**, 486–504.
20 P. Winget, A. H. C. Horn, C. Selcuki, B. Martin and T. Clark, *J. Mol. Model.*, 2003, **9**, 408–414.
21 J. J. P. Stewart, *MOPAC2007, Stewart Computational Chemistry*, Colorado Springs, CO, USA, 2007, http://OpenMOPAC.net.
22 C. N. Schutz and A. Warshel, *Proteins: Struct., Funct., Genet.*, 2001, **44**, 400–417.
23 M. K. Gilson and B. H. Honig, *Biopolymers*, 1986, **25**, 2097–2119.
24 W. L. De Lano, *The PyMOL Molecular Graphics System*, ed. D. L., Scientific, San Carlos, CA, 2002, http://www.pymol.org.
25 M. J. Field, M. Albe, C. Bret, F. Proust-De Martin and A. Thomas, *J. Comput. Chem.*, 2000, **21**, 1088–1100.
26 S. D. Lahiri, G. F. Zhang, D. Dunaway-Mariano and K. N. Allen, *Science*, 2003, **299**, 2067–2071.
27 J. Y. Choe, C. V. Iancu, H. J. Fromm and R. B. Honzatko, *J. Biol. Chem.*, 2003, **278**, 16015–16020.
28 F. Gil-Ortiz, S. Ramon-Maiques, I. Fita and V. Rubio, *J. Mol. Biol.*, 2003, **331**, 231–244.
29 T. Simonson and C. L. Brooks, *J. Am. Chem. Soc.*, 1996, **118**, 8452–8458.
30 L. I. Krishtalik, A. M. Kuznetsov and E. L. Mertz, *Proteins: Struct., Funct., Genet.*, 1997, **28**, 174–182.
31 I. T. Suydam, C. D. Snow, V. S. Pande and S. G. Boxer, *Science*, 2006, **313**, 200–204.
32 M. K. Prakash and R. A. Marcus, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 15982–15987.

33 E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay and D. Kern, *Nature*, 2005, **438**, 117–121.
34 K. Henzler-Wildman and D. Kern, *Nature*, 2007, **450**, 964–972.
35 M. Strajbl, A. Shurki, M. Kato and A. Warshel, *J. Am. Chem. Soc.*, 2003, **125**, 10228–10237.
36 A. Warshel, P. K. Sharma, M. Kato, Y. Xiang, H. B. Liu and M. H. M. Olsson, *Chem. Rev.*, 2006, **106**, 3210–3235.
37 G. Náray-Szabó, M. Fuxreiter and A. Warshel, in *Computational Approaches to Biochemical Reactivity*, ed. G. Naray-Szabo and A. Warshel, Kluwer Academic Publishers, 2006, pp. 237–293.
38 F. Maseras and K. Morokuma, *J. Comput. Chem.*, 1995, **16**, 1170–1179.
39 M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber and K. Morokuma, *J. Phys. Chem.*, 1996, **100**, 19357–19363.
40 C. E. Webster, *J. Am. Chem. Soc.*, 2004, **126**, 6840–6841.

4.1.3. Pentacoordinated Phosphorus in β-phosphoglucomutase

So far we have studied small compounds of pentacoordinated phosphorus as models of putative intermediates in phosphoryl transfer reactions. The experience acquired in the two previous studies provided some predictive power on the likelihood of formation of stable pentacoordinated phosphorus species, so that the next step was the application of this knowledge to an enzymatic reaction of interest. As mentioned in the introduction, the first report of a X-ray structure of a phosphorane intermediate formed in the course of an enzymatic reaction (β-phosphoglucomutase) was polemical from the very beginning[1]. Several authors questioned the nature of the pentacoordinated species observed in the crystal (Figure 4) and suggested that a MgF_3^- salt mimicking the phosphoryl moiety was present in the structure as a transition state analogue. This was an interesting case for applying our background with the aim to shed light into this controversy.

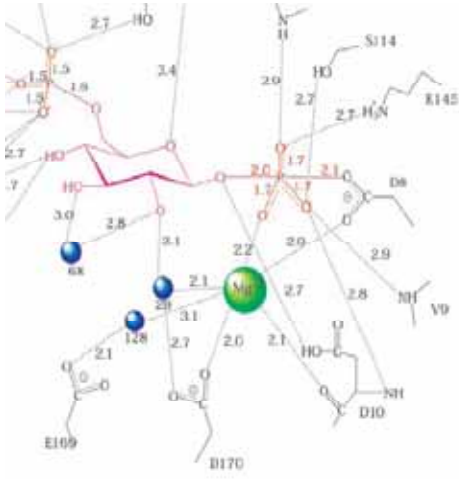


Figure 4. Schematic representation of the X-ray structure of the pentacoordinated intermediate of the β-phosphoglucomutase enzyme[1].

A qualitative and preliminary inspection of the ligands of the putative pentacoordinated phosphorus suggests that the formation of a stable phosphorane was unlikely for two reasons. Firstly, the difference in the electrodonor character between the two apical groups (an alcohol and a carboxylic group bound to a Mg^{2+} cofactor) made questionable the observation in the X-ray structure that the two apical distances are very similar. This skepticism was based on our previous observations in small phosphorane compounds that important differences in the electrodonor character of the two apical groups make the two apical bond distances markedly different. Secondly, in the X-ray structure, the P-O bond distances in the equatorial plane were ~1.7 Å, whereas P-O bonds in phosphoranes are

[1]Lahiri S.D. et al. (2003) *Science* 299: 2067

typically ~1.5 Å in length and Mg-F bond distances are ~ 1.9 Å. This again questioned the identity of the actual species in the structure.

To provide a clear understanding of the reaction mechanism, we calculated the reaction path of the phosphoryl transfer step that may involve the observed phosphorane with the QM/MM formalism. The mechanism turns out to be concerted (not step-wise) with an energy barrier in agreement with kinetic experiments. Furthermore, a MgF_3^- salt in the position of the equatorial PO_3^- fragment was found to be stable exhibiting apical distances very close to those observed in the aforementioned X-ray structure pointing to an efficient transition state analogue. Then, how could the phosphorane be wrongly assigned in the original X-ray structure? To answer this question, we refined the original diffraction map placing a MgF_3^- in the position of the equatorial PO_3^- fragment without constraining its bond distances. The fit was improved with respect to the original one, being the Mg-F bond distances ~ 1.9 Å. This confirms the idea that the anomalous P-O equatorial bond distances of the original X-ray structure resulted from fitting the diffraction map of a MgF_3^- with a phosphoryl moiety, falling the bond distances somewhere in between. Overall, our results confirm the idea that the phosphorane had been wrongly assigned in the original X-ray structure.

A detailed presentation of the results and methodologies used in this study can be found in the article: *Pentacoordinated phosphorus revisited by high-level QM/MM calculations* (2010) Proteins, 78, 2405-2411

PROTEINS

SHORT COMMUNICATION

Pentacoordinated phosphorus revisited by high-level QM/MM calculations

Enrique Marcos,[1] Martin J. Field,[2] and Ramon Crehuet[1]*

[1]Departament de Química Biológica i Modelització Molecular, Institut de Química Avançada de Catalunya (CSIC), Barcelona, Spain
[2]Laboratoire de Dynamique Moléculaire, Institut de Biologie Structurale, CNRS-CEA-UJF, Grenoble, France

ABSTRACT

Enzymes catalyzing phosphoryl transfer reactions are extremely efficient and are involved in crucial biochemical processes. The mechanisms of these enzymes are complex due to the diversity of substrates that are involved. The reaction can proceed through a pentacoordinated phosphorus species that is either a stable intermediate or a transition state (TS). Because of this, the first X-ray structure of a pentacoordinated phosphorus intermediate in the β-phosphoglucomutase enzyme aroused great interest but also much controversy. To provide new insights into the nature of that structure, we have determined the reaction path of the phosphorylation step using high-level QM/MM calculations, and have also calculated the geometry of a complex with a transition state analogue (TSA) that has been suggested to be the actual species in the crystal. The protein crystalline environment has been modeled so as to mimic the experimental conditions. We conclude that the pentacoordinated phosphorus formed in this enzyme is not a stable species but a TS, which gives an activation energy for phosphorylation in agreement with kinetic results. We also show that the TSA is a good mimic of the true TS. We have performed a new crystallographic refinement of the original diffraction map of the pentacoordinated phosphorus structure with the MgF_3^- TSA. The new fit improves significantly with respect to the original one, which strongly supports that Allen and coworkers wrongly assigned the X-ray structure to a pentavalent phosphorane.

Proteins 2010; 78:2405–2411.

Key words: enzyme catalysis; NEB; pentacoordinated phosphorus; phosphorylation; QM/MM; phosphoglucomutase.

INTRODUCTION

The phosphoryl transfer reaction is ubiquitous in nature.[1] However, its mechanism is still the subject of debate because nucleophilic substitution in phosphorus is more complex than in carbon or other first row elements.[2] Understanding the mechanism of this reaction and the elucidation of possible intermediates is crucial for the design of enzyme inhibitors, which often provide the starting points for the development of new drugs.[3]

In 2003, the first X-ray structure of a pentacoordinated phosphorus intermediate was published.[4] It corresponded to a phosphoryl transfer between a glutamate residue and β-glucose-6-phosphate (G6P) in the β-phosphoglucomutase enzyme (β-PGM) to give β-glucose-1,6-bisphosphate (β-G16BP). This structure was highly controversial and it stimulated several comments[5,6] and both computational[7,8] and experimental[9–12] studies. In particular, the existence of a phosphoryl moiety in the crystal was questioned. It was suggested that the position of the PO_3^{2-} fragment was actually occupied by an MgF_3^- ion formed under crystallization conditions.[5] This was supported by Webster's calculations,[7] which showed that the pentacoordinated species was the transition state (TS) separating the tetracoordinated reactant and product phosphates. These results were subsequently challenged by Allen and coworkers,[9] who claimed that a phosphoryl was present in the active site based on quantitative analytical methods. Subsequent NMR[10,11] and kinetic[12] experiments

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: Spanish MEC; Grant number: CTQ2009-08223; Grant sponsor: Catalan AGAUR; Grant number: 2005SGR00111; Grant sponsor: JAE Program of the Consejo Superior de Investigaciones Científicas (CSIC)
*Correspondence to: Ramon Crehuet, Departament de Química Biológica i Modelització Molecular, Institut de Química Avançada de Catalunya (CSIC), c/Jordi Girona 18-26, E-08034 Barcelona, Spain. E-mail: ramon.crehuet@iqac.csic.es
Received 16 December 2009; Revised 15 April 2010; Accepted 19 April 2010
Published online 28 April 2010 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22758

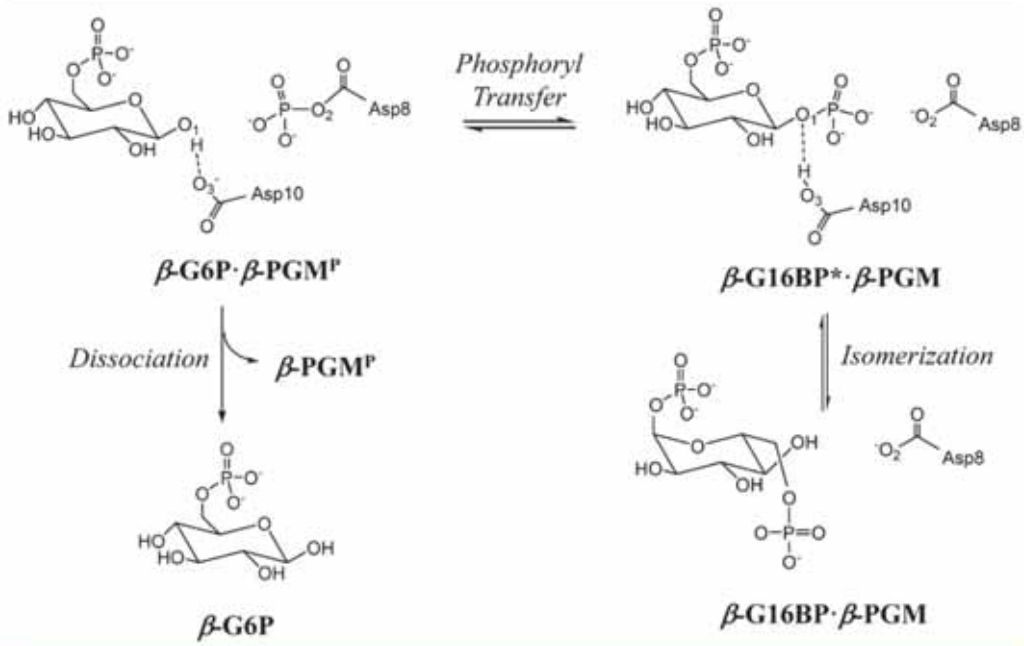


Figure 1

Reaction scheme for the reaction of β-PGM complexed with β-G16BP.

in solution carried out by Waltho and coworkers eventually confirmed that the MgF_3^- ion is a transition state analogue (TSA) that replaces the phosphoryl moiety. These studies raised the question of how well the TSA mimics the geometry and interactions of the true TS.[12] The kinetic study[12] determined the overall rate for the process shown in Figure 1 in which β-G16BP, which can be in two conformations in equilibrium, phosphorylates the enzyme to render a β-G6P·β-PGM^P complex that subsequently undergoes a fast dissociation. The measured rate constant involved the phosphorylation step and product complex dissociation, but whether it also involved isomerization between the equilibrium conformations of the β-G16BP·β-PGM complex could not be ascertained from the kinetic experiments. Likewise the identity of the rate limiting step could not be determined. The activation energy calculated by Webster[7] was much higher than that indicated by the measured kinetic data for the set of steps that included phosphoryl transfer. Was this discrepancy due to the accuracy of the theoretical models employed or is the crystal structure different from the conformation in solution within which the phosphoryl transfer takes place?

In this communication, we revisit the mechanism of the phosphoryl transfer catalyzed by the enzyme β-PGM and compare it with the geometry of the MgF_3^- complex. We report the results of hybrid QM/MM reaction path calculations and put special emphasis on the accuracy of the techniques employed. This is also a feature of the work of other authors. Several studies[13–23] have stated the importance of using very precise quantum chemical methods to describe the kinetic and thermodynamic stability of potential pentacoordinated intermediates. Moreover, Warshel and coworkers[18–23] have also highlighted the need to analyze the free energy surface to describe the reaction paths of phosphoryl transfers. In addition, we have also tried to ensure that the electrostatic effect of the crystal environment on the active site region is properly represented.

MATERIALS AND METHODS

The coordinates of the enzyme were obtained from the crystallographic structure with PDB id.1O08. Hydrogen atoms were added to the structure with standard protonation states at pH = 7, and Asp8 and Asp10 were protonated according to Fig. 1. Geometry optimizations were then performed on the structure. In these, the positions of the atoms were constrained with harmonic forces

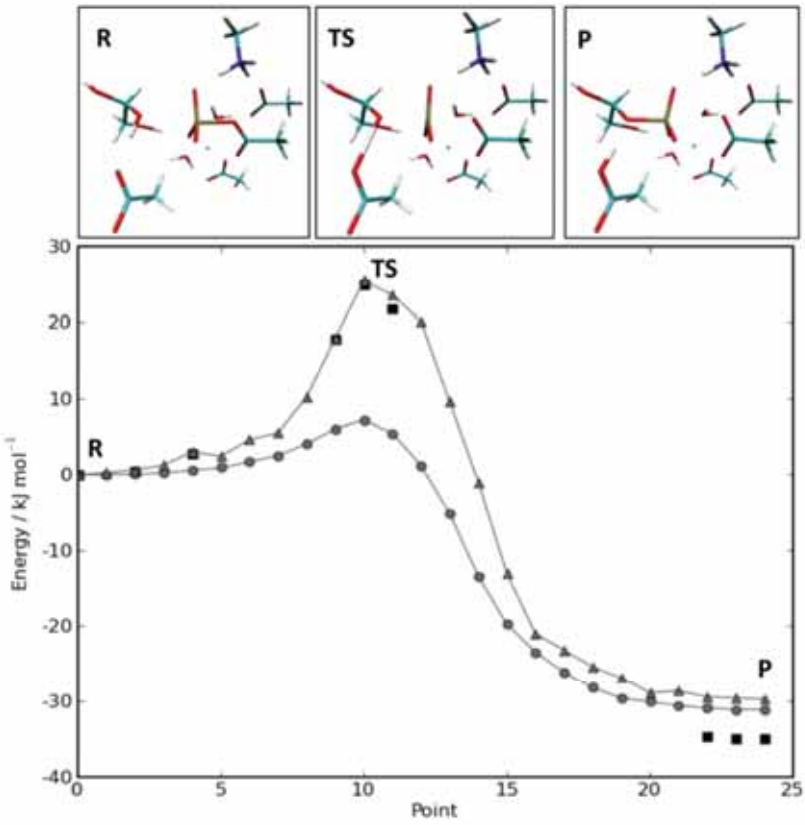


Figure 2

The most reliable energy profile in this work was obtained at the levels of theory *m*PW1PW/b2/q2 (triangles), and SCS-MP2/b2/q2 (squares) with the NEB method. For the latter, only the structures around the stationary points were calculated. More calculations that illustrate convergence to this final profile can be found in the supporting information. The *x*-axis corresponds to structures along the reaction path separated by a constant length. The structures of the QM region of the reactant, transition state, and product are also displayed (the atomic coordinates are given in the Supporting Information).

that were progressively relaxed. We did not perform molecular dynamics simulations as we wanted the structure to remain as close as possible to the experimental one.

For the QM/MM calculations, the system was partitioned between QM and MM regions as depicted in Figure 4. A small QM region was used for the geometry optimizations and nudged elastic band (NEB) calculations. This is denoted q1 in the text and had 58 atoms. A larger QM region, q2, with 94 atoms was employed for single point calculations to verify the results with q1.

A comparison and assessment of the different methods can be found in the Supporting Information. The best estimates are obtained with methods **m7** and **m8**, and these are the ones that are discussed hereafter.

The OPLS-AA force field[24] was employed for the MM region whereas the SCS-MP2[25] and different DFT methods were used for the QM region. Geometry optimizations and NEB calculations were performed with the *m*PWPW functional,[26] which has been parameterized to describe noncovalent interactions. Our previous calculations[16] have shown the good performance of the hybrid version of this functional (*m*PW1PW)[26] in describing the geometry and energetics of pentacoordinated phosphorus compounds. The results of this work (see Supporting Information) show that the pure and hybrid DFT functionals give similar geometries but that the calculations for the former are much faster because of the use of the resolution of the identity approximation. The

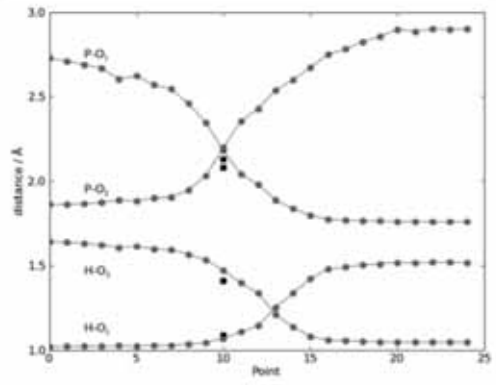


Figure 3

Variations of some relevant distances along the reaction path (gray circles). We also depict the analogous distances in the MgF_3^- TSA complex (black squares). The atomic coordinates of the TSA complex are given in the Supporting Information.

basis set used for the optimization was Ahlrichs-VDZ[27] with polarization functions on nonhydrogen atoms, and a diffuse function on oxygen.[28] It is denoted b1 in the text. Single point calculations were performed with b1 and also a larger Ahlrichs-VTZ[27,29] basis with polarization and diffuse functions on all atoms.[28,30] This is denoted b2. All the system setup and the QM/MM calculations were performed with the pDynamo software[31] using a modified version of the interface to the ORCA program,[32–34] which was employed for the QM calculations.

The reaction path was optimized with the NEB method[35–38] implemented in pDynamo. The utility of NEB can be seen from inspection of Figure 3 in which four distances change asynchronously along the reaction path. Such behavior is obtained automatically in a NEB calculation, but is difficult or impossible to reproduce using a predefined set of reaction coordinate variables.

The MgF_3^- TSA complex was constructed by substitution of the PO_3^- moiety in both reactant and product structures. Subsequent optimizations revealed a convergence to a unique structure as discussed below. In the crystallographic refinement, the four new atoms were treated independently as unbound atoms. A new structure was obtained after several cycles of anisotropic B_{factor} restrained refinement using the program REFMAC[39] yielding a final improved R_{factor} of 12.8%, and an R_{free} of 16.9%. The final 2Fo-Fc electron density map fitting can be seen in Supporting Information Figure S3.

For the crystal calculations, the periodic images of the molecule were generated with Pymol[40] (see Supporting Information Fig. S1). OPLS-AA MM charges on the atoms were used to represent the electrostatic field of the first-shell protein molecules, whereas for more distant periodic copies of the protein a dipolar representation was employed. Full details are given in the Supporting Information.

RESULTS AND DISCUSSION

Our results indicate that the pentacoordinated phosphorus is not a stable species. The complete energy profile that we calculate for the reaction is given in Figure 2, and the reactant and product species are depicted in Figure 1. Although we looked for additional intermediates, including those with different protonation states, that could lead to an alternative stepwise mechanism,[21] the only two stable species that are predicted to occur are those depicted in Figure 1 (phosphorylation step). In relation to this, Waltho and coworkers showed that the β-PGM enzyme prioritizes anionic charge over geometry in aluminum and magnesium fluoride TS analogs.[11] Therefore, the deprotonated form of the phosphoryl moiety, which bears a −1 formal charge, will be preferred over other protonation states. In support of this, we have been unable to locate a stable pentacoordinated intermediate, either with the deprotonated form or with forms in which different oxygen of the phosphoryl moiety are protonated.

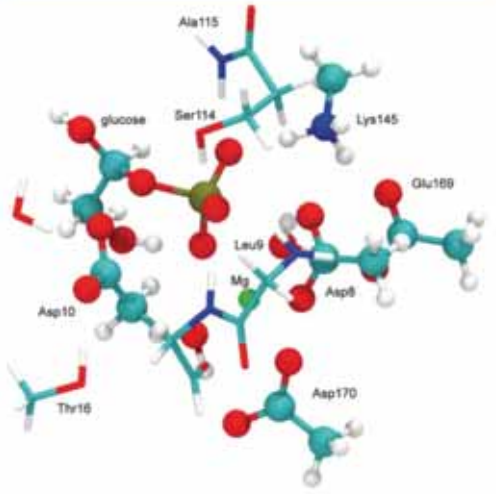


Figure 4

QM regions used in this work: q1 (ball and stick) and the larger q2 (sticks only). The MM region is constituted by all the remaining atoms in the crystal structure, including the water molecules. No additional water molecules were added. Environmental effects due to the crystal were simulated by the addition of periodic protein images as described in the main text and Supporting Information. This figure was created with VMD.[41]

In the reactants, the P atom is bound to the O2 carboxylic oxygen of the aspartate residue (Asp8) at a distance of 1.86 Å, whereas the glucose O1 atom is at 2.73 Å from P. This cannot be considered a pentacoordinated structure as it differs significantly from the crystal structure, in which the P—O1 and P—O2 distances are 2.0 and 2.1 Å, respectively. Likewise, the P atom is not pentacoordinated in the products, as it is bound to glucose O1 at a distance of 1.76 Å (the latter having been deprotonated by Asp10) and is at 2.90 Å from the Asp8 oxygen. The shorter P—O distance in the products is because hydroxylate is a better nucleophile than carboxylate.[16] This, together with the nature of the proton-accepting group, accounts for the exothermicity of 35.0–29.7 kJ mol^{-1} that we observe for the reaction.

Figure 2 shows that the reaction has a single TS with an energy barrier of 24.9–25.5 kJ mol^{-1} with respect to reactants. The TS corresponds to a pentacoordinated species with P—O1 and P—O2 distances of 2.18 and 2.19 Å, respectively (see Fig. 3). Clearly, this species cannot correspond to the one that is observed in the crystal structure as it is unstable.

Figure 3 illustrates the variation in certain interatomic distances along the reaction path and shows that nucleophilic attack occurs before proton transfer. This is better understood by considering the reaction in the reverse direction. The carboxylate group of Asp8 is a weak nucleophile, all the more so when it is coordinated with Mg^{2+}. As such it can only attack the phosphate when the glucose oxygen is protonated, because this hydroxyl is then also a weak nucleophile of similar strength. Figure 3 also shows the interatomic distances for the TSA complex in which an Mg atom replaces P. It is worth emphasizing that the MgF_3 complex and the structure of the TS have strikingly similar bond distances between the central atom and the nucleophile and leaving groups. Remarkably, the distances for the G6P proton that is transferred to Asp10 are also in close agreement with that of the TS. The main difference lies in the Mg—F bond distances, which we predict to be in the range of 1.91–1.94 Å, whereas the equivalent P—O distances in the calculated TS structure are 1.53–1.56 Å. (see Supporting Information Fig. S2 for a comparison between both structures). The reported P—O distances in the PDB are 1.66–1.70 Å. This disagreement led us to analyze the original experimental diffraction map of structure 1O08. We refined the structure using an MgF_3 moiety in place of the PO_3 without constraining its bond distances. The fit is improved using MgF_3 with respect to the original PO_3 (see "Materials and Methods" section and Supporting Information Figure S3). The optimized distances are now in the range of 1.85–1.91 Å, in close agreement with the calculated ones. Therefore, the density map fitting also supports our computational result that MgF_3^- is the species present in the active site.

In the kinetic study by Waltho and coworkers a rate constant was obtained (k = 105 s^{-1}) for the formation of β-G6P from the β-G16BP·β-PGM complex (see Fig. 1), but they were unable to assign this rate constant to a specific step. A direct application of TS theory gives an upper value for the activation free energy for the process of 61.5 kJ mol^{-1}. It is worth reminding the reader that the calculated energy profile shown in Figure 2 corresponds to the phosphoryl transfer step of the inverted process, that is, the formation of β-G16BP·β-PGM from β-G6P·β-PGM[P]. Therefore, if we make the approximation of equating free energy with potential energy, we can compare the energy difference between the TS and products with the experimental energy barrier (61.5 kJ mol^{-1}). This value is in very good agreement with our findings, which gives a result of 55.2–59.9 kJ mol^{-1} using our best two calculation methods, **m7** and **m8**. Thus, we can identify the experimental rate constant as arising from the chemical process of phosphoryl transfer, and conclude that the remaining processes will be faster (or as fast as) this step. Webster's results gave an energy barrier for this process of 76.6 kJ mol^{-1}, which is in less good agreement with the experimental rate constant. It is not obvious to ascertain the reasons for the discrepancy between Webster's result and ours, given the differences in computational procedures between the two studies (e.g., the functional, basis sets, number of degrees of freedom that are minimized, the embedding between the high and low-level parts of the model, etc.). Although a difference of 15 kJ mol^{-1} cannot be considered large, the better agreement of our value with the experimental one and its relative insensitivity to parameter changes (see Supporting Information), provide support for the validity of our computational approach.

Waltho and coworkers also studied the kinetics of inhibition with fluoride and magnesium. They concluded that the ions first enter an open form of the enzyme, which then closes tightly.[12] This open-close motion of the enzyme is important for turnover and product release (see Fig. 1). The good agreement for the phosphoryl transfer activation energy is proof that MgF_3^- is a good TSA, that is, that the structure in the crystal is close to the conformation that forms β-G16BP in solution. The matching geometries of the calculated PO_3^- and MgF_3^- complexes also suggest that the latter can be a better TSA than other commonly used ones, such as vanadate, which sometimes fails to reproduce the same interactions as the actual TS in other kinases.[12] Further agreement with experiment can be obtained with the calculated NMR chemical shifts for the fluorine in the active site: −213, −203, and −195 ppm. Although the absolute values differ, the relative values compare well with the measured ones[10]: −159, −152, and −147 ppm. Again, this confirms that the crystal structure contains MgF_3^- in a conformation close to that in solution. Details of these calculations are reported in the Supporting Information.

We have already stated the importance of having a converged calculation, both in terms of the quality of the

Table I
Relative Energies of Reactants, TS and Products in kJ mol^{-1}

Name	Method[a]	Reactants	TS[b]	Products
m7	*m*PW1PW/b2/q2	0.0	25.5	−29.7
m8	SCS-MP2/b2/q2	0.0	24.9	−35.0
m9	SCS-MP2/b2/q2[c]	0.0	27.8	−21.3
m10	SCS-MP2/b2/q2[d]	0.0	27.2	−24.1

[a]The QM method, the basis set, and the QM region used is shown.
[b]The energy of structure 10 of the reaction path.
[c]Including the point charges of the 14 neighboring molecules.
[d]Same as footnote "c" plus the dipoles of 6897 neighboring molecules.

method used for the quantum mechanical region and in its size. Figure 4 shows the two QM regions that were used, one for the optimization, and the larger one for the best energy profile. Our best results are listed in Table I. Supporting Information discusses the validity of this choice with additional results.

The results discussed so far correspond to calculations with an isolated enzyme molecule. In the crystal, periodic effects can create a different electrostatic environment. Because of the importance of polarizability effects in pentacoordinated phosphorus species, the electrostatic field generated by the crystal could potentially change the energetics of the profile.[16–18] To investigate this point, we included, as a first step, the point charges of the first shell of 14 protein molecules that surround the one under study (**m9**). And, in a second step, we added the dipole moments of the 6897 closest crystal images of the central protein beyond the first shell (**m10**, see Supporting Information for details). Table I shows that these environments do not change significantly the results for a single molecule. We also hypothesize that solvent effects will have a minor influence, because the active site of this molecule is buried far from the solvent.

CONCLUSIONS

In the present investigation, we have analyzed the structure reported by Allen and coworkers[4] in their crystallographic study of the β-phosphoglucomutase enzyme, both with respect to the reaction mechanism of phosphoryl transfer and the geometry of the MgF_3^- TSA complex. To have a proper theoretical description of the reacting species during phosphoryl transfer, we have employed a series of high-level QM/MM simulation models, and we have also used a number of representations of the protein's crystalline environment so as to reproduce the experimental conditions as closely as possible. Our results reveal that the rate limiting step for the production of G6P from the phosphorylated enzyme is the chemical process of phosphoryl transfer, with an activation energy that corresponds well to the experimental rate constant obtained by Waltho and coworkers. Although we conclude that having a stable, pentacoordinated phosphorus intermediate for this enzyme is impossible, we have shown that the MgF_3^- present in the crystal structure is a good TSA that can give insight into the geometry of the phosphoryl transfer TSs. The good agreement between the experimental and calculated energy barrier and chemical shifts supports our conclusion that the crystal structure is equivalent to the closed conformation that binds the β-G16BP in solution.

ACKNOWLEDGMENTS

This work has been supported by grants from the JAE-predoc programme of the Consejo Superior de Investigaciones Cientificas (CSIC), the Spanish MEC (CTQ2009-08223) and the Catalan AGAUR (2005SGR00111). The calculations were performed, in part, with CESCA and CESGA resources. The authors thank Xavier Carpena for help and suggestions in the analysis of the crystallographic data and one of the referees whose comments helped to improve the article.

REFERENCES

1. Cleland WW, Hengge AC. Enzymatic mechanisms of phosphate and sulfate transfer. Chem Rev 2006;106:3252–3278.
2. Allen KN, Dunaway-Mariano D. Phosphoryl group transfer: evolution of a catalytic scaffold. Trends Biochem Sci 2004;29:495–503.
3. Cohen P. The role of protein phosphorylation in human health and disease—delivered on June 30th 2001 at the FEBS Meeting in Lisbon. Eur J Biochem 2001;268:5001–5010.
4. Lahiri SD, Zhang GF, Dunaway-Mariano D, Allen KN. The pentacovalent phosphorus intermediate of a phosphoryl transfer reaction. Science 2003;299:2067–2071.
5. Blackburn GM, Williams NH, Gamblin SJ, Smerdon SJ. Comment on "The pentacovalent phosphorus intermediate of a phosphoryl transfer reaction." Science 2003;301:1184.
6. Allen KN, Dunaway-Mariano D. Response to comment on "The pentacovalent phosphorus intermediate of a phosphoryl transfer reaction." Science 2003;301:1184.
7. Webster CE. High-energy intermediate or stable transition state analogue: theoretical perspective of the active site and mechanism of beta-phosphoglucomutase. J Am Chem Soc 2004;126:6840–6841.
8. Berente I, Beke T, Naray-Szabo G. Quantum mechanical studies on the existence of a trigonal bipyramidal phosphorane intermediate in enzymatic phosphate ester hydrolysis. Theor Chem Acc 2007; 118:129–134.
9. Tremblay LW, Zhang GF, Dai JY, Dunaway-Mariano D, Allen KN. Chemical confirmation of a pentavalent phosphorane in complex with beta-phosphoglucomutase. J Am Chem Soc 2005;127:5298–5299.
10. Baxter NJ, Olguin LF, Golicnik M, Feng G, Hounslow AM, Bermel W, Blackburn GM, Hollfelder F, Waltho JP, Williams NH. A Trojan horse transition state analogue generated by MgF3-formation in an enzyme active site. Proc Natl Acad Sci USA 2006;103:14732–14737.
11. Baxter NJ, Blackburn GM, Marston JP, Hounslow AM, Cliff MJ, Bermel W, Williams NH, Hollfelder F, Wemmer DE, Waltho JP. Anionic charge is prioritized over geometry in aluminum and magnesium fluoride transition state analogs of phosphoryl transfer enzymes. J Am Chem Soc 2008;130:3952–3958.
12. Golicnik M, Olguin LF, Feng GQ, Baxter NJ, Waltho JP, Williams NH, Hollfelder F. Kinetic analysis of beta-phosphoglucomutase and its inhibition by magnesium fluoride. J Am Chem Soc 2009;131: 1575–1588.

13. Elsasser B, Valiev M, Weare JH. A dianionic phosphorane intermediate and transition states in an associative A(N)+D-N mechanism for the ribonuclease A hydrolysis reaction. J Am Chem Soc 2009;131:3869–3871.
14. de Vivo M, Dal Peraro M, Klein ML. Phosphodiester cleavage in ribonuclease H occurs via an associative two-metal-aided catalytic mechanism. J Am Chem Soc 2008;130:10955–10962.
15. Range K, McGrath MJ, Lopez X, York DM. The structure and stability of biological metaphosphate, phosphate, and phosphorane compounds in the gas phase and in solution. J Am Chem Soc 2004;126:1654–1665.
16. Marcos E, Crehuet R, Anglada JM. Inductive and external electric field effects in pentacoordinated phosphorus compounds. J Chem Theory Comput 2008;4:49–63.
17. Marcos E, Anglada JM, Crehuet R. Description of pentacoordinated phosphorus under an external electric field: which basis sets and semi-empirical methods are needed? Phys Chem Chem Phys 2008;10:2442–2450.
18. Klahn M, Rosta E, Warshel A. On the mechanism of hydrolysis of phosphate monoesters dianions in solutions and proteins. J Am Chem Soc 2006;128:15310–15323.
19. Rosta E, Kamerlin SCL, Warshel A. On the interpretation of the observed linear free energy relationship in phosphate hydrolysis: a thorough computational study of phosphate diester hydrolysis in solution. Biochemistry 2008;47:3725–3735.
20. Kamerlin SCL, Florian J, Warshel A. Associative versus dissociative mechanisms of phosphate monoester hydrolysis: on the interpretation of activation entropies. Chem Phys Chem 2008;9:1767–1773.
21. Kamerlin SCL, Williams NH, Warshel A. Dineopentyl phosphate hydrolysis: evidence for stepwise water attack. J Org Chem 2008;73:6960–6969.
22. Kamerlin SCL, Haranczyk M, Warshel A. Are mixed explicit/implicit solvation models reliable for studying phosphate hydrolysis? A comparative study of continuum explicit and mixed solvation models. Chem Phys Chem 2009;10:1125–1134.
23. Kamerlin SCL, McKenna CE, Goodman MF, Warshel A. A computational study of the hydrolysis of dGTP analogues with halomethylene-modified leaving groups in solution: implications for the mechanism of DNA polymerases. Biochemistry 2009;48:5963–5971.
24. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
25. Grimme S. Improved second-order Moller-Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies. J Chem Phys 2003;118:9095–9102.
26. Adamo C, Barone V. Exchange functionals with improved long-range behavior and adiabatic connection methods without adjustable parameters: the mPW and mPW1PW models. J Chem Phys 1998;108:664–675.
27. Schafer A, Horn H, Ahlrichs R. Fully optimized contracted gaussian-basis sets for atoms Li to Kr. J Chem Phys 1992;97:2571–2577.
28. Krishnan R, Binkley JS, Seeger R, Pople JA. Self-consistent molecular-orbital methods. 20. Basis set fro correlated wave-functions. J Chem Phys 1980;72:650–654.
29. Weigend F, Ahlrichs R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy. Phys Chem Chem Phys 2005;7:3297–3305.
30. McLean AD, Chandler GS. Contracted gaussian-basis sets for molecular calculations. I. 2nd row atoms, $Z = 11–18$. J Chem Phys 1980;72:5639–5648.
31. Field MJ. The pDynamo program for molecular simulations using hybrid quantum chemical and molecular mechanical potentials. J Chem Theory Comput 2008;4:1151–1161.
32. Neese F. ORCA—an ab initio, Density Functional and Semiempirical program package, version 2.6. Bonn: University of Bonn; 2008.
33. Neese F, Wennmohs F, Hansen A, Becker U. Efficient, approximate and parallel Hartree-Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree-Fock exchange. Chem Phys 2009;356:98–109.
34. Neese F. An improvement of the resolution of the identity approximation for the formation of the Coulomb matrix. J Comput Chem 2003;24:1740–1747.
35. Jónsson H, Mills G, Jacobsen KW. Nudged elastic band method for finding minimum energy paths of transitions. In: Berne BJ, Ciccotti G, Coker DF, editors. Classical and Quantum Dynamics in Condensed Phase Simulations. Singapore: World Scientific; 1998. pp 385–404.
36. Henkelman G, Jónsson H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. J Chem Phys 2000;113:9978–9985.
37. Crehuet R, Field MJ. A temperature-dependent nudged-elastic-band algorithm. J Chem Phys 2003;118:9563–9571.
38. Galvan IF, Field MJ. Improving the efficiency of the NEB reaction path finding algorithm. J Comput Chem 2008;29:139–143.
39. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr Sect D-Biol Crystallogr 1997;53:240–255.
40. DeLano WL. The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific; 2002.
41. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14:33–38.

SUPPORTING INFORMATION

QM/MM calculations

In this section we provide a detailed description of the methods used and justify our confidence in the best results, reported in the main text. Table S1 displays the results. All calculation codes **m1–10** in this section refer to this table.

Choice of functional:

The *m*PWPW functional underestimates the energy barrier, as can be seen in **m1**. Optimisation with its hybrid version, *m*PW1PW, gives geometries that are very close (see table S2) and an energy difference that is very similar to the one obtained using the *m*PWPW geometry (cf. **m2** with **m3**).

Choice of basis set:

Expansion of the basis set from b1 to b2 with DFT methods does not change considerably the energy barrier (Compare **m2** with **m4**). It is also well known that geometries are less sensitive to basis sets than energies. Ab initio correlated methods, such as MP2, do need larger basis sets, and the b1 result shows a significant discrepancy with the b2 result, justifying the use of the larger basis set for these calculations. **m6** therefore is much more reliable than **m5**, the latter being displayed for the sake of coherence.

Comparison of **m2** and **m4** shows that a larger basis set does not render the pentacoordinated structure stable enough to become an intermediate, and thus, that the inability to locate a phosphorane is not a limitation of the basis set. In our previous work, we showed that a split-valence basis set with polarisation functions on oxygens and phosphorus gave correct geometries for stable phosphoranes and transition states[1].

Choice of QM region:

Expanding the QM region to include further residues will improve the result, but the numbers do not change significantly. Compare **m4** with **m7**, and **m6** with **m8**.

Overall:

Our best methods are **m7** and **m8**. We have shown that the error for each method arising from the basis set and the QM region is around 1-2 kJ/mol. The barriers for **m7** and **m8** also differ by a similar amount (1.4 kJ/mol difference).The exothermicity is slightly larger for SCS-MP2 (-29.7 *vs* -35.0 kJ/mol). SCS-MP2 is considered a more reliable method but we have no benchmarks to compare with. A difference of 5.3 kJ/mol,

however, should be considered small when all the approximations of the computational setup are taken into account.

Table S1. Relative energies of reactants, TS and products in kJ mol^{-1}.

	Method[a]	Reactants	TS[b]	Products
m1[d]	*m*PWPW/b1/q1	0.0	7.2	–31.1
m2	*m*PW1PW/b1/q1	0.0	24.2	–31.2
m3	*m*PW1PW/b1/q1[c]	0.0	-	–32.4
m4	*m*PW1PW/b2/q1	0.0	25.0	–30.6
m5[d]	SCS-MP2/b1/q1	0.0	17.3	–36.4
m6	SCS-MP2/b2/q1	0.0	26.0	–36.6
m7[e]	*m*PW1PW/b2/q2	0.0	25.5	–29.7
m8[e]	SCS-MP2/b2/q2	0.0	24.9	–35.0

[a] The QM method, the basis set and the QM region used for single-point calculations at the geometries optimized with the *m*PWPW/b1/q1 scheme are indicated. [b] The energy of structure 10 of the reaction path. [c] Geometry of the reactants and products optimized with the *m*PW1PW/b1/q1 scheme. [d] These are results of limited value. See text. [e] These are the most reliable results. See main text.

Table S2. Relevant distances in the optimized reactant and product structures in Angstroms.

Method [a]	d (glucose-P)		d (P-Asp8)		d (Asp10-H [d])		d (H[d]-glucose)	
	R[b]	P[c]	R	P	R	P	R	P
mPWPW/b1/q1	2.73	1.76	1.86	2.90	1.64	1.05	1.02	1.52
mPW1PW/b1/q1	2.79	1.73	1.81	2.89	1.67	1.02	1.00	1.54

[a] The QM method, the basis set and the QM region used is shown. [b] Reactants. [c] Products. [d] The hydrogen atom that is transferred from Glucose to Asp10 along the reaction path.

NMR chemical Shifts

We have calculated the NMR chemical shifts with the IGLO method[2,3]. We used the B3LYP functional, and the IGLOIII basis set for all atoms except for Mg, for which it is not defined. Instead we substituted the TVZP basis set. The calculation was done with the q2 definition of the QM/MM region. The reference was $CFCl_3$, which was optimized with the SVP basis set and a COSMO solvation model[4,5] and the fluorine chemical displacement was calculated with B3LYP/IGLOIII.

Figure S1. Representation of the first 14 molecules surrounding a given β-PGM molecule (in green) in the crystal under study. This first shell of proteins was generated with Pymol. For these molecules the OPLS-AA MM charges of all atoms were included. A homogeneous background charge was added to make the system neutral. For additional images, 6 cells in each of the (i,j,k) directions were considered. To reduce the computational cost, the dipole moment of each chain was calculated and represented by two equivalent point charges. The unit cell contains 4 molecules and so the total number of dipoles added was $(6x2)^3x4-15=6897$. Because the energy changes introduced by this shell were negligible, we did not consider any more distant crystallographic images.

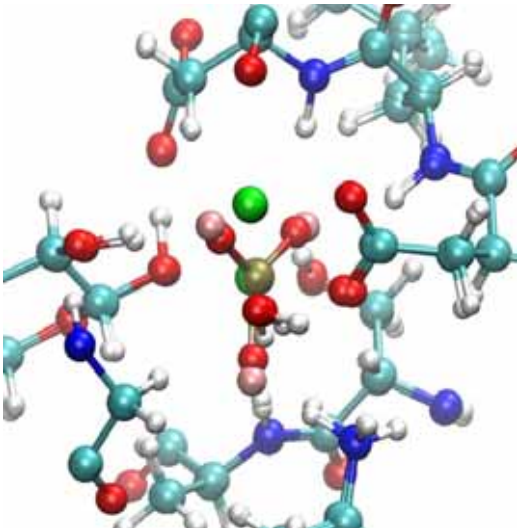


Figure S2. Representation of the calculated structures for the transition state of the phosphoryl transfer and the MgF_3^- complex. Striking similarities are observed in the position of active site residues, thus supporting the efficiency of the MgF_3^- ion as a transition state analogue.

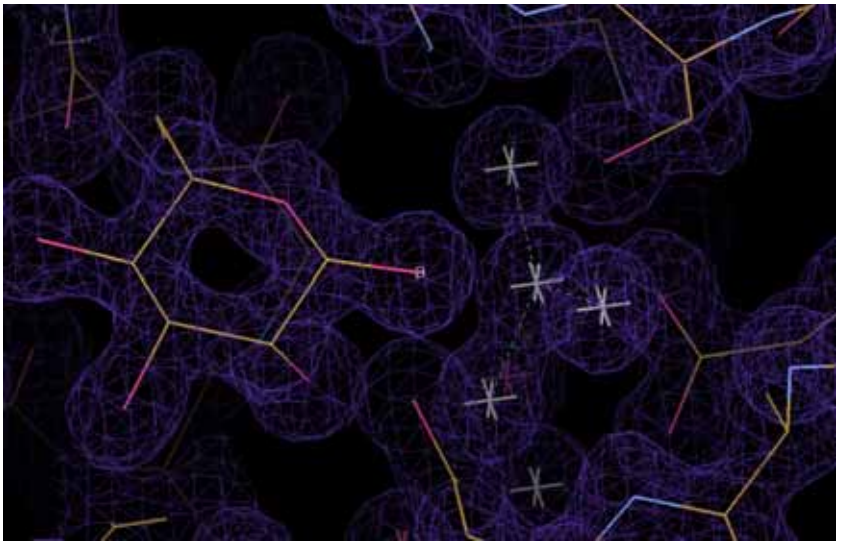


Figure S3. 2Fo-Fc electron density map fitting of the MgF_3^- region.

1. Marcos E, Crehuet R, Anglada JM. Inductive and external electric field effects in pentacoordinated phosphorus compounds. J Chem Theory Comput 2008; 4:49-63.

2. Kutzelnigg W. Theory of magnetic-susceptibilities and NMR chemical-shifts in terms of localized quantities. Isr J Chem 1980; 19:193-200.

3. Schindler M, Kutzelnigg W. Theory of magnetic-susceptibilities and NMR chemical-shifts in terms of localized quantities. 2. Application to some simple molecules. J Chem Phys 1982; 76:1919-1933.

4. Klamt A, Schüürmann G. COSMO - A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. J Chem Soc-Perkin Trans 2 1993: 799-805.

5. Barone V, Cossi M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. J Phys Chem A 1998; 102:1995-2001.

4.2. Dynamical properties of the Amino Acid Kinase family

The previous section focuses on phosphoryl transfer reactions and, in particular, on some factors determining the reaction mechanism with special emphasis on the putative pentacoordinated phosphorus intermediate in enzymes. This was necessary for a clear understanding of how the phosphoryl transfer reaction can proceed in enzymes. In this section, we are also focused on phosphoryl transfer enzymes but covering another aspect extremely important for enzymatic function: *large-amplitude dynamics*, also known as slow conformational dynamics. We have studied how large-amplitude motions necessary for the catalytic process emerge from the 3D-protein structure. For this purpose, we have investigated the large-amplitude motions of the Amino Acid Kinase (AAK) family of enzymes. In particular, this focuses on slow conformational motions linked to substrate binding and allosteric regulation. Interestingly, AAK members present different oligomeric states, so that this family represents, in addition, a suitable case for exploring the role of the oligomeric architecture in determining functional motions. This family of enzymes has been widely characterized with X-ray crystallography by our collaborators Rubio and co-workers at the Instituto de Biomedicina de Valencia (CSIC). To study large-amplitude motions we have applied the broadly used Elastic Network Models developed by Bahar and co-workers at University of Pittsburgh.

4.2.1. NAGK as a paradigm of large-amplitude motions in the Amino Acid Kinase family

The homodimeric enzyme N-Acetyl-Glutamate kinase from *E. Coli* (*Ec*NAGK) is regarded as the structural paradigm of the AAK family, so we decided to study the large-amplitude dynamics (low-frequency modes) of this enzyme and compare it with other family members: carbamate kinase (CK), UMP kinase (UMPK) and hexameric NAGK.

We first analyzed the low-frequency modes of motion of *Ec*NAGK and found that the intrinsic dynamics strongly correlates with the conformational transition between the bound and unbound forms of the enzyme as observed by crystallography. This demonstrates that the conformational change in NAGK necessary for ATP and NAG binding is encoded in the enzyme fold. As a second goal, we evaluated the degree to which the low-frequency modes of the other AAK enzymes considered in this study resemble those of NAGK. We found that AAK members exhibit well-defined dynamic patterns that are encoded in their shared arquitecture pointing to similar mechanisms of function as originally purposed by Rubio and co-workers. The developed approach is readily applicable to other families of proteins and indeed permits to identify dynamic fingerprints.

A detailed presentation of the results and methodologies used in this study can be found in the article: *On the conservation of the slow conformational dynamics within the Amino Acid Kinase family: NAGK the paradigm* (2010) PLoS Comput. Biol., 6:e1000738.

OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

On the Conservation of the Slow Conformational Dynamics within the Amino Acid Kinase Family: NAGK the Paradigm

Enrique Marcos[1], Ramon Crehuet[1]*, Ivet Bahar[2]*

1 Department of Biological Chemistry and Molecular Modelling, IQAC-CSIC, Barcelona, Spain, **2** Department of Computational Biology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

Abstract

N-Acetyl-L-Glutamate Kinase (NAGK) is the structural paradigm for examining the catalytic mechanisms and dynamics of amino acid kinase family members. Given that the slow conformational dynamics of the NAGK (at the microseconds time scale or slower) may be rate-limiting, it is of importance to assess the mechanisms of the most cooperative modes of motion intrinsically accessible to this enzyme. Here, we present the results from normal mode analysis using an elastic network model representation, which shows that the conformational mechanisms for substrate binding by NAGK strongly correlate with the intrinsic dynamics of the enzyme in the unbound form. We further analyzed the potential mechanisms of allosteric signalling within NAGK using a Markov model for network communication. Comparative analysis of the dynamics of family members strongly suggests that the low-frequency modes of motion and the associated intramolecular couplings that establish signal transduction are highly conserved among family members, in support of the paradigm sequence→structure→dynamics→function.

Citation: Marcos E, Crehuet R, Bahar I (2010) On the Conservation of the Slow Conformational Dynamics within the Amino Acid Kinase Family: NAGK the Paradigm. PLoS Comput Biol 6(4): e1000738. doi:10.1371/journal.pcbi.1000738

Editor: Michael Levitt, Stanford University, United States of America

Received November 20, 2009; **Accepted** March 5, 2010; **Published** April 8, 2010

Funding: This work was supported by grants from the JAE-predoc programme of the Consejo Superior de Investigaciones Científicas (CSIC), the Spanish MEC (CTQ2009-08223) and the Catalan AGAUR (2005SGR00111). Support from NIH 5R01LM007994-06 is gratefully acknowledged by IB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rcsqtc@iqac.csic.es (RC); bahar@ccbb.pitt.edu (IB)

Introduction

Many recent studies, both experimental and computational, point to the inherent ability of proteins to undergo, under native state conditions, large-amplitude conformational changes that are usually linked to their biological function. Proteins have access, via such equilibrium fluctuations, to an ensemble of conformers encoded by their 3-dimensional (3D) structure; and ligand binding essentially shifts the population of these pre-existing conformers in favour of the ligand-bound form [1–4]. With the accessibility of multiple structures resolved for a given protein in different forms, it is now possible to identify the principal changes in structure assumed by a given protein upon binding different ligands, which are observed to conform to those intrinsically accessible to the protein prior to ligand binding [5–7]. The observations suggest the dominance of proteins' intrinsic dynamics in defining the modes of interactions with the ligands. This is in contrast to the induced-fit model [8] where the ligand 'induces' the change in conformation. Instead, the Monod-Wyman-Changeux (MWC) [9] model of allostery where a selection from amongst those conformers already accessible is triggered upon ligand binding.

Yet, the choice between intrinsic *vs* induced dynamics, and the correlations between dynamics and function, are still to be established, and presumably depend on the particular systems of study [10]. NMR relaxation experiments provide evidence, for example, for the existence of correlations between the time scales of large-amplitude conformational motions and catalytic turnover [11,12]; and collective motions in the low frequency regime appear to be potentially limiting reaction rates. On the other hand, other studies point to the different time scales and events that control catalysis and binding events [13,14]. Furthermore, while the intrinsic dynamics in the unbound form is observed to be the dominant mechanism that facilitates protein-protein or protein-ligand complexation, the ligand may also promote structural rearrangements on a local scale at the binding site [2,15,16]. Given that proteins' collective dynamics, and thereby potential functional motions, are encoded by the structure, proteins grouped in families on the basis of their fold similarities would be expected to share relevant dynamical features [17–21]. It is of paramount importance, in this respect, to have a clear understanding of collective motions and their relationship to binding or catalytic activities, if any, toward gaining deeper insights into functional mechanisms shared by members of protein families.

Protein dynamics can be explored by means of all-atom force fields and simulations, or by coarse-grained (CG) models and methods. All-atom simulations such as Molecular Dynamics (MD) describe the conformational fluctuations of the system over a broad range of timescales. Except for small proteins, the main limitation of MD is that the timescales computationally attainable (below hundreds of nanoseconds) do not allow for accurate sampling of slow and large-amplitude motions (low-frequency modes) that are usually of biological interest. CG approaches, on the other hand, lack atomic details but provide insights into global movements. Among them, Elastic Network Models (ENMs) have

Author Summary

During the last 20 years both the experimental and computational communities have provided strong evidence that proteins cannot be regarded as static entities, but as intrinsically flexible molecules that exploit their fluctuation dynamics for catalytic and ligand-binding events, as well as for allosteric regulation. This intrinsic dynamics is encoded in the protein structure and, therefore, those proteins with similar folding should share dynamic features essential to their biological function. In this work, we have applied an Elastic Network Model to predict the large-amplitude dynamics of different enzymes belonging to the same protein family (Amino Acid Kinase family). Subsequent comparison of the dynamics of these proteins reveals that this protein family follows the same dynamic pattern. The present results are strongly supported by experimental data and provide new insights into the performance of biological function by these enzymes. The investigation presented here provides us with a useful framework to identify dynamic fingerprints among proteins with structural similarities.

found wide use in conjunction with normal mode analyses (NMAs) in the last decade [22]. ENMs describe the protein as a network, the nodes of which are usually identified by the spatial positions of C^{α}-atoms. Elastic springs of uniform force constant connect the nodes in the simplest (most broadly used) ENM, referred to as the anisotropic network model (ANM) [23–25]. Despite the oversimplified description of the protein conveyed by the ENMs, a surge of studies have shown that the predicted low-frequency modes describe well experimentally observed conformational changes and provide insights into potential mechanisms of function and allostery [5–7,24–27], in accord with NMAs performed [28,29] with more detailed models and force fields. Additionally, recent studies by Orozco and co-workers [30], and Liu et al [31] point to the similarities of the conformational space described by the low-frequency modes obtained from MD and that from CG NMA, provided that MD runs are long enough to accurately sample the collective motions.

The present study focuses on the amino acid kinase (AAK) family. This family comprises the following enzymes on the basis of sequence identity and structural similarities: N-acetyl-L-glutamate (NAG) kinase (NAGK), carbamate kinase (CK), glutamate-5-kinase (G5K), UMP kinase (UMPK), aspartokinase (AK) and the fosfomycin resistance kinase (FomA). Rubio and co-workers [32] have exhaustively studied this family and proposed that the shared fold among the members is likely to give rise to a similar mechanism of substrate binding and catalysis. NAGK is the most widely studied member of this family taking into account the large amount of structural information gathered [32,33]. This enzyme indeed serves as a structural paradigm for the AAK family, such that studying its structure-encoded dynamics can shed light on the mechanisms shared by family members to perform their function [32].

NAGK catalyzes the phosphorylation of NAG, which is the controlling step in arginine biosynthesis. The hallmark of this biosynthetic route in bacteria is that it proceeds through N-acetylated intermediates, as opposed to mammals that produce non-acetylated intermediates. Consequently, NAGK activity may be selectively inhibited and, taking into account that it is the controlling enzyme of arginine biosynthesis, it is a potential target for antibacterial drugs. In many organisms, NAGK phosphorylation is the controlling step in arginine biosynthesis. In these cases, NAGK is feedback inhibited by the end product arginine, and recent studies shed light on this mechanism of inhibition [34,35]. NAGK from *Escherichia Coli* (*Ec*NAGK), on the other hand, is arginine-insensitive. Its mechanism of phosphoryl transfer has been the most thoroughly characterized among the enzymes that catalyze the synthesis of acylphosphates (EC group 2.7.2). In particular, crystallographic studies by Rubio and coworkers [32,33] have provided insights into its mechanisms of binding and catalysis. *Ec*NAGK is a homodimer of 258 residues, each monomer being folded into an αβα sandwich (Figure 1). The N-domain of each subunit/monomer makes intersubunit contacts and hosts the NAG binding site (NAG lid), whereas the C-domain binds the ATP. The phosphoryl transfer reaction takes place at the interface between the two domains within each subunit. Kinetic studies show no evidence of cooperativity between subunits [36], suggesting that the dimeric structure provides thermodynamic

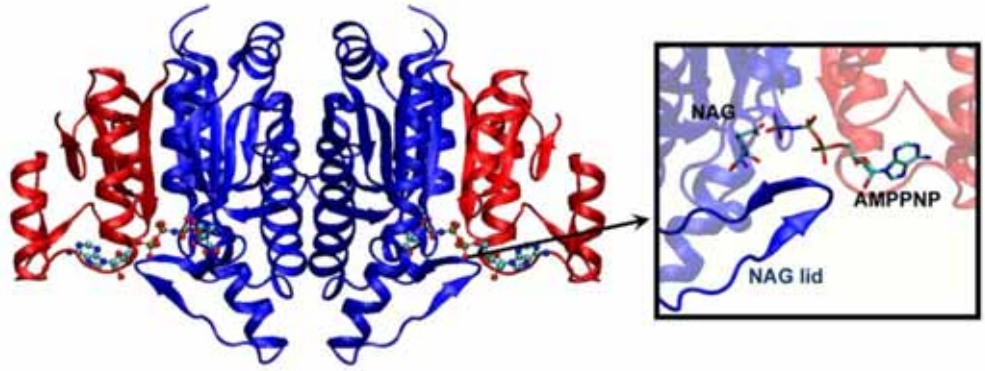


Figure 1. Structure of the closed form of NAGK. The NAGK dimer is complexed with the ATP analogue AMPPNP and the amino acid NAG (PDB code 1GS5) [32]. The substrates ATP and NAG (ball and sticks) bind to the C- (red) and N-domains (blue) respectively. N-terminal domains form the interface between the two monomers. Inset shows a closer view of the NAG lid and both ligands.
doi:10.1371/journal.pcbi.1000738.g001

stability, only, to the monomeric fold that has been evolutionary selected to perform the catalytic function.

The diverse crystallographic structures solved for the bound state of this enzyme indicate two types of functional motions [33]: (1) X-ray structures of *Ec*NAGK complexed with either ADP or with the inert ATP analogue AMPPNP (PDB codes 1GS5, 1OH9, 1OHA and 1OHB) have a too narrow active site to let the substrates bind directly; whereas the unbound structure (PDB code 2WXB; kindly provided by the authors prior to release) has a more open active site. This suggests that the enzyme undergoes a conformational closure that is likely to be triggered upon nucleotide binding, since all these complexes display a closed structure whether NAG is bound or not. (2) The ternary complex with ADP and NAG displays the ability to exchange NAG with a sulphate ion in solution without opening the active site. The NAG lid therefore must be able to open and close independently of other structural elements.

The aim of the present study is two-fold. Firstly, given the interest in acquiring deeper knowledge on the enzymatic mechanism of *Ec*NAGK and the potential role of slow dynamics in the pre-disposition of the enzymatic function, we analyze here the low-frequency modes of motion of *Ec*NAGK. Secondly, using *Ec*NAGK as the paradigm of AAK family, we assess to what extent the slow modes of motion are shared by other members of the AAK family.

Results/Discussion

Plan

The results are organized as follows. First, results from GNM analysis are presented, which give insights into the functional significance of residue fluctuations and underlying sizes of motions in the most readily accessible (i.e. softest) collective modes. Second, ANM modes are described to analyze the directionality/mechanism of these modes. Note that GNM does not provide information on 3N-dimensional structural changes, but on N-dimensional properties such as the mean-square fluctuations (MSFs) of residues, their cross-correlations, or movements along normal mode axes, hence the use of the ANM for exploring and visualizing the 3D motions (see *Methods*). Third, communication properties of the *Ec*NAGK enzyme are assessed based on graph theoretical examination of shortest paths between network nodes representative of the enzyme. Finally, a comparative analysis of the ANM dynamics of different members of the AAK family is made. GNM and ANM modes of *Ec*NAGK are computed for the open form in general, except for the analysis of the intrinsic dynamics of the closed form; and ligands are not included in the calculations. The predicted motions therefore reflect the intrinsic dynamics of the enzyme in the absence of bound ligands.

Mobility profiles for the EcNAGK open form

Figure 2 displays the results from the GNM analysis of the equilibrium dynamics of *Ec*NAGK. Panel (B) compares the MSFs of residues, $<(\Delta R_i)^2>$, predicted by the GNM with those indicated by X-ray crystallographic *B*-factors $B_i = 8\pi^2/3 \ <(\Delta R_i)^2>$. For clarity, the different structural elements are numbered and color-coded along the upper abscissa bar in accord with the colors in panel (A). Results for chains A and B are identical as a result of the dyadic axis of symmetry at the intersubunit surface. Calculations and experimental data refer to the open form of *Ec*NAGK. The high correlation coefficient ($r = 0.75$) between the experimental and theoretical curves in Figure 2 is remarkable in view of the simplicity of the GNM, but it is also worth noting that in the case of the closed conformation the correlation coefficient drops to $r = 0.61$. Indeed, ENMs tend to provide a better description of the dynamics of open forms [24].

The mobility profile in Figure 2B permits us to identify the most mobile and rigid regions of the protein from the maxima and minima, respectively. Mainly two dynamical features are distinguished. First, the β3–β4 hairpin, which corresponds to the NAG-binding site lid, is the most mobile part of the N-domain (region 3). Second, the β12–β13 hairpin and αF and αG helices (regions 9 and 10) emerge as the most mobile parts of the C-domain; notably, these structural elements are involved in ATP binding. It is remarkable that the topology of the structure provides flexibility near the two binding sites, which may be a functional requirement to accommodate ligand binding. Rigid/constrained elements, on the other hand, include the αC helices making intersubunit contacts, along with the strands β8 and β10 in the N-domain core. Moreover, the N-termini of αB and αE helices (regions 2 and 8), which point toward the γ-phosphate in the closed form, also show reduced mobility. The lack of mobility in these NAGK sequence motifs [32] is presumably a dynamic requirement to optimally perform their functional role in orienting their dipoles to withdraw negative charge from the transferring phosphate group. These results are consistent with those inferred by Rubio and co-workers from their crystallographic studies [32,33],

Global modes point to intrinsic mechanisms for opening/closing ligand binding sites

The decomposition of the global dynamics into a set of GNM modes permits us to identify the different kinds of motions allowed by the structure as well as the couplings between different parts of the protein. Moreover, minima in the low-frequency mode-profiles reveal mechanically important residues. When residues surrounding a given site move in opposite directions, the latter site serves as a hinge. Hinge sites at low-frequency modes, also called soft modes, usually serve as key mechanical site at the interface between domains subject to concerted movements [37].

Figure 3 shows the mobility of different parts of the protein in the first three softest modes. The diagrams are color coded from red (most rigid) to blue (most mobile), in accord with the size of motions undergone by the residues along these examined modes' axes (shown on the right panels). All three modes appear to induce motions symmetrically distributed about the inter-subunit interface. In the 1st mode, the mobility increases with distance away from the dyadic axis, such that the C-domain, and in particular the β12–β13 hairpin and αF helix, undergo the largest movements.

The 2nd slowest mode involves movements of the C- and N-domains with respect to each other within each monomer. A hinge site at residue A174 is observed, where previous crystallographic studies had exactly set the boundary between the C- and N-domains [32]. This hinge presumably enables the opening/closure of the active site in each subunit. A mutant on residue D162 [36], which is close to this hinge site, disrupted function and thus confirms this hinge as a key element in the functionality of the enzyme.

On the other hand, the 3rd mode involves mainly the β3–β4 hairpin, i.e., the NAG lid, and suggests an intrinsic ability at this region to move independently with respect to the ATP site (note that all modes are orthonormal and independent). Such local flexibility is consistent with an ability of the NAG lid to open and close the NAG binding site, in support of the hypothesis inferred from crystallographic studies [33]. The anticorrelated motion of the C-domain, due to a hinge at residue E181, is minimal but, as in mode 2 and together with the movement of NAG lid, might lead to the opening/closure of the active site. Interfacial residues

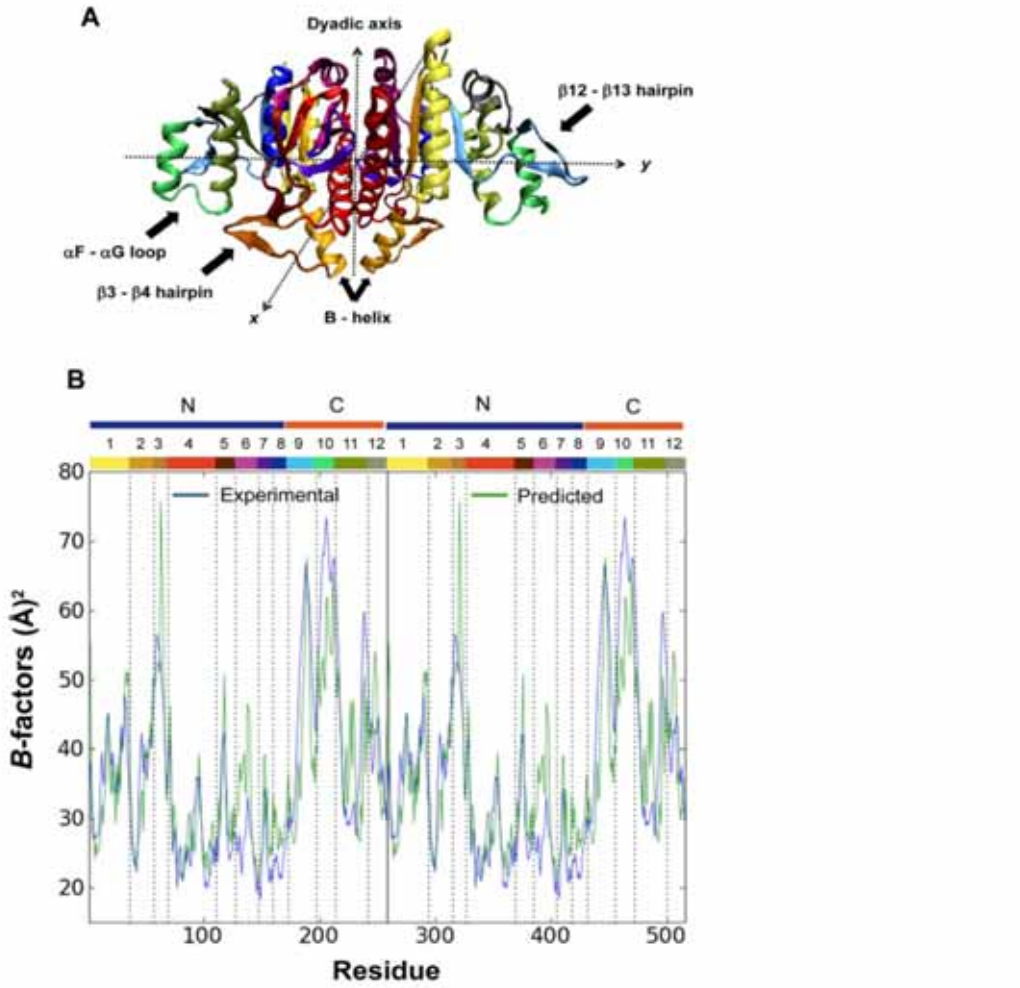


Figure 2. Structure and fluctuation dynamics of the open conformation of NAGK. (A) Color-coded ribbon diagram of NAGK where regions involved in substrate binding are indicated by arrows. All secondary structure elements are highlighted with different colors. (B) Comparison of experimentally observed (blue) and computationally predicted (green) *B*-factors. The theoretically predicted *B*-factors are rescaled based on the experimentally observed *B*-factors averaged over all residues. Experimental data refer to the PDB structure 2WXB (to be published). The range of residues of N and C domains is highlighted. The different parts of the protein have been numbered as follows: (1) β1+αA; (2) β2+αB; (3) β3–β4, the NAG lid; (4) αC+β5; (5) β6–β7; (6) αD+β8; (7) β9–β10;(8) αE; (9) β11+β12–β13+β14; (10) αF+αF–αG; (11) αG+β15; (12) αH+β16. The color code of the numbered parts of the protein is the same in both subunits, and indicated along the upper abscissa.
doi:10.1371/journal.pcbi.1000738.g002

making intersubunit contacts exhibit low mobilities, suggesting that the tightly packed hydrophobic contacts are essential to the thermodynamic stability of the dimeric protein.

How do global modes correlate with experimentally observed change in structure?

With the aim of gaining insights into the directionality of these modes, one can map GNM modes into ANM modes by comparing the mean-square fluctuations. A one-to-one correspondence would not be expected, due to differences in the number of modes as well as underlying potentials of the two ENMs. We found that the first GNM mode correlates with the 1st and 3rd ANM modes; the 2nd with the 1st, 2nd and 4th ANM modes; and the 3rd is the counterpart of the 5th ANM mode.

The directionality provided by the ANM approach helps us ascertain how well the slowest ANM modes describe the conformational difference observed between the open and closed structures resolved by X-ray crystallography. To this aim, the

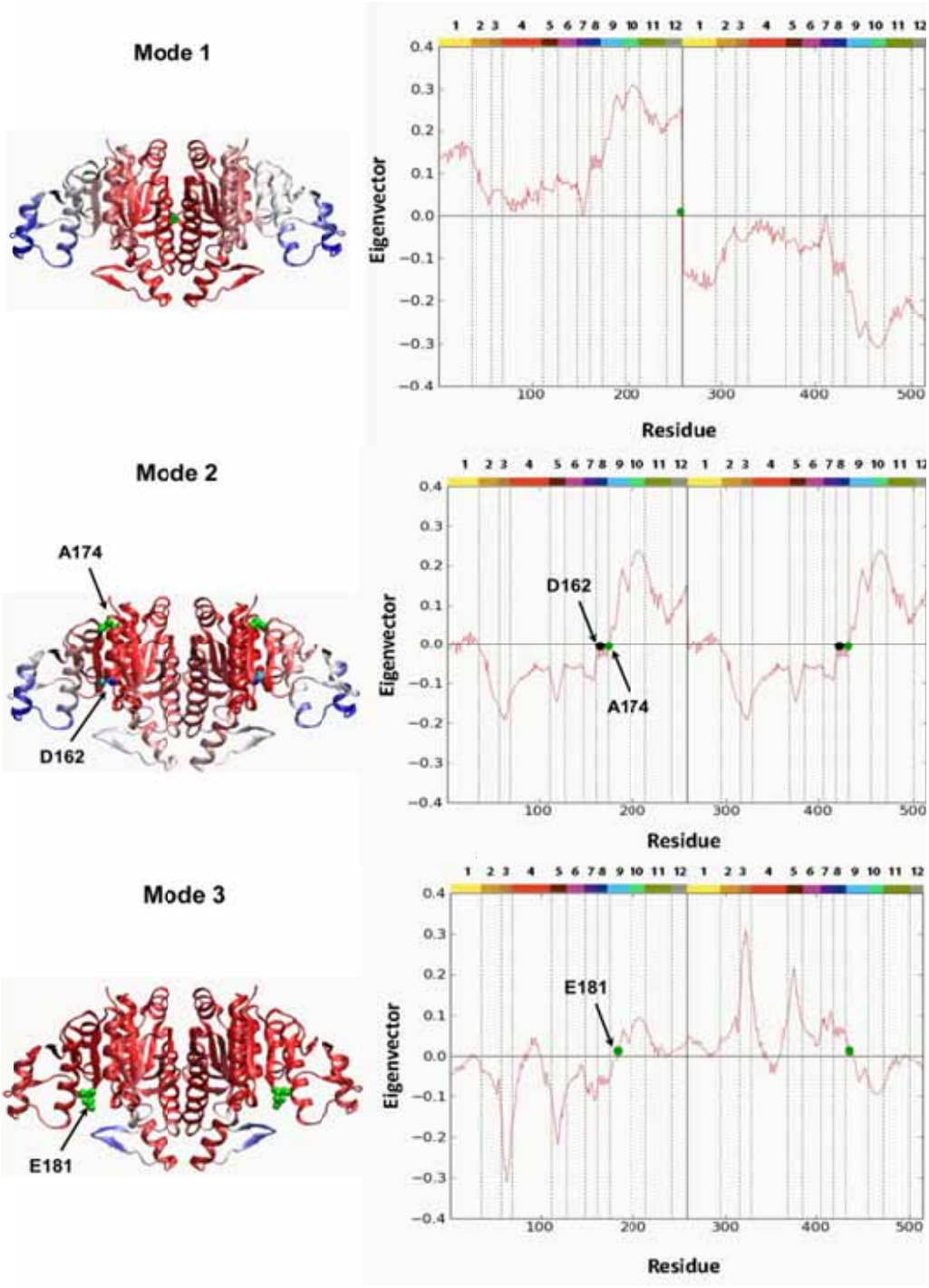

Mode 1
Eigenvector
Residue
Mode 2
A174
D162
Eigenvector
Residue
Mode 3
E181
Eigenvector
Residue

Figure 3. Mobilities of residues. Results are presented for the softest three GNM modes. On the left, the ribbon diagrams of NAGK, color-coded according to the mobilities of residues in the respective modes are displayed. The most mobile residues are colored blue, and the most rigid ones, red. The green dots on the diagrams indicate the position of the hinge sites. Note that in mode 2 the hinge sites closely neighbour the residue D162, which is a key catalytic residue.
doi:10.1371/journal.pcbi.1000738.g003

ANM modes are projected into the deformation vector $\Delta \boldsymbol{r}$ obtained from the open and closed conformations. Figure 4 displays the cumulative overlap (see equation (7)) between $\Delta \boldsymbol{r}$ and the ANM modes for both passages (from closed to open, and *vice versa*). It is worth emphasizing that the first 10 ANM modes of the open and closed forms are able to describe 84% and 76% of the observed conformational change, respectively. On the other hand, the open form requires a smaller set of modes to describe the deformation vector to a given extent. This is in agreement with the fact that the dynamics of open conformations are usually better described with ENM as also noted above. Modes 1, 3 and 5 are the main contributors to the cumulative overlap and thus the most relevant modes to describe the global dynamics of NAGK (see Figure S1).

Changes induced near active sites by global modes

As mentioned above three modes play a dominant role in enabling the functional changes in NAGK. Modes 1 and 3 drive a symmetrical opening and closing of the active site. In both modes, the most mobile region is the C-domain, which binds ATP, while the N-domain is practically rigid. Mode 5 ensures the opening/closing of the NAG-binding site by the β3–β4 hairpin that serves as a lid.

It is of the utmost importance to examine how active site residues move in these modes. Do they possess an intrinsic ability to adopt the conformation of the bound state, or does ligand binding trigger the conformational change? To explore this issue, we generated a series of conformations driven by these modes. Figure 5 illustrates the results for mode 1. This mode entails an anticorrelated movement of the two subunits as shown in panel (A). The following features are distinguished by a closer examination of functional sites. The catalytic residues K8, K217 and D162, and those in the vicinity, such as N158, exhibit minimal changes in their coordinates as seen in panel (B). On the other hand, a number of hydrophobic residues near the ATP binding site move concertedly in a direction required for coordinating ATP, as the subunit reconfigures from the open to the closed form (see panel (C)). This mode accessible in the absence of ATP binding thus facilitates the suitable re-positioning of these residues upon ATP binding. NAG-binding residues undergo minimal change during this global motion (panel (D)). The movement of residues in mode 3 complement those in the 1st mode to reach the bound conformation from the open form (Figure S2).

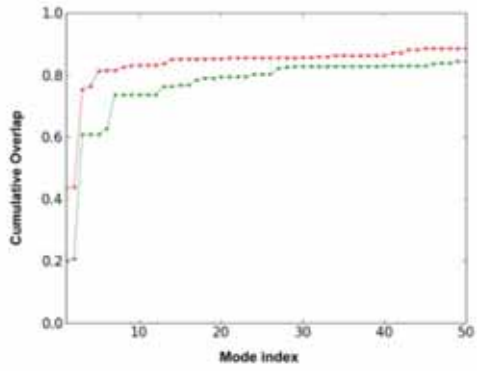


Figure 4. Comparison with experimental conformational changes. Cumulative overlaps CO(*m*) between ANM modes and the experimentally observed deformation between the open and closed forms of NAGK are plotted for subsets of m modes, in the range $1 \leq m \leq 50$ (see equation (7) in *Methods*). We note that the first 3 ANM modes (among $3N\text{-}6 = 1542$ modes) accessible to the open form (red) yield an overlap of 0.75 with the experimentally observed reconfiguration from open to closed state of the enzyme. In the case of the closed form, the first 3 modes yield an overlap of 0.61. In either case, a small subset of modes intrinsically accessible to the structure attain a cumulative overlap of >0.80, pointing to the pre-disposition of the structure to undergo its functional changes in conformation between the open and closed forms.
doi:10.1371/journal.pcbi.1000738.g004

K8 and D162 are key catalytic residues on the basis of structural [32,33] and mutational [36] studies. D162 is inferred from these studies to play a critical role in properly positioning two lysines (K8, K217) that stabilize the negative charge of ATP. The minimal displacements of D162 and K8 in these global modes, and the intrinsic tendency of K217 to move toward D162, are presumably dynamic requirements to optimally perform their catalytic roles (note that D162 is located close to the hinge site of GNM mode 2, as pointed out above, and thus its mobility is rather constrained). This rigidity is confirmed by the striking similarity in the orientation of these residues in different bound states of the enzyme [33] that characterize the entire catalytic process. The rearrangements of catalytic residues may be necessary to optimally orient, or pre-organize, the ligands to catalyze the chemical reaction [13,14,38]. In this case, some additional changes appear to occur in the bound form, such as the change in the side chain conformation of K217, which exchanges a salt bridge between residues E181 and D162. These rearrangements would presumably take place upon ligand binding, since E181 interacts with ATP via hydrogen bonds.

In relation to the NAG binding process, mutants on the NAG binding site revealed that N158 and R66 are key residues that underlie the affinity of *Ec*NAGK for NAG [36]. By examining the NAG binding mode (5th ANM mode, see Figure S3), R66 was found to be far more flexible than N158. This suggests that R66 may play a role in the recognition of the ligand, whereas the less exposed residue N158 might subsequently aid to fix the position of NAG at the active site. Furthermore, upon NAG binding, the size of the hydrophobic pocket (L65, R66, V122 and N160) that hosts the methyl group of NAG is reduced upon correlated movements between R66 and L65 toward the closed form. Binding of R66 and N158 to NAG, thus, apparently guides L65 toward the methyl group of the substrate. The correlated movements of L65 together with the rigidity of V122 and N160 fix the size of the hydrophobic pocket, which has been observed to be unable to bind glutamate derivatives with larger N-acyl groups [32,39] (Figure S3).

Communication properties

Using the Markov model described in the *Methods*, we computed the *hitting times* H_{ij}. H_{ij} provides a measure of the average path length over all possible combinations of edges, required to send

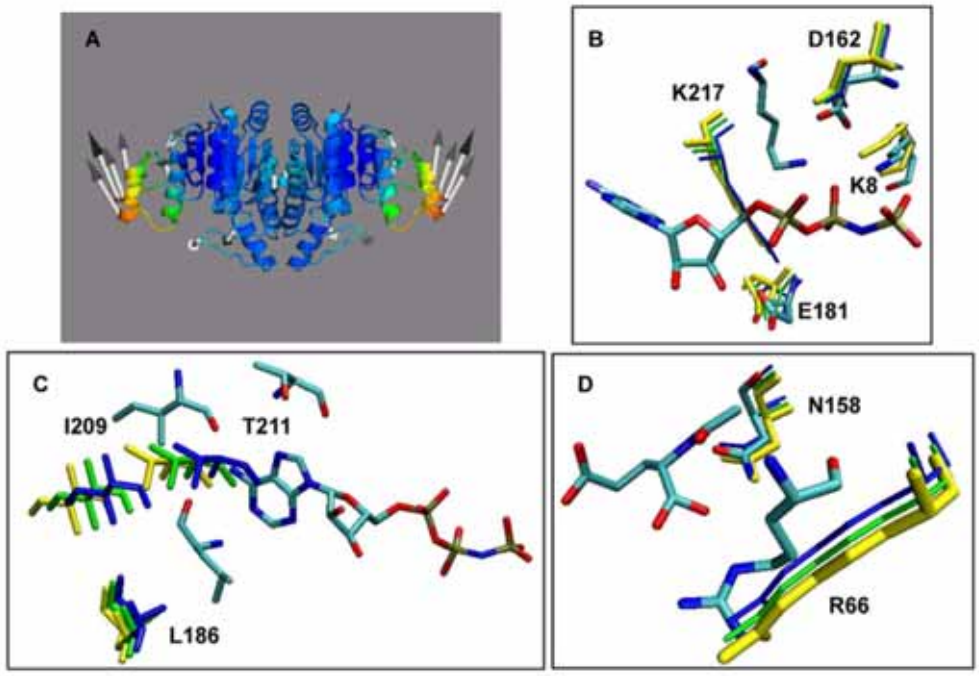


Figure 5. Movement along the slowest ANM mode. Motion of active site residues between open and closed conformers along the 1st ANM mode accessible to the open form. The position of these residues in different conformations is shown: open conformation (yellow), intermediate positions (green and blue) and closed conformation (atom-colored). (A) Color-coded ribbon diagram for motions along the 1st mode (generated with the ANM web server [25] and Pymol [64]). (B) Movement of catalytic residues with respect to the ATP analogue. (C) Movement of ATP binding residues with respect to the nucleotide. (D) Movement of NAG binding residues with respect to NAG.
doi:10.1371/journal.pcbi.1000738.g005

information to a given node *j*, or 'hit' residue *j*, starting from node *i*. The hitting times for all pairs of residues were evaluated for three different cases: open form (NAGK(O)), closed from without ligands (NAGK(C)) and closed form with ligands (NAGK(C)+ligands). Toward gaining an understanding of the communication propensity of individual residues, results have been consolidated, by calculating the mean hitting time, $<H_i> = (1/N)\ \Sigma_j\ H_{ji}$, for each residue *i*.

Figure 6A displays the mean hitting times of all residues in the three cases. The main contribution to H_{ji} arises from the MSF of the target residue itself (via the term $[\mathbf{\Gamma}^{-1}]_{jj}$ in equation (8), which in turn is proportional to $<(\Delta R_j)^2>$ - see equation (3)). As a result, the average hitting time profile shares some characteristics with the MSF profile shown in Figure 2B. The minima correspond to the most efficient receivers; these exhibit minimal fluctuations in their positions. It is worth noting that catalytic residues are among those with the lowest hitting times. These results suggests that the structural position and contact topology of the active site have been evolutionary designed to effectively receive signals from the binding sites and other parts of the protein so as to optimize the catalytic activity of the enzyme. On the other hand, the ligand-binding residues exhibit a broader variety of hitting times.

A closer comparison of the results obtained for the three structures revealed an interesting feature upon examination of the average hitting times between different substructures. The results in Figure 6C display the average path lengths evaluated for the communication between such particular domains in each structure: the average path lengths over all residue pairs belonging to the respective C-terminal and N-terminal domains (blue curves), those over all the N-terminal domain and catalytic site residues (red), and those over the C-terminal domain and catalytic site residues (green). These results clearly demonstrate that the closure of the structure enhances the communication of residues (decreases the average hitting times or path lengths) and upon ligand binding the communication shows a further improvement. Panel (D) demonstrates that not only the average path lengths, but the variance in the path lengths decrease upon domain closure, and ligand binding. In all cases, the N- and C-domains exhibit average path lengths longer than those connecting either domain to the catalytic site. This is a natural consequence of the location of catalytic residues - at the inter-domain region, where the phosphoryl transfer takes place.

Figure 7 panels (A) and (B) illustrates the three types of communication pathways in the open and closed states. Three residues have been selected as endpoints representative of the N-terminal domain NAG-binding site (R66), C-terminal domain nucleotide-binding site (L209) and the catalytic site at the inter-domain interface (D162), and the residues along the shortest paths

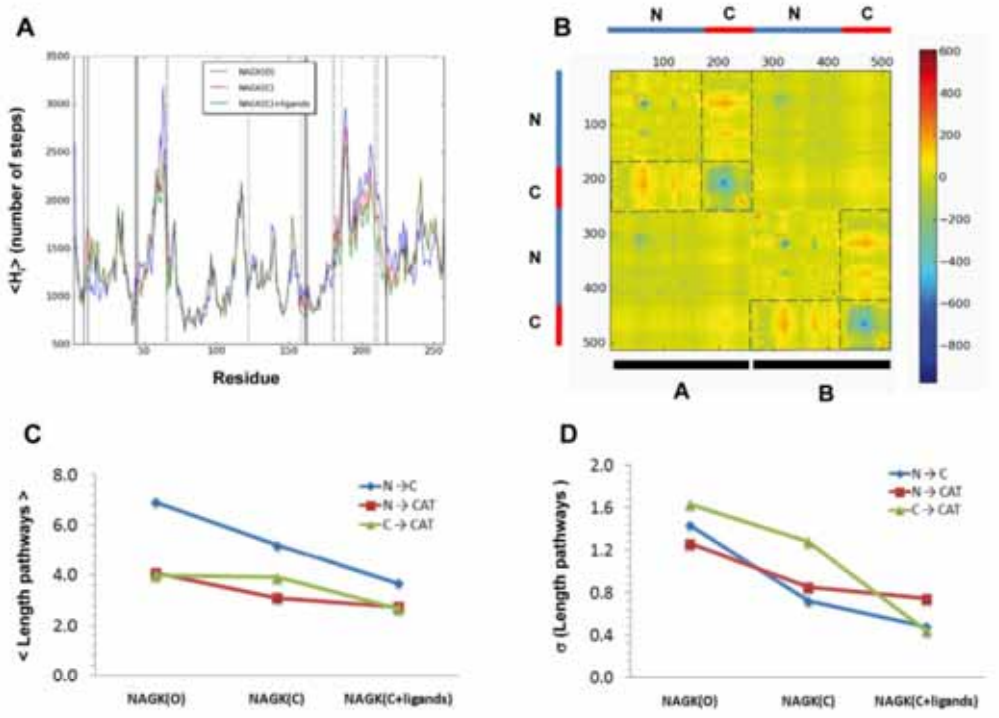


Figure 6. Communication properties of NAGK. (A) Mean hitting time profile for the open and closed (with and without ligands) forms of NAGK. Vertical lines indicate the positions of catalytic (solid line) and ligand-binding residues (dotted line). Note that catalytic residue tend to occupy minima positions, indicative of their efficient communication properties. (B) Difference map between the contribution to hitting times from cross-correlations ($\mathbf{\Gamma}_{ij}^{-1}$) (equation (8) in *Methods*) of the open and ligand-bound closed forms. Dashed lines set the boundaries of N- and C- domains and also enclose those pairs of domains that undergo the largest changes in the contribution from cross-correlations upon ligand binding. (C) Mean path lengths for linking different parts of the protein: N- and C-domains (blue), N-domain and catalytic site (red), and C-domain and catalytic site (green). (D) Standard deviation in the mean paths displayed in panel (C).
doi:10.1371/journal.pcbi.1000738.g006

evaluated using the Dijkstra's algorithm (see *Methods*) are shown by different dots in each case (see caption). We note in panel (C) that the ligand in the closed+liganded structure effectively spans the optimal communication pathway.

The enhancement of communication observed in panels (C) and (D) of Figure 6 is a consequence of the rigidity imparted by the closure of the structure and by ligand binding. The structure obviously becomes more cohesive in the closed conformation and consequently the couplings between residue fluctuations are increased, or the fluctuations in inter-residue distances are reduced. As summarized in the methods and derived in detail in our previous work [40], the commute times τ_{ij} between residue pairs directly scale with the fluctuations in the corresponding inter-residue distances (see equation (9)). Restrictions in inter-residue distance fluctuations acquired upon closure of the structure thus necessarily induce an enhancement in communication. From a catalytic point of view, the closure of the structure upon substrate binding is presumably an efficient way to optimize signal transduction and facilitate the catalytic process. Panel (B) in Figure 6 displays the changes in the contribution to hitting times from cross-correlations ($\mathbf{\Gamma}_{ij}^{-1}$) evaluated using equation (8) (see *Methods*) twice, for the open and closed+liganded states, and taking their difference. It is observed that upon domain closure and ligand binding the contribution from cross-correlations within the C-domain decrease (blue points in the panel), whereas those between the N- and C-domains within each subunit increase (red points in the panel). The resultant shorter and more homogeneous communication pathways suggest that ligands tend to centralize the communication between the C- and N- domains. Therefore, the transmission of conformational signals between the flexible domains and the more rigid catalytic residues takes place across the substrates. This might indicate a way to cooperatively optimize substrate binding (or product release) or even couple the intrinsic enzyme dynamics to the catalysis of the chemical reaction.

Comparison of EcNAGK dynamics with other members of the AAK family

It is of interest to determine if the dynamical features observed for *Ec*NAGK are shared by other members of the AAK family. Various approaches can be adopted to this aim [19,21,41]. Here,

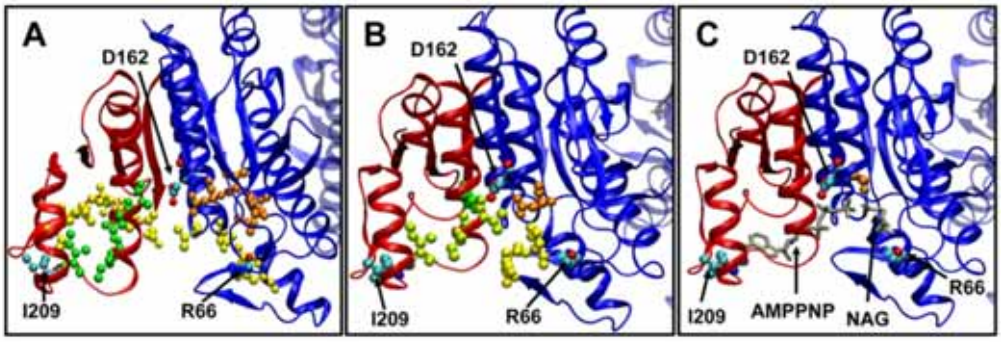


Figure 7. Communication pathways between catalytic site and ligand-binding residues in *Ec*NAGK. The communication pathways are represented by the network of residues (each atom is shown as a dot) along the shortest paths of communication between the following residue pairs: R66-L209 (yellow), R66-D162 (orange) and L209-D162 (green). These three cases are representative of the communication between a NAG-binding residue (R66) on the N-domain, a catalytic site residue (D162) and an AMPPNP binding residue on the C-domain (L209). These residues at the endpoints of the pathways are colored by atom name. N- and C- domains are colored blue and red, respectively. These pathways are shown for the three states considered: (A) Open state. (B) Closed state. The pathways R66-L209 and L209-D162 have three residues in common (colored in light green). (C) Closed state with ligands AMPPNP and NAG. The ligands are shown with gray sticks for a better visualization and participate in all three pathways considered in the figure. Note that the ligands directly establish the communication between the pairs of residues considered.
doi:10.1371/journal.pcbi.1000738.g007

we focus on global dynamics and compare the ANM modes predicted for *Ec*NAGK with those predicted for each AAK member. First, each pair of enzymes is structurally aligned using the DALI server [42]. Second, we define our 'subsystem' as the aligned portions of the families members and constructed the Hessian submatrices for these portions, and the remaining chain segments are considered as environment, similar to the approach adopted by Zheng & Brooks [43], described in *Methods*). Third, NMA is performed using equation (11) for the Hessian of the examined system. The correlation cosines between the collective modes evaluated for each subsystem and those calculated for the *Ec*NAGK dimer are presented in Figure 8. Results are shown for the top-ranking 10 modes, calculated for the corresponding dimers of three different members of the AAK family (structures shown in Figure 8): Carbamate kinase from *Pyrococcus furiosus* (*Pf*CK), NAGK from *Thermotoga Maritima* (*Tm*NAGK) and UMPK from *Pyrococcus furiosus* (*Pf*UMPK). Further information on the structural and dynamical pairwise comparisons is provided in Table 1.

***EcNAGK vs. PfCK* (Figure 8A).** The crystallographic structure of *PfCK* (PDB code 1E19) represents the open form of the enzyme [44]. Remarkably high correlations are obtained between the slow modes accessible to the two enzymes as may be seen in Figure 8A.

***EcNAGK vs. TmNAGK* (Figure 8B).** *Tm*NAGK is a hexamer that can be regarded as a trimer of *Ec*NAGK-like dimers (PDB code 2BTY). Thus, we compared the slow modes of *Ec*NAGK with those sampled by the *Ec*NAGK-like dimer from *Tm*NAGK. The five lowest modes of *Ec*NAGK are almost identically observed in *Tm*NAGK, except for a change in the order (or relative frequencies) of the modes.

***EcNAGK vs. PfUMPK* (Figure 8C).** UMPK is also a trimer of dimers [45] (PDB code 2BRI), but the monomeric subunits within these dimers are not arranged in the same manner as *Ec*NAGK or *Tm*NAGK. Thus, the comparison has been made between the dynamics of the monomeric subunits of *Ec*NAGK and *Pf*UMPK, including as environment, *via* equation (11), the remaining residues of the respective dimers. The mode-mode correspondences are not as clear as in the previous cases, but there is still some discernible correlation. The weaker correspondence can be attributed to the fact that the percentage of aligned residues (77%) is lower than that of the previous two cases (above 90%) (see Table 1).

It is worth pointing out that the modes of motion of the dimeric scaffold of the hexameric enzymes (*Tm*NAGK and *Pf*UMPK) may well be affected by the interfaces with the rest of subunits. The analysis of oligomerization effects on the dynamics of these proteins, however, is beyond the scope of this article and will be published elsewhere. The three cases studied here illustrate how the slow conformational dynamics of the *Ec*NAGK dimer is preserved to a large extent among the members of the AAK family. It is worth emphasizing that some of the modes are remarkably well conserved. In accordance with other studies [19], the lowest frequency modes prove here to be robust to sequence and structural variations within a given protein family; and the shared dynamics may be viewed as a dynamic fingerprint of the AAK family.

Conclusions

The present study focused on the *Ec*NAGK dimer in order to provide new insights into the competition between intrinsic *vs* induced dynamics in controlling enzymatic activity, assessing which residues play a key role in mediating the collective motions, or which conformational mechanisms are shared among members of the AAK family. The most probable modes of motion encoded by the structure have been determined using the available structural data for *Ec*NAGK dimer, which is used as a prototype, as well as other members of the family. The present study illustrates that this family, not only has important sequence and structure similarities, but also shares relevant dynamical features (Figure 8).

The results in Figure 4 demonstrate that the conformational change observed between the open and closed forms of *Ec*NAGK are essentially accomplished by movements along a small subset of modes (among the complete set of 1542 modes accessible to the enzyme); these are the modes predicted by the GNM to be the softest, i.e., they incur the least ascent in energy for a given size of

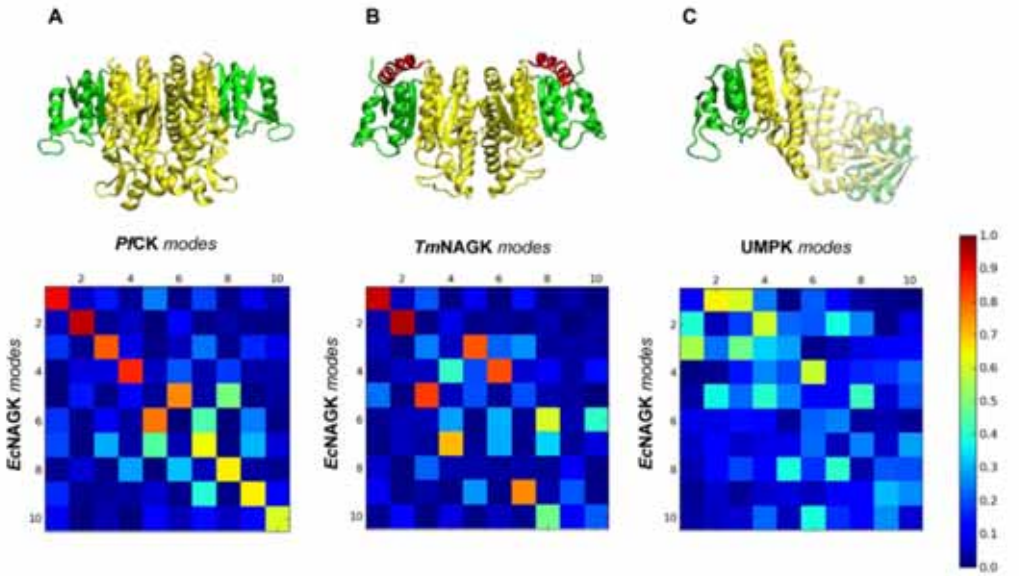


Figure 8. Conservation of the lowest frequency modes of motion among the AAK family members. Correlation cosines between the ten slowest modes of *Ec*NAGK and those of three other members of the AAK family. (A) Dimeric *Pf*CK. (B) *Ec*NAGK-like dimer of *Tm*NAGK. (C) Monomer of *Pf*UMPK. Ribbon diagrams represent the structure of the corresponding dimers of the AAK family that are compared with *Ec*NAGK. The C- and N-domains are colored green and yellow. In panel B, the N-terminal helix (red) is a key element for the hexamerization of *Tm*NAGK. In panel C, one of the monomers is shadowed to highlight that it has been considered as the coupling environment of the monomer that is being compared with a monomer of *Ec*NAGK.
doi:10.1371/journal.pcbi.1000738.g008

motion. This shows that the 'easiest' movements intrinsically favoured by the enzyme structure are actually those underlying its ligand binding mechanism, suggesting an evolutionary optimization of structure to favour functional dynamics.

Closer examination of the changes on a local scale, on the other hand, shows that the changes induced at residue level, may have the right directions to facilitate the interactions with the substrate, but are not sufficiently large and specific enough to explain the repositioning of amino acids near active sites. The ATP-bound conformation appears, for example, to result from the intrinsic dynamics of the protein combined with local rearrangements induced by the ligand to optimize its interactions in the complex (Figure 5).

Table 1. Structural and dynamical similarities among the AAK family members.

Pairwise comparison	Slow modes overlap [a] (%)	RMSD (Å)	Structurally aligned positions [b] (%)	Sequence identity at aligned positions (%)
*Ec*NAGK/*Pf*CK	88	3.9	91	20
*Ec*NAGK/*Tm*NAGK	81	3.3	99	32
*Ec*NAGK/*Pf*UMPK	70	3.3	77	16

(a) $\langle \mathrm{COI}(m) \rangle$, for m = 10, averaged over the first 10 modes.
(b) Residue pairs of proteins A and B are aligned by minimizing the deviations of intramolecular α-carbon distances (r_{ij}^A and r_{ij}^B) relative to their arithmetic mean. A threshold of similarity is set to 20%.
doi:10.1371/journal.pcbi.1000738.t001

The low frequency modes shared among the examined family members are shown here to enable access to the active site, by opening/closing to the environment the cleft where catalytic residues reside. Notably, these movements have minimal effects on the organization of catalytic residues, which are located near the hinge center that allows for the relative movements of the N- and C-domains in each subunit. The restricted mobility of catalytic residues is consistent with previous observations where catalytic sites have been pointed out to be highly constrained in the global modes and occupy key positions (near or coinciding with global hinges) in the structure. The collective modes do not therefore induce distinctive rearrangements at the catalytic residues. However, they appear to modulate their communication to the environment, i.e. they provide exposure to solvent, and/or a loosening or tightening of the packing density in the neighbourhood which apparently plays a role in controlling the propagation of structural or energetic perturbations to/from the active site.

Perhaps the most striking results concern the communication properties and the role of domain movements and ligand binding in enhancing allosteric effects. The decrease in the mean values and dispersion of the hitting times and the communication path lengths upon ligand-binding (Figure 6) suggests that the ligands optimize the coupling of domain movements that are already characteristic of the intrinsic protein dynamics. Indeed, the structure may have been evolutionary selected to bind the substrates in an optimal position to maximize the allosteric couplings.

Methods

The protein structure is represented as an ENM. The coordinates in the native structure are assumed to define the equilibrium positions of network nodes/residues. Pairs of nodes within a cutoff distance are coupled by elastic springs. Although the inter-residue interactions are non-specific, the collective dynamics of the protein in the low frequency regime is primarily and robustly determined by the overall fold [17,46,47], which permits us to explore the cooperative motions of NAGK using ENMs. Different ENMs have been developed [23,48–51], and Phillips and co-workers [52] demonstrated that these provide a consistent description of the lowest-frequency modes. Here we adopt the GNM [51,53] and the ANM [23].

Gaussian Network Model (GNM)

The GNM potential depends on the vectorial distance between each pair of nodes as

$$V_{GNM} = \frac{\gamma}{2}\sum_{i,j}^{N} \Gamma_{ij}\left[\left(\boldsymbol{R}_{ij}-\boldsymbol{R}_{ij}^{0}\right)\cdot\left(\boldsymbol{R}_{ij}-\boldsymbol{R}_{ij}^{0}\right)\right] \quad (1)$$

where N is the total number of residues, γ the uniform force constant for all springs in the network, $\boldsymbol{\Gamma}_{ij}$ is the ij^{th} element of the $N\times N$ Kirchhoff matrix $\boldsymbol{\Gamma}$ that defines the connectivity of the network, equal to -1 if residues i and j are within a cutoff distance R_c, zero otherwise, $\mathbf{R}_{ij}$ and $\mathbf{R}_{ij}^{0}$ are the instantaneous and equilibrium distance vectors between residues i and j, residue positions being identified by those of their α-carbons in the PDB files. The GNM approach allows for decomposing the dynamics of the protein into a set of normal modes of motion upon eigenvalue decomposition of $\boldsymbol{\Gamma}$. The contribution of the k^{th} mode to the MSF of residue i is expressed as

$$\left[(\boldsymbol{\Delta R}_i)^2\right]_k = \frac{3k_BT}{\gamma}\left[\lambda_k^{-1}(\boldsymbol{u}_k)_i^2\right] \quad (2)$$

where λ_k and $\boldsymbol{u_k}$ are the k^{th} eigenvalue and eigenvector of $\boldsymbol{\Gamma}$, respectively; $(\boldsymbol{u}_k)_i$ designates the mobility of residue i along the k^{th} mode. The low-frequency modes usually have the highest degree of collectivity, and they make the largest contribution to the observed MSFs

$$\langle(\boldsymbol{\Delta R}_i)^2\rangle = \sum_{k=1}^{N-1}\left[(\boldsymbol{\Delta R}_i)^2\right]_k = \frac{3k_BT}{\gamma}\left[\Gamma_{ii}^{-1}\right] \quad (3)$$

where the summation is performed over all non-zero modes and $[\boldsymbol{\Gamma}^{-1}]_{ii}$ designates the i^{th} diagonal element of the inverse of $\boldsymbol{\Gamma}$. Therefore, the lowest frequency modes usually provide insights into the cooperative motions involved in biological function [22].

GNM provides information on the relative sizes of residue motions in different modes (equation (2)), the MSFs of individual residues (equation (3)), or their cross-correlations

$$\langle\boldsymbol{\Delta R}_i\cdot\boldsymbol{\Delta R}_j\rangle = \frac{3k_BT}{\gamma}\left[\Gamma_{ij}^{-1}\right] \quad (4)$$

(obtained by rewriting equation (3) for pairs of residues i and j), as but not on their directionality; the fluctuations are implicitly assumed to be isotropic. The 3D characterization of the normal modes is provided by the ANM.

Anisotropic Network Model (ANM)

The ANM potential is a function of the scalar distance between the interacting pair of nodes and is given by [23]

$$V_{ANM} = \frac{\gamma}{2}\sum_{i,j}^{N}\left(|\boldsymbol{R}_{ij}|^2-|\boldsymbol{R}_{ij}^{0}|^2\right) \quad (5)$$

Both GNM and ANM penalize inter-residue distance changes, but GNM also takes into account the orientational change of the inter-residue vector, which leads to better agreement with experimental B-factors [54]. NMA is carried out using the $3N\times3N$ Hessian matrix $\mathbf{H}$ derived from the ANM potential. The diagonalization of $\mathbf{H}$ yields $3N$-6 non-zero modes (as opposed to N-1 in the GNM). A given ANM mode (eigenvector) thus contains information on the x-,y- and z-components of the motion undergone by each residue, thus describing the spatial directionalities of the collective motions.

Generation of large-amplitude conformational changes

ANM modes can be used to generate alternative conformations sampled along most easily accessible (lowest frequency) mode directions. Due to the harmonic character of the potential, two sets of conformers are obtained for a given mode k:

$$[\boldsymbol{R}(\pm s)]_k = \boldsymbol{R^0} \pm s_k\lambda_k^{-1/2}\boldsymbol{u}_k^{ANM} \quad (6)$$

where λ_k and $\boldsymbol{u}_k^{ANM}$ are the eigenvalue and eigenvector for mode k respectively, $\boldsymbol{R^0}$ is the $3N$-dimensional vector representing the initial coordinates and s_k is a parameter that scales the amplitude of the deformation induced by mode k. No sidechain atomic coordinates are included in the ANM calculations. An all-atom model for the deformed structure is generated by displacing the backbone and side chain atoms of each residue along the mode component of the corresponding C^{α}-atom and subsequent energy minimization. Such energy minimization performed with Gromacs [55] was verified to involve negligible conformational change in the backbone.

Comparison of experimental conformational changes with ANM modes

The degree of similarity between a conformational change $\boldsymbol{\Delta r}$ observed by crystallography and the theoretically predicted direction of the k^{th} mode can be quantified with the correlation cosine, $cos(\boldsymbol{\Delta r}\cdot\boldsymbol{u}_k^{ANM})$. Here $\boldsymbol{\Delta r}$ refers to the 3N-dimensional difference vector between the α-carbon coordinates of the open form and closed form of NAGK, for example, after optimal superimposition of the two structures to eliminate the external degrees of freedom. The cumulative overlap between the experimentally observed deformation $\boldsymbol{\Delta r}$ and that accounted for by a subset of m modes ($m < 3N$-6) is given by a summation of squared correlation cosines as

$$[CO(m)]^2 = \sum_{k=1}^{m}\cos^2\left(\boldsymbol{\Delta r}\cdot\boldsymbol{u}_k^{ANM}\right) \quad (7)$$

The summation of the squared cosines over all $3N$-6 nonzero modes is identically equal to unity as the eigenvectors form a complete orthonormal basis set in the $3N$-6 dimensional space of internal conformational changes. In the absence of correlation between $\boldsymbol{\Delta r}$ and $\boldsymbol{u}_k^{ANM}$, the average correlation cosine squared contributed by mode k will thus be $1/(3N$-6). Note the strong departure from this random behaviour in Figure 4.

Communication propensities. Hitting times and relation to collective dynamics

Inter-residue communication has been suggested as an essential mechanism in the allosteric regulation of protein function and enzymatic catalysis [56], and explored in diverse computational studies [57–59]. A network-based Markov model has been recently developed [40,60], which reconciles the NMA-based predictions on allosteric changes in conformations (global modes) with the shortest path(s) analyses based on graph theoretical methods [28]. We use this method to identify communication paths. The interactions between residue pairs are defined therein by the affinity matrix **A**. The elements of this matrix are defined as [60] $a_{ij} = N_{ij}/(N_i\,N_j)^{1/2}$ where N_{ij} is the number of atom-atom contacts between residues i and j, based on a threshold distance of 4 Å, and N_i,N_j are the number of heavy atoms of both residues. **A** is related to the Kirchhoff matrix $\mathbf{\Gamma_M}$ (same as GNM $\mathbf{\Gamma}$, except for the adoption of affinities, instead of γ, for the weights of the edges) as $\mathbf{\Gamma_M} = \mathbf{D} - \mathbf{A}$, where **D** is the diagonal matrix of elements $d_i = \sum_{j=1}^{N} a_{ij}$. In the simplified case where $a_{ij} = 1$ for all $R_{ij} < R_c$, $\mathbf{\Gamma_M}$ reduces to the GNM $\mathbf{\Gamma}$. The network communication is controlled by the Markov transition matrix $\mathbf{M} = \{m_{ij}\}$, where $m_{ij} = a_{ij}/d_j$ represents the conditional probability of transmitting a signal from residue j to residue i in one time step [60]. We define $-\log(m_{ij})$ as the 'distance' between two residues, in terms of communication, and the maximum-likelihood paths associated with each residue pair are evaluated using the Dijkstra's algorithm [60]. This permits us to evaluate a basic communication property: hitting time $\mathbf{H}_{ji}$ as the average path length for the passage of signals from node i to node j [40]. $\mathbf{H}_{ji}$ can be expressed in terms of the elements of $\mathbf{\Gamma}^{-1}$ (or $\mathbf{\Gamma}_{\mathbf{M}}^{-1}$) as [40]

$$\mathbf{H}_{ji} = \sum_{k=1}^{N} \left(\mathbf{\Gamma}_{ki}^{-1} - \mathbf{\Gamma}_{ji}^{-1} - \mathbf{\Gamma}_{kj}^{-1} + \mathbf{\Gamma}_{jj}^{-1} \right) d_k \qquad (8)$$

Given that the elements of $\mathbf{\Gamma}^{-1}$ scale with the MSFs of residues (diagonal elements) or the cross-correlations between residue fluctuations (off-diagonal elements), the above equation establishes the link between the signal transduction properties of the protein and its collective dynamics [40]. Note that the commute time $\tau_{ij} = \mathbf{H}_{ij} + \mathbf{H}_{ji}$ assumes an even simpler form, using equation (8) twice, i.e.,

$$\tau_{ij} = \left(\sum_{k=1}^{N} d_k \right) \left(\mathbf{\Gamma}_{ii}^{-1} - 2\mathbf{\Gamma}_{ji}^{-1} + \mathbf{\Gamma}_{jj}^{-1} \right) = \left(\sum_{k=1}^{N} d_k \right) \langle (\Delta \boldsymbol{R}_{ij})^2 \rangle \qquad (9)$$

This equation simply states that the communication between two residues takes longer if their inter-residue distances have higher fluctuations [40]. The inter-residue distances, in turn, are readily evaluated from the difference $\langle (\Delta \boldsymbol{R}_{ij})^2 \rangle = \langle (\Delta \boldsymbol{R}_i)^2 \rangle + \langle (\Delta \boldsymbol{R}_j)^2 \rangle - 2\langle \Delta \boldsymbol{R}_i \bullet \Delta \boldsymbol{R}_j \rangle$, where the respective terms are evaluated using the equations (3) and (4).

NMA of a subsystem coupled to a dynamic environment

If the dynamics of a part of the protein (subsystem, S) in the presence of an environment (E) is of interest, a useful approach is to partition the Hessian into four submatrices [43]:

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{SS} & \mathbf{H}_{SE} \\ \mathbf{H}_{ES} & \mathbf{H}_{EE} \end{pmatrix} \qquad (10)$$

where $\mathbf{H}_{SS}$ is the matrix referring to the subsystem, $\mathbf{H}_{EE}$ to that associated with the interactions within the environment and $\mathbf{H}_{SE}$ (and $\mathbf{H}_{ES}$) to those coupling the subsystem to the environment. An effective Hessian for the subsystem $\mathbf{H}_{SS}^{eff}$ can be constructed from these elements as

$$\mathbf{H}_{SS}^{eff} = \mathbf{H}_{SS} - \mathbf{H}_{SE}\mathbf{H}_{EE}^{-1}\mathbf{H}_{ES} \qquad (11)$$

The NMA of $\mathbf{H}_{SS}^{eff}$ effectively describes the collective dynamics of the subsystem in the presence of coupling to the environment. This approach proved useful in determining the allosteric potential of residues [61] or the location of transition states of chemical reactions by defining a reduced potential energy surface [62].

The above method has been used for evaluating the overlap between the collective modes of different family members in different environments. The cumulative overlap (expressed in terms of percentage in Table 1) between subsets of modes is evaluated using equation (7) where a double summation over the particular subsets of modes of interest, e.g. top ranking 10 modes of the two systems.

All figures depicting molecular structures have been generated with the VMD visualization software [63].

Supporting Information

Figure S1 Representation of the movement undergone by EcNAGK in ANM modes 1,3 and 5. Ribbon diagrams represent deformed conformations generated using Eq 5. Different perspectives of the enzyme (see rotation of the reference axes) have been displayed to highlight the main deformation of each mode: front view (mode 1), lateral view (mode 3) and bottom view (mode 5). C and N domains are colored in red and blue respectively.
Found at: doi:10.1371/journal.pcbi.1000738.s001 (3.62 MB TIF)

Figure S2 Movement of active site residues between open and closed conformers along the 3rd ANM mode accessible to the open form. The position of these residues in different conformations is shown: open conformation (yellow), intermediate positions (green and blue) and closed conformation (atom-colored). (A) Color-coded ribbon diagram for motions along the 3rd mode (generated with the ANM web server[1] and Pymol[2]). (B) Movement of catalytic residues with respect to the ATP analogue. (C) Movement of ATP binding residues with respect to the nucleotide. (D) Movement of NAG binding residues with respect to NAG. 1. Eyal E, Yang LW, Bahar I (2006) Anisotropic network model: systematic evaluation and a new web interface. Bioinformatics 22: 2619–2627. 2. DeLano WL (2002) The PyMOL Molecular Graphics System. San Carlos, CA: DeLano Scientific.
Found at: doi:10.1371/journal.pcbi.1000738.s002 (2.06 MB TIF)

Figure S3 Movement of NAG binding residues between open and closed conformers along the 5th ANM mode accessible to the open form. (A) The position of these residues in different conformations is shown: open conformation (yellow), intermediate positions (green and blue) and closed conformation (atom-colored). (B) Schematic representation of the conformational change of the hydrophobic pocket at the NAG binding site along the 5th ANM mode. Red dots show the interaction sites between NAG and residues R66 and 158. Residues R66 and L65 move concertedly toward the interaction site of N158, which fixes the size of the hydrophobic pocket.
Found at: doi:10.1371/journal.pcbi.1000738.s003 (0.64 MB TIF)

Acknowledgments

The authors thank Drs. V. Rubio, F. Gil-Ortiz and S. Ramón-Maiques for providing the PDB file of the open structure of *Ec*NAGK prior to release

and other crystallographic information. EM acknowledges fruitful discussions with members of Bahar lab and also thanks Dr. Chennubhotla for assistance with implementation issues.

Author Contributions

Conceived and designed the experiments: EM RC IB. Performed the experiments: EM. Analyzed the data: EM RC IB. Contributed reagents/materials/analysis tools: EM RC IB. Wrote the paper: EM RC IB.

References

1. Changeux JP, Edelstein SJ (2005) Allosteric mechanisms of signal transduction. Science 308: 1424–1428.
2. Tobi D, Bahar I (2005) Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. Proc Natl Acad Sci U S A 102: 18908–18913.
3. Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. Nature 450: 964–972.
4. Lange OF, Lakomek NA, Fares C, Schroder GF, Walter KFA, et al. (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science 320: 1471–1475.
5. Bakan A, Bahar I (2009) The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. Proc Natl Acad Sci U S A 106: 14349–14354.
6. Yang L, Song G, Carriquiry A, Jernigan RL (2008) Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. Structure 16: 321–330.
7. Yang LW, Eyal E, Bahar I, Kitao A (2009) Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics. Bioinformatics 25: 606–614.
8. Koshland DE (1958) Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci U S A 44: 98–104.
9. Monod J, Wyman J, Changeux JP (1965) On nature of allosteric transitions - A plausible model. J Mol Biol 12: 88–&.
10. Okazaki KI, Takada S (2008) Dynamic energy landscape view of coupled binding and protein conformational change: Induced-fit versus population-shift mechanisms. Proc Natl Acad Sci U S A 105: 11182–11187.
11. Eisenmesser EZ, Bosco DA, Akke M, Kern D (2002) Enzyme dynamics during catalysis. Science 295: 1520–1523.
12. Wolf-Watz M, Thai V, Henzler-Wildman K, Hadjipavlou G, Eisenmesser EZ, et al. (2004) Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. Nat Struct Mol Biol 11: 945–949.
13. Villa J, Warshel A (2001) Energetics and dynamics of enzymatic reactions. J Phys Chem B 105: 7887–7907.
14. Warshel A, Sharma PK, Kato M, Xiang Y, Liu HB, et al. (2006) Electrostatic basis for enzyme catalysis. Chem Rev 106: 3210–3235.
15. James LC, Tawfik DS (2005) Structure and kinetics of a transient antibody binding intermediate reveal a kinetic discrimination mechanism in antigen recognition. Proc Natl Acad Sci U S A 102: 12730–12735.
16. Sullivan SM, Holyoak T (2008) Enzymes with lid-gated active sites must operate by an induced fit mechanism instead of conformational selection. Proc Natl Acad Sci U S A 105: 13829–13834.
17. Bahar I, Rader AJ (2005) Coarse-grained normal mode analysis in structural biology. Curr Opin Struct Biol 15: 586–592.
18. Agarwal PK (2006) Enzymes: An integrated view of structure, dynamics and function. Microb Cell Fact 5.
19. Zheng WJ, Brooks BR, Thirumalai D (2006) Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc Natl Acad Sci U S A 103: 7664–7669.
20. Zen A, de Chiara C, Pastore A, Micheletti C (2009) Using dynamics-based comparisons to predict nucleic acid binding sites in proteins: an application to OB-fold domains. Bioinformatics 25: 1876–1883.
21. Zen A, Carnevale V, Lesk AM, Micheletti C (2008) Correspondences between low-energy modes in enzymes: Dynamics-based alignment of enzymatic functional families. Protein Sci 17: 918–929.
22. Cui Q, Bahar I, eds (2006) Normal Mode Analysis: Theory and applications to biological and chemical systems. Chapman & Hall, CRC Press, Taylor & Francis Group ed. Boca Raton.
23. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, et al. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80: 505–515.
24. Tama F, Sanejouand YH (2001) Conformational change of proteins arising from normal mode calculations. Protein Eng 14: 1–6.
25. Eyal E, Yang LW, Bahar I (2006) Anisotropic network model: systematic evaluation and a new web interface. Bioinformatics 22: 2619–2627.
26. Krebs WG, Alexandrov V, Wilson CA, Echols N, Yu HY, et al. (2002) Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. Proteins-Structure Function and Genetics 48: 682–695.
27. Xu CY, Tobi D, Bahar I (2003) Allosteric changes in protein structure computed by a simple mechanical model: Hemoglobin T <-> R2 transition. J Mol Biol 333: 153–168.
28. Thomas A, Field MJ, Perahia D (1996) Analysis of the low-frequency normal modes of the R state of aspartate transcarbamylase and a comparison with the T state modes. J Mol Biol 261: 490–506.
29. Mouawad L, Perahia D (1996) Motions in hemoglobin studied by normal mode analysis and energy minimization: Evidence for the existence of tertiary T-like, quaternary R-like intermediate structures. J Mol Biol 258: 393–410.
30. Rueda M, Chacon P, Orozco M (2007) Thorough validation of protein normal mode analysis: A comparative study with essential dynamics. Structure 15: 565–575.
31. Liu L, Koharudin LMI, Gronenborn AM, Bahar I (2009) A comparative analysis of the equilibrium dynamics of a designed protein inferred from NMR, X-ray, and computations. Proteins-Structure Function and Bioinformatics 77: 927–939.
32. Ramon-Maiques S, Marina A, Gil-Ortiz F, Fita I, Rubio V (2002) Structure of acetylglutamate kinase, a key enzyme for arginine biosynthesis and a prototype for the amino acid kinase enzyme family, during catalysis. Structure 10: 329–342.
33. Gil-Ortiz F, Ramon-Maiques S, Fita I, Rubio V (2003) The course of phosphorus in the reaction of N-acetyl-L-glutamate kinase, determined from the structures of crystalline complexes, including a complex with an AlF4- transition state mimic. J Mol Biol 331: 231–244.
34. Ramon-Maiques S, Fernandez-Murga ML, Gil-Ortiz F, Vagin A, Fita I, et al. (2006) Structural bases of feed-back control of arginine biosynthesis, revealed by the structures of two hexameric N-acetylglutamate kinases, from Thermotoga maritima and Pseudomonas aeruginosa. J Mol Biol 356: 695–713.
35. Fernandez-Murga ML, Rubio V (2008) Basis of arginine sensitivity of microbial N-acetyl-L-glutamate kinases: Mutagenesis and protein engineering study with the Pseudomonas aeruginosa and Escherichia coli enzymes. J Bacteriol 190: 3018–3025.
36. Marco-Marin C, Ramon-Maiques S, Tavarez S, Rubio V (2003) Site-directed mutagenesis of Escherichia coli acetylglutamate kinase and aspartokinase III probes the catalytic and substrate-binding mechanisms of these amino acid kinase family enzymes and allows three-dimensional modelling of aspartokinase. J Mol Biol 334: 459–476.
37. Yang LW, Bahar I (2005) Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes. Structure 13: 893–904.
38. Roca M, Liu H, Messer B, Warshel A (2007) On the relationship between thermal stability and catalytic power of enzymes. Biochemistry 46: 15076–15088.
39. Haas D, Leisinger T (1975) N-Acetylglutamate 5-phosphotransferase of Pseudomonas Aeruginosa - Catalytic and regulatory properties. Eur J Biochem 52: 377–383.
40. Chennubhotla C, Bahar I (2007) Signal propagation in proteins and relation to equilibrium fluctuations. PLoS Comput Biol 3: 1716–1726.
41. Kondrashov DA, Van Wynsberghe AW, Bannen RM, Cui Q, Phillips GN (2007) Protein structural variation in computational models and crystallographic data. Structure 15: 169–177.
42. Holm L, Kaariainen S, Rosenstrom P, Schenkel A (2008) Searching protein structure databases with DaliLite v.3. Bioinformatics 24: 2780–2781.
43. Zheng WJ, Brooks BR (2005) Probing the local dynamics of nucleotide-binding pocket coupled to the global dynamics: Myosin versus kinesin. Biophys J 89: 167–178.
44. Ramon-Maiques S, Marina A, Uriarte M, Fita I, Rubio V (2000) The 1.5 angstrom resolution crystal structure of the carbamate kinase-like carbamoyl phosphate synthetase from the hyperthermophilic archaeon Pyrococcus furiosus, bound to ADP, confirms that this thermostable enzyme is a carbamate kinase, and provides insight into substrate binding and stability in carbamate kinases. J Mol Biol 299: 463–476.
45. Marco-Marin C, Gil-Ortiz F, Rubio V (2005) The crystal structure of Pyrococcus furiosus UMP kinase provides insight into catalysis and regulation in microbial pyrimidine nucleotide biosynthesis. J Mol Biol 352: 438–454.
46. Tama F, Brooks CL (2006) Symmetry, form, and shape: Guiding principles for robustness in macromolecular machines. Annu Rev Biophys Biomol Struct 35: 115–133.
47. Nicolay S, Sanejouand YH (2006) Functional modes of proteins are among the most robust. Phys Rev Lett 96.
48. Hinsen K, Petrescu AJ, Dellerue S, Bellissent-Funel MC, Kneller GR (2000) Harmonicity in slow protein dynamics. Chem Phys 261: 25–37.
49. Li GH, Cui Q (2002) A coarse-grained normal mode approach for macromolecules: An efficient implementation and application to Ca2+-ATPase. Biophys J 83: 2457–2474.
50. Suhre K, Sanejouand YH (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res 32: W610–W614.
51. Bahar I, Atilgan AR, Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding & Design 2: 173–181.

52. Kundu S, Melton JS, Sorensen DC, Phillips GN (2002) Dynamics of proteins in crystals: Comparison of experiment with simple models. Biophys J 83: 723–732.
53. Haliloglu T, Bahar I, Erman B (1997) Gaussian dynamics of folded proteins. Phys Rev Lett 79: 3090–3093.
54. Riccardi D, Cui Q, Phillips GN (2009) Application of Elastic Network Models to Proteins in the Crystalline State. Biophys J 96: 464–475.
55. Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. Journal of Molecular Modeling 7: 306–317.
56. Goodey NM, Benkovic SJ (2008) Allosteric regulation and catalysis emerge via a common route. Nat Chem Biol 4: 474–482.
57. Agarwal PK, Billeter SR, Rajagopalan PTR, Benkovic SJ, Hammes-Schiffer S (2002) Network of coupled promoting motions in enzyme catalysis. Proc Natl Acad Sci U S A 99: 2794–2799.
58. Suel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 10: 59–69.
59. Bode C, Kovacs IA, Szalay MS, Palotai R, Korcsmaros T, et al. (2007) Network analysis of protein dynamics. FEBS Lett 581: 2776–2782.
60. Chennubhotla C, Bahar I (2006) Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. Mol Syst Biol 2.
61. Ming DM, Wall ME (2006) Interactions in native binding sites cause a large change in protein dynamics. J Mol Biol 358: 213–223.
62. Anglada JM, Besalu E, Bofill JM, Crehuet R (2001) On the quadratic reaction path evaluated in a reduced potential energy surface model and the problem to locate transition states. J Comput Chem 22: 387–406.
63. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. Journal of Molecular Graphics 14: 33–38.
64. DeLano WL (2002) The PyMOL Molecular Graphics System. San CarlosCA: DeLano Scientific.

4.2.2. Oligomerization effecs on large-amplitude dynamics

A similar approach to that used for comparing low-frequency modes in the previous study was adopted to analyze the effects of oligomerization on the dynamics of AAK members, which exhibit different oligomeric states (dimeric and hexameric). This study showed how the oligomerization confers new cooperative modes of motion that exploit the intrinsic dynamics of the subunits and shed light into the allosteric regulation of hexameric NAGK and UMPK.

By comparing the low-frequency modes of the component subunits of *Ec*NAGK, CK and hexameric NAGK with those of the oligomer, we found that large-amplitude motions intrinsically accessible by the subunits are preserved in the oligomeric state in a high extent. Furthermore, new cooperative modes provided by the design of the interface between subunits are found in the oligomer In particular, the conformational change associated to the allosteric regulation of hexameric NAGK, which involves rigid-body motions of the dimeric subunits, is determined by the structure of the interface between the *Ec*NAGK-like dimers that build the hexamer.

Another case of interest was that of UMPK (hexamer) which presents an assembly of the dimeric subunits that is strikingly different to that present in the rest of the AAK family. By studying the low-frequency modes of the dimeric subunit we found that the unique interface displayed by UMPK allows rigid-body motions of the monomeric subunits, which are not allowed by the *Ec*NAGK-like dimeric arquitecture and have been observed to be involved in the allosteric regulation of the enzyme.

A detailed presentation of the results and methodologies used in this study can be found in the article: *Changes in dynamics upon oligomerization regulate substrate binding and allostery in Amino Acid Kinase family members* (2011) PLoS Comput. Biol., 7: e1002201.

OPEN ACCESS Freely available online PLOS COMPUTATIONAL BIOLOGY

Changes in Dynamics upon Oligomerization Regulate Substrate Binding and Allostery in Amino Acid Kinase Family Members

Enrique Marcos[1], Ramon Crehuet[1]*, Ivet Bahar[2]*

1 Department of Biological Chemistry and Molecular Modelling, IQAC-CSIC, Barcelona, Spain, **2** Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

Abstract

Oligomerization is a functional requirement for many proteins. The interfacial interactions and the overall packing geometry of the individual monomers are viewed as important determinants of the thermodynamic stability and allosteric regulation of oligomers. The present study focuses on the role of the interfacial interactions and overall contact topology in the dynamic features acquired in the oligomeric state. To this aim, the collective dynamics of enzymes belonging to the amino acid kinase family both in dimeric and hexameric forms are examined by means of an elastic network model, and the softest collective motions (i.e., lowest frequency or global modes of motions) favored by the overall architecture are analyzed. Notably, the lowest-frequency modes accessible to the individual subunits in the absence of multimerization are conserved to a large extent in the oligomer, suggesting that the oligomer takes advantage of the intrinsic dynamics of the individual monomers. At the same time, oligomerization stiffens the interfacial regions of the monomers and confers new cooperative modes that exploit the rigid-body translational and rotational degrees of freedom of the intact monomers. The present study sheds light on the mechanism of cooperative inhibition of hexameric *N*-acetyl-L-glutamate kinase by arginine and on the allosteric regulation of UMP kinases. It also highlights the significance of the particular quaternary design in selectively determining the oligomer dynamics congruent with required ligand-binding and allosteric activities.

Citation: Marcos E, Crehuet R, Bahar I (2011) Changes in Dynamics upon Oligomerization Regulate Substrate Binding and Allostery in Amino Acid Kinase Family Members. PLoS Comput Biol 7(9): e1002201. doi:10.1371/journal.pcbi.1002201

Editor: Michael Gilson, University of California San Diego, United States of America

Received March 9, 2011; **Accepted** August 4, 2011; **Published** September 29, 2011

Funding: This work was supported by grants from the JAE-predoc programme of the Consejo Superior de Investigaciones Cientificas (CSIC), the Spanish MEC (CTQ2009-08223) and the Catalan AGAUR (2005SGR00111). IB gratefully acknowledges support from NIH project # 5R01LM007994-06. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bahar@ccbb.pitt.edu (IB); rcsqtc@iqac.csic.es (RC)

Introduction

The biological function of proteins is usually enabled by their dynamics under native state conditions, which, in turn, is encoded by their 3-dimensional (3D) structure. Unraveling this functional code has been the aim of many experimental and theoretical studies [1–9]. In particular the slow conformational dynamics of proteins in the micro-to-milliseconds time scale has been pointed out to be consistent with the changes in structure or domain/subunit movements observed between the substrate-bound and -unbound forms of enzymes [4–7,10], and potentially limit the catalytic turnover rates of enzymes [11–14]. The quaternary structure of oligomeric proteins adds another layer of complexity to this code as the assembly of the subunits entails additional constraints while possibly inducing new types of collective motions. The structural hierarchy in oligomers indeed gives rise to a wide diversity of dynamical events [15]. For instance, in allosteric proteins, such as the paradigmatic hemoglobin [16,17], the coupling between the internal dynamics of the subunits and the intrinsic ability of pairs of dimers to undergo concerted reorientations with respect to each other underlies the cooperative response to ligand binding [18–20]. Analysing the slow conformational dynamics thus emerges as a crucial step towards understanding the structure-function code in oligomeric proteins.

Two classical models have been broadly used in the literature to interpret the conformational changes observed upon ligand binding: the Koshland-Némethy-Filmer (KNF) model [21] where the ligand 'induces' a conformational change in the allosteric protein, in line with the classical induced fit model, and the Monod-Wyman-Changeux (MWC) model [22] where the ligand selects from amongst those pre-existing conformers accessible by the *intrinsic* dynamics of the 3D structure. The former is usually a stepwise process, while the latter is all-or-none. The experimentally observed structural changes appear to result from a combination of intrinsic and induced effects: the intrinsic dynamics of the protein prior to substrate binding is essential to enabling cooperative changes in structure, while induced motions, usually more localized, help optimize and stabilize the bound conformers [4,23].

Protein-protein interfaces are usually characterized by their size, shape complementarity and hydrophobicity [24,25]. The dynamics at the interfacial residues are usually given little attention, although the functional significance of the structural changes triggered by complex formation or oligomerization is widely recognized. The interface between subunits often plays a key role in mediating the activity of each monomeric subunit [25]. Protein-protein interactions provide, not only thermodynamic stability to the folded state of the subunit in the complex (or assembly), but

Author Summary

Protein function requires a three-dimensional structure with specific dynamic features for catalytic and binding events, and, in many cases, the structure results from the assembly of more than one polypeptide chain (also called monomer or subunit) to form an *oligomer* or *multimer*. Proteins such as hemoglobin or chaperonin GroEL are oligomers formed by 2 and 14 subunits, respectively, whereas virus capsids are multimers composed of hundreds of monomers. In these cases, the architecture of the interface between the subunits and the overall assembly geometry are essential in determining the functional motions that these sophisticated structures are able to perform under physiological conditions. Here we present results from our computational study of the large-amplitude motions of dimeric and hexameric proteins that belong to the Amino Acid Kinase family. Our study reveals that the monomers in these oligomeric proteins are arranged in such a way that the oligomer inherits the intrinsic dynamic features of its components. The packing geometry additionally confers the ability to perform highly cooperative conformational changes that involve all monomers and enable the biological activity of the multimer. The study highlights the significance of the quaternary design in favoring the oligomer dynamics that enables ligand-binding and allosteric regulation functions.

also a new spectrum of collective motions. Furthermore, the oligomeric arrangement provides an efficient means of communication that may modulate allosteric regulation [19]. The present study focuses on the following questions: (1) Is the intrinsic dynamics of the component subunit modified by the oligomerization process, and if so, in which ways? (2) What is the role of interfacial interactions and overall contact topology in the functional dynamics of the oligomer and, in particular, in signal transduction or allosteric communication?

The effect of multimerization on protein dynamics is investigated here in the context of the Amino Acid Kinase (AAK) family of enzymes. Members of this family have different degrees of oligomerization (Figure 1). Rubio and co-workers have significantly contributed to our current knowledge of this family of enzymes: they have resolved the X-ray structures of most family members [26-33] and suggested a shared mechanism of action on the basis of their sequence and folding similarities [28]. This mechanism was elucidated by our recent computational study of the softest modes of motion intrinsically accessible to different members of the AAK family of proteins [34].

The most exhaustively studied member of the AAK family is *N*-acetyl-L-glutamate kinase (NAGK) (Figure 1A). NAGK phosphorylates the amino acid *N*-acetyl-L-glutamate (NAG) in the bacterial route of arginine biosynthesis. In many organisms, NAG phosphorylation is the controlling step of the route, as NAGK is feedback inhibited by the end product arginine. Rubio and co-workers [30] characterized the structures of two hexameric NAGKs (from *Thermotoga maritima* (Figure 1B) and *Pseudomonas aeruginosa*) that are cooperatively inhibited by arginine [35]. In *Escherichia coli*, NAGK (*Ec*NAGK) is homodimeric and arginine-insensitive (Figure 1A). Indeed, several studies have proven that the hexameric arrangement is a requirement for the cooperative inhibition by arginine [30,36]. The distinctive feature of this biosynthetic route in bacteria is that it produces *N*-acetylated intermediates, in contrast to mammals that yield non-acetylated intermediates. This turns NAGK into a potential target for antibacterial drugs by selective inhibition. Another member of the AAK family is carbamate kinase (CK; Figure 1C). CK catalyses the formation of ATP from ADP and carbamoyl phosphate (CP; a precursor of arginine and pyrimidine bases), and undergoes a substantial change in its structure upon substrate binding [37]. A third member is the hexameric UMP kinase (UMPK) (Figure 1D). UMPK catalyzes the reaction ATP + UMP $\rightleftharpoons$ ADP + UDP to yield uridine diphosphate (UDP). It is involved in the multistep synthesis of UTP, being regulated by the allosteric activator GTP and inhibited by UTP itself. Its monomer fold is very similar to the rest of family members, but presents a strikingly different assembly of the subunits that has not been explained so far.

Notably, while the AAK family members do not exist in monomeric form, they share the same monomeric fold. This commonly shared monomeric fold is stabilized by oligomerization. The selection of a common monomeric fold in different oligomers suggests that that particular architecture possesses structure-encoded dynamic features that are exploited for enzymatic activity in oligomeric state. It is essential to analyze what the intrinsic dynamics of the monomeric units are, and to what extent, if any, they are maintained in the oligomeric state, or how they are coupled to, or complement, the dynamics of the biologically active (oligomeric) state. Calculations are thus performed for the monomeric fold alone as well as the monomer in the context of different oligomeric states, and the intact oligomers. As will be shown below, the oligomers do maintain some intrinsic dynamic features of the monomeric units, while the different assembly geometries of the monomers give rise to global motions uniquely defined for the particular oligomerization states. The method of analysis presented here is applicable to any protein that functions in different multimeric states. The effect of oligomerization on the dynamics of the component subunits can be experimentally examined provided that the protein exists in monomeric and different oligomeric states, which, in turn, may be controlled by environmental conditions [38] and few mutations at the protein surface [39]. However, such studies may be challenging in practice, and a computational examination emerges as an alternative promising tool.

The most collective movements of biomolecular systems, also called the *global* modes of motions, can be determined using Elastic Network Models (ENMs) in conjunction with Normal Mode Analysis (NMA) at very low computational cost. A wealth of studies have shown the robustness of the global modes predicted by the ENMs (e.g., by the anisotropic network model, ANM [40,41]) and their close relevance to experimentally observed structural transitions related to ligand binding [4-6,10,18,41–46], or to the essential modes extracted from converged molecular dynamics (MD) simulations [47–49]. The global modes are the low-frequency modes extracted from NMA, also referred to as *slow* modes. They correspond to large-amplitude motions taking place at long timescales (e.g. microseconds to milliseconds); and they are also called *soft* modes due to their lower energy cost associated with a given level of fluctuation away from the equilibrium state, compared to other modes. Given their robustness and efficiency, ENMs are uniquely suited for exploring the collective motions and allostery in oligomers. Previous such studies have highlighted the significance of multimeric arrangement in defining the collective dynamics [50–54].

The present study adds new evidences to the role played by multimerization in defining functional dynamics. First, we contrast the low-frequency modes favoured by the *Ec*NAGK and *Pf*CK monomers to those preferentially selected by the corresponding dimers. Secondly, the modes of the monomeric and dimeric components of hexameric *Tm*NAGK are compared to those collectively accessible in the hexameric form. Third, a detailed

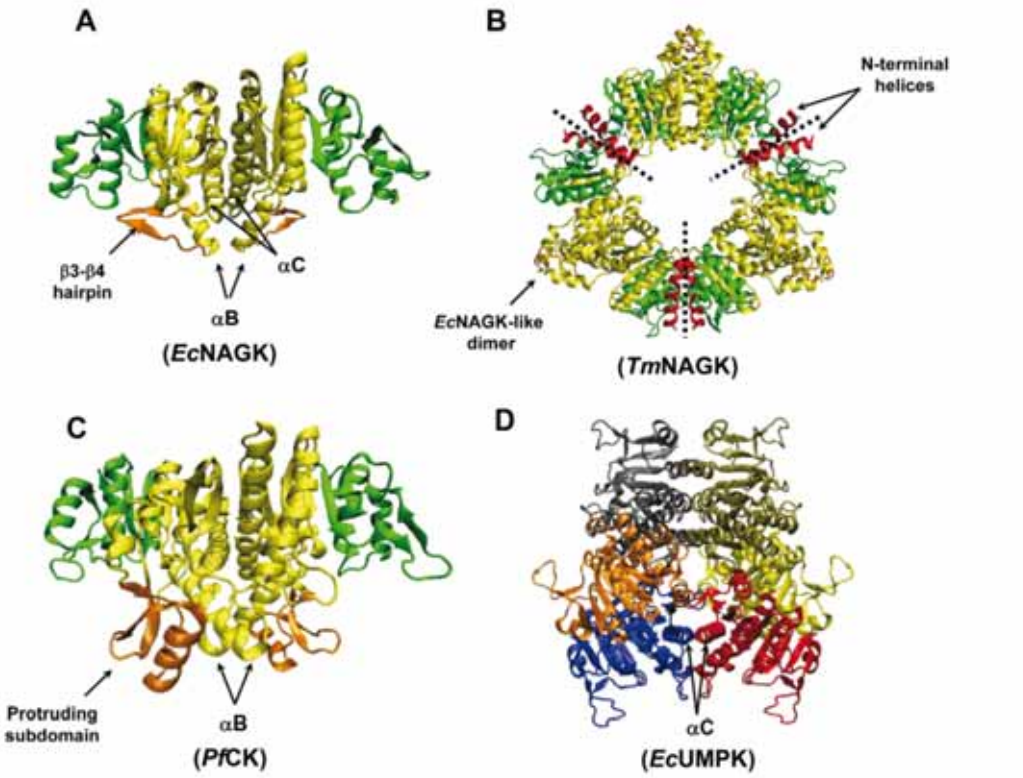


Figure 1. AAK family enzymes examined in the present study. (A) NAGK from *Escherichia coli* (*Ec*NAGK), (B) NAGK from *Thermotoga maritime* (*Tm*NAGK), (C) CK from *Pyrococcus furiosus* (*Pf*CK), (D) UMPK from *Escherichia coli* (*Ec*UMPK). Panels A, B and C show the ATP binding domains in green and N domains in yellow. The NAG-binding sites in *Ec*NAGK (β3–β4 hairpin) and the CK-binding site in *Pf*CK (protruding subdomain (PS) composed of the strand β5, helix D and hairpin β6-β7) are colored orange. The B helices of these two enzymes build part of the intersubunit surface and are very close to the N-domain binding sites. The N-terminal helices of *Tm*NAGK (red) interlink three *Ec*NAGK-like dimers (delimited by dotted lines). This hexameric enzyme is indeed regarded as a trimer of *Ec*NAGK-like dimers. The UMPK is colored by chains. αC helices indicated in panels A and D highlight the difference in the assembly of the monomeric subunits between the two structures.
doi:10.1371/journal.pcbi.1002201.g001

analysis of the softest modes accessible to the *Ec*UMPK dimeric form is presented to shed light onto the role played by different dimeric assemblies found in the AAK family in selecting the functional motions of the family members. Overall, the different designs of interfaces and assembly geometries observed among the members of the AAK family are shown to practically define the collective modes that are being exploited by the oligomers for achieving their particular activities, including substrate binding and allosteric regulation.

Results/Discussion

Soft modes intrinsically accessible to the monomer are selectively utilized or obstructed in compliance with the specific substrate-binding properties of the dimer: *Ec*NAGK vs *Pf*CK

How does the intrinsic dynamics of the monomeric subunits affect the oligomerization process or *vice versa*? To what extent the intrinsic dynamics of the monomers prevail in the oligomers? Or to what extent they are perturbed by oligomerization? To analyse these issues, we have first compared the low-frequency ANM modes of the dimeric *Pf*CK and *Ec*NAGK with those of their respective monomers. The two enzymes exhibit close structural similarities (Figure 2). Their sequence identity is 24%, and their ATP-binding site and catalytic sites exhibit similar structural features. In fact, our previous comparative analysis of their collective dynamics showed that the slowest three ANM modes, which essentially modulate the opening/closure of the ATP-binding site, are commonly shared between these two enzymes; and they yield an overlap of 0.75 with the experimentally observed reconfiguration from open to closed state of NAGK [34].

The main structural difference between *Pf*CK and *Ec*NAGK, on the other hand, resides in their amino acid substrate binding site, and here we focus on the softest modes that control those sites. In *Ec*NAGK, the β3–β4 hairpin serves as the lid of the NAG binding site and interlinks helices B and C, which are key components of the interface (Figure 1A); in *Pf*CK (Figure 1C), a

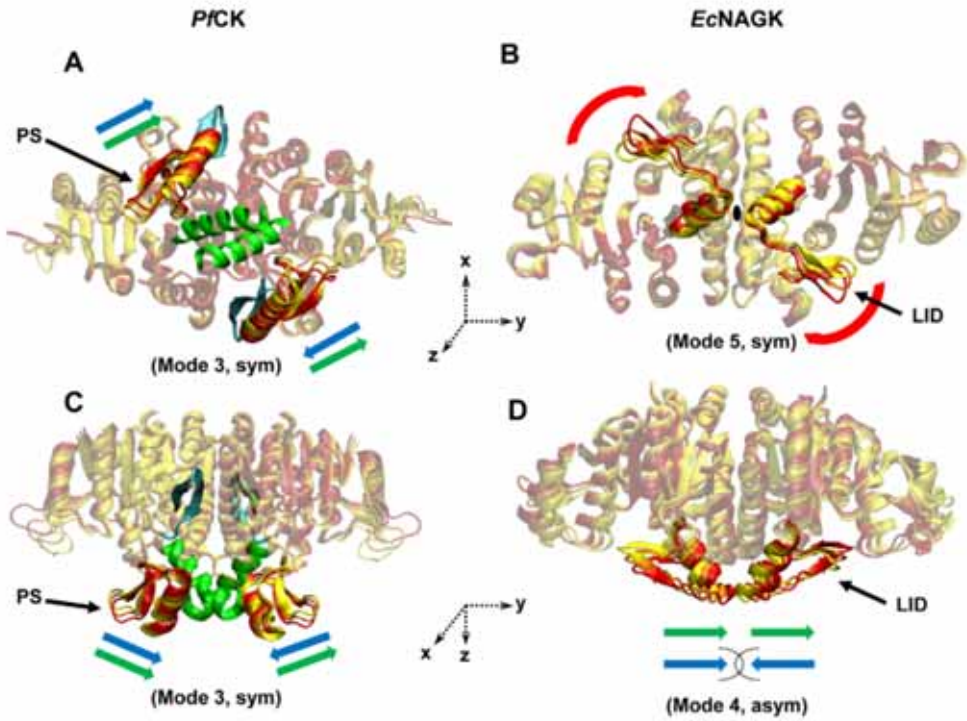


Figure 2. Dynamics of the substrate-binding sites on *Pf*CK and *Ec*NAGK. Panels A and C are bottom and lateral views of *Pf*CK. Panels B and D are bottom and lateral views of *Ec*NAGK. Each panel shows different conformations (in yellow, orange and red) along a given mode (the number and the symmetry are specified in parenthesis). The structural elements involved in substrate-binding are highlighted by the brighter colors: PS in *Pf*CK and the β3–β4 hairpin (lid) in *Ec*NAGK. In *Pf*CK, helix B (green) and β10–β11 hairpin (cyan) are key structural elements at the dimer interface. In *Ec*NAGK, helix B, which is connected to the β3–β4 hairpin, is also highlighted. Green and blue arrows indicate the mechanisms of the modes that induce asymmetric and symmetric opening/closure at the substrate-binding site, respectively. In panels A and C, the green arrows show that the corresponding asymmetric movements of PSs are also allowed (4th mode). In panel B, the black ellipse displayed at the interface shows the axis of rotation (*z*-axis), normal to the plane of the figure. The blue arrows in panel D show that the opposite movement of β3-β4 hairpins is not allowed due to the steric clashes. For better visualization of these modes see Videos S1, S2, S3 and S4.
doi:10.1371/journal.pcbi.1002201.g002

subdomain protruding away from the interface serves as the lid of the CP binding site. This subdomain (PS) is formed by the strand β5, helix αD and hairpin β6–β7. Both lids exhibit significant conformational changes closely linked to substrate binding, as shown by the crystallographic studies performed by Rubio and co-workers [27,32]. Among the ANM modes that affect the substrate-binding sites, those simultaneously leading to closure/opening of the substrate-binding site in both subunits will be called symmetrical modes, and others, asymmetrical (Figure 2).

Description of the modes

In *Ec*NAGK, the symmetrical opening/closure of the substrate-binding sites is enabled by the 5th mode (red arrows in Figures 2B and 2D; see Video S1), whereas the corresponding asymmetrical motion takes place in the 4th (green arrows) mode (Video S2). Note that our previous work [34] showed that ANM modes 1–3 were instrumental in accommodating the structural changes at the ATP-binding site, but had practically no effect on the NAG-binding site. This nicely illustrates how the enzyme takes advantage of different types of motions accessible to its native structure for achieving different types of functional motions. In mode 5, the two β3–β4 hairpins (Figure 1A), the lids of the NAG-binding sites, undergo an almost rigid-body rotation about the dyadic (*z*-) axis of the molecule while the ATP binding domains undergo smaller but coupled anticorrelated rotations. On the other hand, the asymmetrical motion (mode 4) induces a translation along the *y* axis in both lids, along with the C-terminal part of the two helices B which are connected to the lids. No symmetric opening/closing of the lids is observed about the *y*-axis because these movements would be prohibited by steric clashes between the two B-helices (blue arrows in Figure 2D). Rotational motions about the *z*-axis, on the other hand, are favored by the overall architecture of the dimeric enzyme. Indeed, tight interfacial interaction between the two B-helices is considered to be a key element for the stability of the dimer [28]. The interfacial region thus coincides with the central hinge site that mediates the opening/closing of the two monomers. This example emphasizes the effect of inter-subunit surface and topology on the character of the movements allowed/prohibited, or selected, in the oligomer.

As to *Pf*CK, the two substrate-binding subdomains are able to undergo both symmetric (1st and 3rd mode; see Video S3) and asymmetric (4th mode; see Video S4) motions because these two subdomains protrude away from the interface and their rotational rigid-body motions are not constrained by potential clashes between the adjacent B-helices. Indeed, the motion is parallel, rather than normal, to the plane defined by the two B-helices, and the two B-helices remain tightly packed and almost immobile in these modes. Notably, the global fluctuations of two PSs on *Pf*CK dimer appear to modulate the access to the substrate-binding sites, suggesting a role in mediating substrate-binding.

Comparison between the monomer and dimer dynamics

The selection of particular modes by *Ec*NAGK for achieving its specific functions (e.g., modes 1 and 3 enabling ATP-binding; and mode 5, substrate binding) [34] raises the following question: is the rotation of the hairpins an acquired mode of motion originating from the topology of the dimer interface and not accessible to the monomer? Or, is it an intrinsic dynamical ability of the monomer that is conserved and exploited in the dimer? To address this issue, we compared the modes obtained for the isolated monomer with those of the monomer in the dimer, using the subsystem/environment coupling method described in the Methods. The monomer is the *subsystem*, and the second monomer stands for the *environment* in this case. For the sake of clarity, herein the modes that include the coupling to the environment are indicated with a superscript, i.e., monomer$^{(dimer)}$ refers to the behaviour of the monomer within the dimer.

The results are presented in Figure 3 (and Supplementary Tables S1 and S2). Therein the overlaps between the eight lowest-frequency modes accessible to the monomer in the isolated state (y-axis) and within the dimer (x-axis) are displayed for *Ec*NAGK (panel A) and *Pf*CK (panel B), and Tables S1 and S2 lists the corresponding values. The orange-red entries along the diagonal in panel A demonstrate that the modes intrinsically accessible to the *Ec*NAGK are closely maintained in the dimeric enzyme. Notably, both the order of the modes (i.e., their relative frequency and size, as defined by the respective eigenvalues), and their shapes are closely conserved.

The picture is different in the case of the *Pf*CK dimer (panel B). While in *Ec*NAGK all of the top-ranking seven modes are maintained with an overlap of 0.70 or above, in *Pf*CK significantly fewer global modes favored by the isolated monomer are maintained, and with a weaker correlation and reordering of the modes. Thus, the *Pf*CK monomer dynamics is strongly affected by dimerization. Examination of the individual modes showed that the monomer modes that induce high fluctuations at particular secondary structural elements such as the helix B and the β10–β11 hairpin (shown in cyan in Figures 2A and C) are practically absent in the dimer. As shown in Figure 2 these are key elements at the intersubunit interface, and dimerization imposes high constraints quenching their motion. The intersubunit surface of *Pf*CK (2453 Å^2) [27] is remarkably bigger than that of *Ec*NAGK (1279 Å^2) [28]. This higher surface area, and ensuing closer association of the two monomers, may be partly responsible for the larger perturbation of the intrinsic dynamics of the monomer upon dimerization in *Pf*CK, compared to *Ec*NAGK.

Figure 2 and videos S3 and S4 in the Supporting Information demonstrate that the global motions preferentially undergone by the two PSs in the *Pf*CK dimer induce conformational changes near the substrate-binding site; and Figure 3 shows that the global dimer dynamics departs from that of the isolated monomers. So, dimerization promotes in this case collective motions that affect substrate recognition and/or binding. The PS has been proposed to have evolved, together with the intersubunit interface, to play a key role in the specificity of CK for its substrate carbamate, as opposed to more abundant analogues, i.e., acetate, bicarbonate or acetylphosphate [37]. This conjecture originally inferred from the examina-

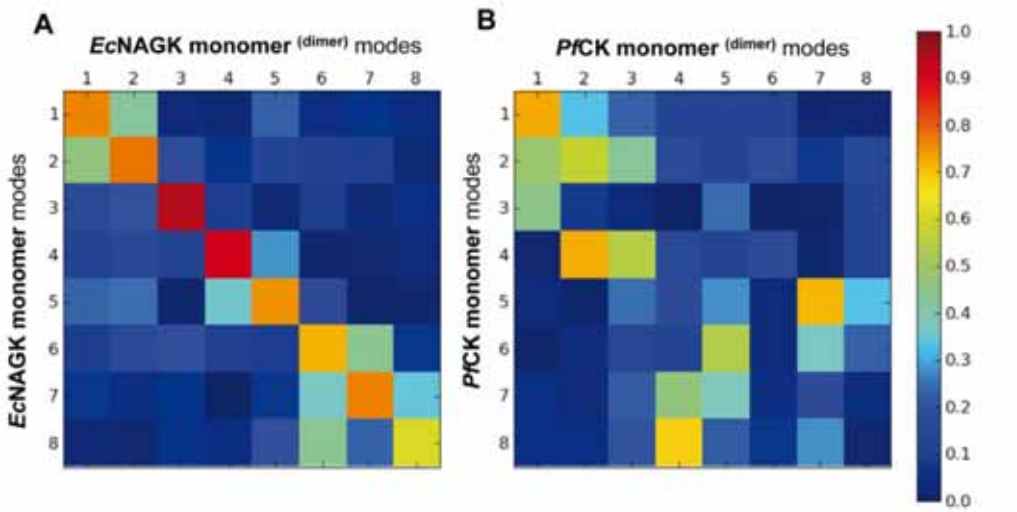


Figure 3. Comparison of the global dynamics of *Ec*NAGK and *Pf*CK monomers in the dimer with those of the isolated monomeric components. Overlaps between the eight slowest modes of the monomers and dimers of (A) *Ec*NAGK and (B) *Pf*CK are shown in the heat map. Dimerization has minimal effect on the intrinsic global dynamics of *Ec*NAGK, while that of *Pf*CK appears to be more strongly affected, presumably due to its larger intersubunit interface.
doi:10.1371/journal.pcbi.1002201.g003

tion of crystal structure alone is supported by our examination of *Pf*CK dynamics. ANM global modes clearly indicate the ability of the PS to undergo movements toward the substrate-binding site, and the enhanced mobility at this particular region may indeed underlie the adaptability of CK to bind its substrate.

Conservation and creation of functional modes: the hexameric *Tm*NAGK

The next case we studied is the hexameric form of the NAGK enzyme from *Thermotoga maritima* (*Tm*NAGK). The higher degree of multimerization of *Tm*NAGK will permit us to contrast the dynamics of the whole enzyme with those of its dimeric and monomeric components.

On the basis of the X-ray crystallographic structure, the hexameric arrangement of *Tm*NAGK is considered to be a trimer of *Ec*NAGK-like dimers [30], herein called the AB dimer (see Figures 1B and 4A). The dimeric scaffolds are interlaced by a mobile N-terminal helix, not present in the dimeric *Ec*NAGK, and organized with a ring shape. An alternative dimeric building block being considered is the one constituted by the two monomers that interlink two adjacent AB dimers, herein called the AF dimer (see Figure 4A). In the present study, we have compared the 20 lowest-frequency modes of the hexamer with those of the monomeric subunit and the two different dimeric building blocks.

The results are presented in the panels B–F of Figure 4. In each panel, the *x*-axis refers to the modes observed in the oligomer (hexamer or dimer), and the y-axis refers to those intrinsically accessible to the components (dimers or monomers) that make these oligomers, e.g., panel B compares the global modes of the AB dimer in the hexamer (x-axis) to those accessible to the AB dimer itself when examined in isolation (y-axis). The comparative examination of these maps discloses two distinctive patterns: panels C and E reveal the conservation of global modes, in general, between the entities that are being compared, while panels B, D and F reveal that about ⅓ of the modes accessible to the substructures when examined in isolation are *not* represented in the assemblies. This behavior is clearly seen, and quantified, by the dashed lines on the maps, which represent a linear fit by weighted least squares regression to the entries that exhibit a correlation of 0.5 of higher. The dashed line in the former groups lies along the diagonal (slope -1.04 and -1.01 in the respective panels C and E), whereas in the latter case, the slope varies as -1.31 (panel B), -1.72 (D) and -1.44 (F).

Let us first examine the 1st group more closely: panel C essentially tells us that the monomers participating in the AB dimer maintain in the dimer their intrinsic dynamics favored by their monomeric architecture. As to panel E, it simply reflects that AF dimer in the hexamer behaves practically in the same way as in

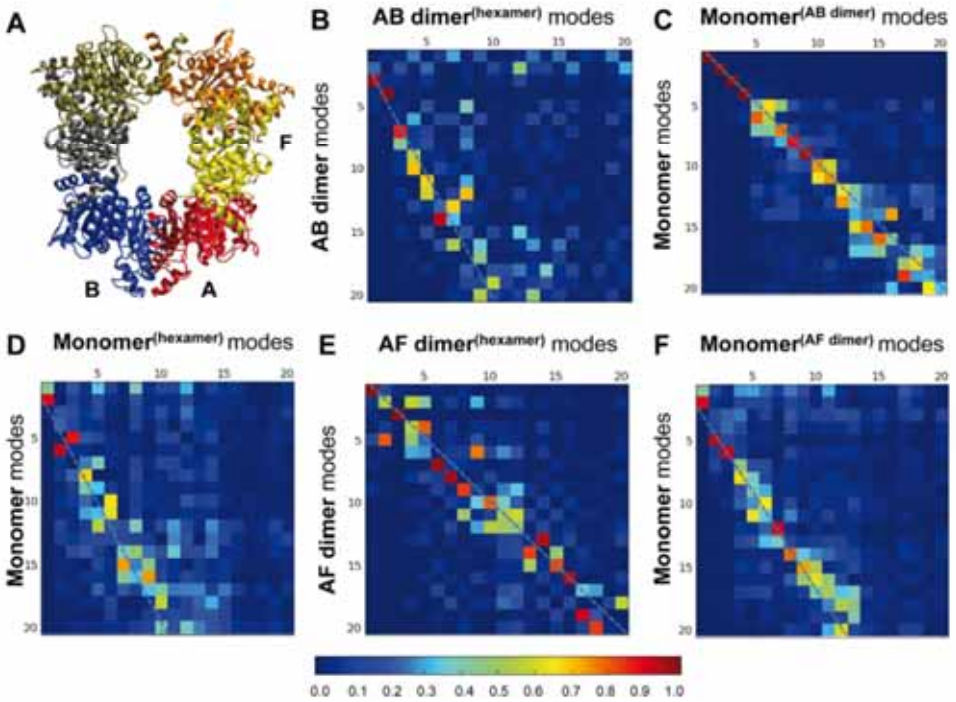


Figure 4. Comparison of the global dynamics of *Tm*NAGK monomers in the hexamer with those of the isolated monomeric and dimeric components. (A) Cartoon representation of *Tm*NAGK, illustrating the packing of the individual subunits, resulting in two different types of intersubunit interfaces represented by those in the AB and AF dimers. (B)-(F) Overlaps between the 20 slowest modes of different building blocks (monomer, AB and AF dimers) and the hexamer, as labelled in the individual heat maps. In all five maps, the weighted linear fit to overlaps above 0.5 is shown by the dashed line.
doi:10.1371/journal.pcbi.1002201.g004

the isolated AF dimer, indicating that multimerization does not alter the global dynamics favored by the AF dimeric structure. In other words, the *Tm*NAGK hexamer exploits the intrinsic dynamics of the AF dimer; and likewise, the AB dimer takes advantage of the structure-encoded dynamics of its monomers. Notably, the top four modes are conserved in this case with a correlation of more than 0.95. This is in agreement with the high conservation of the monomer dynamics in the *Ec*NAGK dimer, as pointed out in Figure 3A, given the structural and dynamical similarities [34] between the AB dimer and *Ec*NAGK.

We now turn our attention to the 2nd group. Here we see the dimer AB in the hexamer which is unable to sample several modes that are accessible to the same dimer in isolation (panel B). Thus, the environment provided by the hexamer constrains the intrinsic dynamics of the AB dimer. Why is the AB dimer rigidified in the hexamer? We note that in the hexamer, these *Ec*NAGK-like (AB) dimers make close, interlacing interactions with the adjacent dimer by swapping their *N*-terminal helices and also making contacts with the *C*-domain, i.e. the interactions of AB-type dimers with the adjacent dimer through the AF interface impose topological constraints that impair several modes in the hexamer (panel B). Likewise, the monomer in the hexameric environment is more restricted than the isolated monomer, such that many modes accessible to the isolated monomer cannot be effectuated in the hexamer (panel D). Given the different degree of conservation of the dynamics of the AB and AF dimers within the hexamer (panels B and E), we can add a complementary perspective to the structural view of *Tm*NAGK as a trimer of *Ec*NAGK-like dimers. The stronger conservation of the dynamics of the AF dimer supports a dynamical view of *Tm*NAGK as a trimer of AF-like dimers.

Finally, it is worth pointing out that the surface area of the AF interface (1186 Å^2) is slightly smaller than that of the AB interface (1381 Å^2) [30]. This might suggest that the monomeric modes would be more severely constrained in the AB dimer, but this does not hold true as explained above. The small difference in the surface area is therefore not sufficient to explain the observed behavior. The major determinant of accessible global motions is not the surface area but the topology of the interfacial contacts, or the overall shape/architecture of the dimer. In the present case, the overall architecture of the hexamer selectively hinders a number of global modes accessible to the AB dimer, while those of the AF dimer are mostly preserved. It is widely accepted that the size of the interface is closely linked to the thermodynamic stability of the oligomer [25,55]. The dynamics of the oligomer, on the other hand, is suggested by the present analysis to be predominantly controlled by the quaternary arrangement and contact topology of the subunits.

New modes of motion and cooperativity

The results discussed above focus on the preservation or the obstruction of the global motions of the subunits upon oligomerization. Nevertheless, in many cases, oligomeric proteins are subject to cooperative processes that regulate the biological activity. This raises the question whether such cooperative processes are linked to new modes of motion unique to oligomeric arrangement.

*Tm*NAGK is cooperatively inhibited by arginine in contrast to the dimeric *Ec*NAGK and *Pf*CK, which do not exhibit an allosteric regulation. The available X-ray crystallographic structure of *Tm*NAGK represents the T state of the enzyme, which is bound to arginine. The apo form of the enzyme (R state) has not been structurally resolved, but the X-ray structure of the same enzyme from *Pseudomonas aeruginosa* (*Pa*NAGK) serves as a suitable model for the R state on the basis of sequence and structural similarities [30]. Taking into account that the transition of *Tm*NAGK between the R and T states is intimately linked to its allosteric regulation, those modes of motion that favor this conformational change will be the most functional. Therefore, the cumulative overlap of the lowest modes with the deformation vector between the R and T states has been calculated. Given that the T and R states correspond to proteins with different sequences, we have structurally aligned the two structures with DALI [56] and used the subsystem/environment coupling method (see *Methods*) to compute the ANM modes of *Tm*NAGK, considering as subsystem those residues of *Tm*NAGK structurally aligned to *Pa*NAGK. Likewise, the deformation vector was calculated for the structurally aligned residues.

Strikingly, a single non-degenerate mode (6th) accessible to *Tm*NAGK is found to describe 75% of the R↔T deformation (see Figure 5D showing the cumulative overlap). A deeper analysis of this mode can shed light on the structural origin of the functionality of this enzyme. The aim is to ascertain whether this mode arises from the intrinsic dynamics of the subunits or is acquired in the hexameric state. Mode 6 is an expansion/contraction of the ring, accompanied by cooperative rotational and twisting motions of each monomer (see Video S5). The axis of rotation goes through each AF interface (Figure 5A) and performs an almost rigid rotation of the *Ec*NAGK-like dimers (Figure 5C). Residues close to these axes of rotation form minima in the mode fluctuations profile (Figure 5B) and belong to the AF interface. The axis involves a part of the *N*-terminal helix (6–20) of chains A and F, where the two helices interact tightly. Indeed, this interface stabilizes the hexameric arrangement and no NAGK dimer has been structurally characterized with an AF-like interface. The AF interface is unique to the hexameric arrangement.

As shown in Figure 4, the hexamer dynamics is affected by the intrinsic dynamics of the component subunits. Therefore, mode 6 could be associated with particular global modes accessible to the AB and/or AF dimers. We have examined the inter-residue distance variations maps induced by the low-frequency modes of the isolated AB and AF dimers to explore this possibility. AF dimer proves to be the major source of the rigid body movements of monomers observed in the hexamer (see Videos S6 and S7). The distance variation maps of the 1st and 4th modes of the AF dimer (Figure S1) illustrate that the internal motions within a given subunit are negligible, but the relative movements between the two subunits are significant. The AF interface, thus, emerges as a key mechanical region that confers to the two linked subunits suitable flexibility to undergo functional changes in their relative orientations. This dynamic feature of the AF interface, whose size is smaller than the AB interface, is in accord with Hubbard and co-workers [57], who stated that those interfaces that are not optimally packed may confer functional mobility to the oligomer. This inherent dynamical ability of the AF interface is therefore exploited in the hexameric arrangement to couple the rigid-body movements of the subunits, complementing their intrinsic internal dynamics.

Communication across the structure

The topology of the AF interface appears to be evolutionary selected to provide two essential features for the functionality of the enzyme: (1) flexibility to allow for the cooperative reorientations of the dimers, which is inextricably linked to allostery, and (2) thermodynamic stability of the whole hexamer. Taking into account the crucial role of the AF interface and with the aim of providing further insights into the allosteric regulation of this enzyme, we considered the maximum likelihood pathway (MLP)

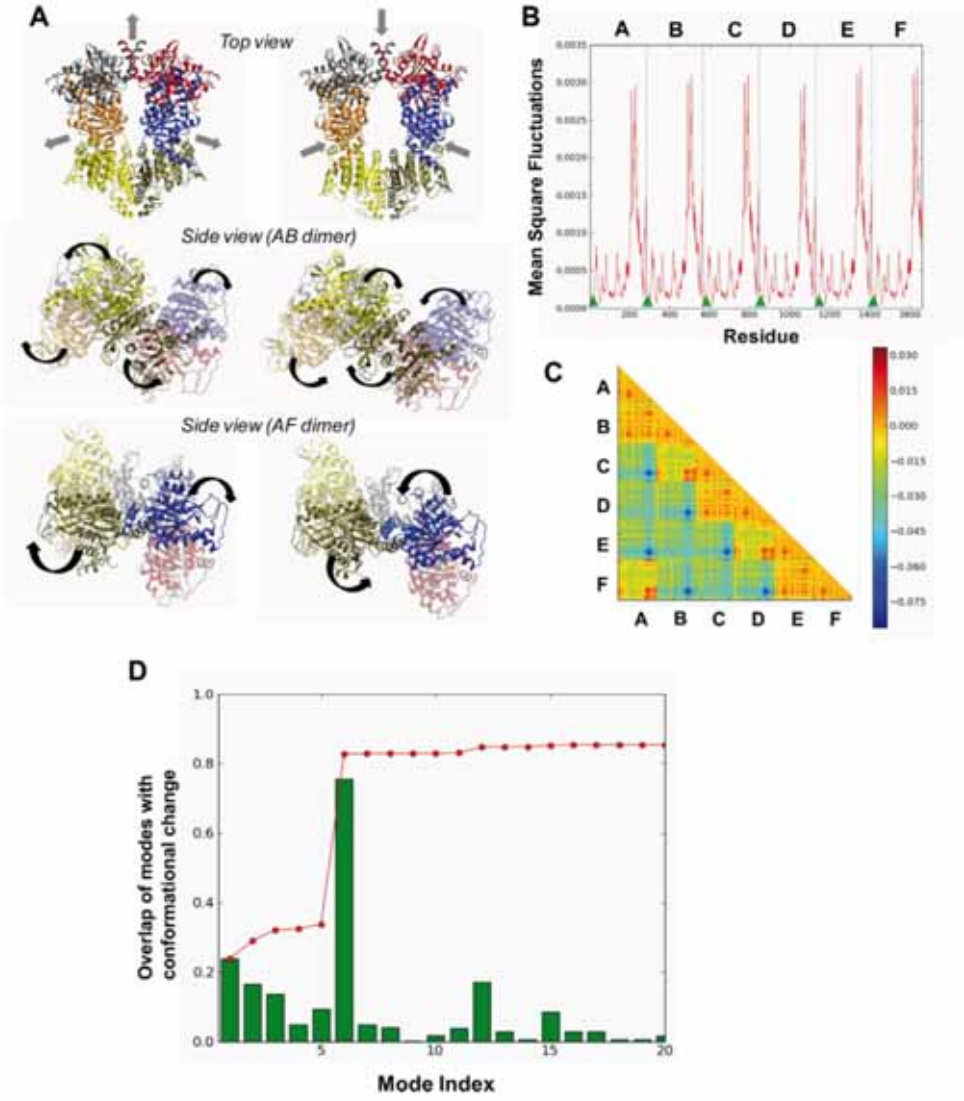


Figure 5. The cooperative mode of motion that enables the T→R transition of hexameric *Tm*NAGK. (A) Schematic description of the *Tm*NAGK mode 6 which yields a remarkably high overlap with the structural change involved in the T→R transition of the enzyme. Two structures have been generated using Eq.3. The top view of the deformed structures shows the opening/closure of the ring. The arrows show the direction of motion. The side views of the AB and AF dimers show a rotational movement of both dimers that make the hexameric ring flatter when it opens. See Video S5 for better visualization. (B) Mean-square displacements of residues in the 6th mode. Hinge sites are indicated by solid triangles. (C) Inter-residue distance variation map in mode 6. Blue/red/orange entries refer to distances that decrease/increase/remain unchanged. If the inter-residue distances within a given subunit remain constant, this indicates a rigid-body motion of the subunit (see Eq. 5). (D) Overlap of individual *Tm*NAGK modes with the allosteric change in structural coordinates between the T and R states. *Red line*: cumulative overlaps CO(*m*) between ANM modes and the experimentally observed conformational transition between R and T states (see Eq. 4), calculated for the 20 lowest-frequency modes. *Green bars*: overlap of each mode. This subset of 20 modes accounts for 85% of the conformational change, predominantly contributed by the 6th mode (overlap of 75%).

doi:10.1371/journal.pcbi.1002201.g005

for each combination of pairs of residues (endpoints) belonging to the respective chains A and F, and evaluated the fractional occurrence of each residue in the ensemble of MLPs (see Methods). Figure 6A displays the percent occurrence of each residue, which also provides a measure of the relative allosteric potential of the residues. Peaks are observed at K17, E18, F19, Y20, K50 and Y51 (ribbon diagram color-coded from blue (peaks) to red (minima) in Figure 6B). The significance of this first set in allosteric communication could be anticipated due to their location at the tightest part of the AF interface and proximity to the arginine inhibitor (Figure 6B). However, our approach helps to identify other distal residues important for the communication, which behave as hubs. In particular, K196 and I162 channel most of the pathways to the AF interface via interactions with F19 (and the arginine inhibitor) and K50, respectively.

The communication across the AF interface can be summarized namely by two symmetric pathways distinguished by the MLP analysis: $I162_A \rightarrow K50_A \rightarrow Y51_A \rightarrow K17_F \rightarrow E18_F \rightarrow F19_F \rightarrow K196_F$ and its counterpart $I162_F \rightarrow \ldots \rightarrow K196_A$ (colored yellow and green in Figure 6B). Aromatic residues tend to be favored at protein interfaces [25], and in this case, F19 and Y20 play a critical role. Not surprisingly, F19 is highly conserved among arginine-sensitive NAGKs [30] and, together with Y20 (violet in Figure 6B), it establishes an efficient communication pathway of the form $F19_{(A/F)} \rightarrow Y20_{(A/F)} \rightarrow Y20_{(F/A)} \rightarrow F19_{(F/A)}$.

Differences in the dimer organization point to different functional mechanisms: *Ec*NAGK vs *Ec*UMPK

The structure of the monomeric subunit of *Ec*NAGK is preserved among all family members, but the assembly geometry is less conserved. The arrangement of the monomeric subunits of NAGKs and CKs is strikingly similar, as shown above, but has significant differences with the assembly of UMP Kinases. Structurally, UMPKs are trimers of dimers in which the two helices that build the intersubunit surface of each dimer are parallel (Figure 7C and D), whereas in NAGK (and CK) these helices at the interface make an angle of ~65° (Figure 7A and B). To our knowledge, a clear functional reason for this difference in monomer-monomer packing has not been reported so far. Although this difference has been argued to be necessary for hexameric assembly [58], there might be another functional reason since *Tm*NAGK is an example of a hexameric assembly that selectively adapts the *Ec*NAGK-like dimer packing (AB dimer). Here we compute the ANM modes of the UPMK dimer from *Escherichia Coli* (*Ec*UMPK) in order to examine whether such a difference in packing geometry gives rise to significant changes in the global dynamics.

The first mode of motion of the isolated *Ec*UMPK dimer entails a rotational rigid-body movement with respect to an axis across the αC helices (Figure 7, panels C and D, and Video S8). The anticorrelated motion of both subunits leads to an opening/closure movement of the whole dimer. This is in sharp contrast to the *Ec*NAGK dimer dynamics, whose low-frequency modes do not exhibit rigid-body movements of the subunits. Does this dynamic feature of the *Ec*UMPK dimer play a functional role?

Gilles and co-workers determined the X-ray crystal structure of *Ec*UMPK complexed with GTP (PDB code 2VRY) [59], which is an allosteric activator, and characterized a functional conformational change. They argued that GTP induces a rearrangement of the quaternary structure that involves a rigid-body rotation of 11° that opens the UMPK dimer. Strikingly, the first ANM mode predicted for the UDP-bound dimer describes the structural transition between the UDP- and GTP-bound forms. The overlap is outstandingly high (0.78) (see Figure 8E for cumulative overlap). Moreover, it is worth pointing out that we have checked that this mode of motion is totally conserved in the hexamer (see Figure S2).

Why does the different assembly in the UMPK dimer give rise to a normal mode with a rigid-body character not present in *Ec*NAGK? In UMPK the interface between the monomers is constituted mainly by two long parallel helices (αC) able to build a rotational axis that promotes an *en bloc* motion of both subunits. In contrast, the crossed orientation of the helices of NAGK (~65°)

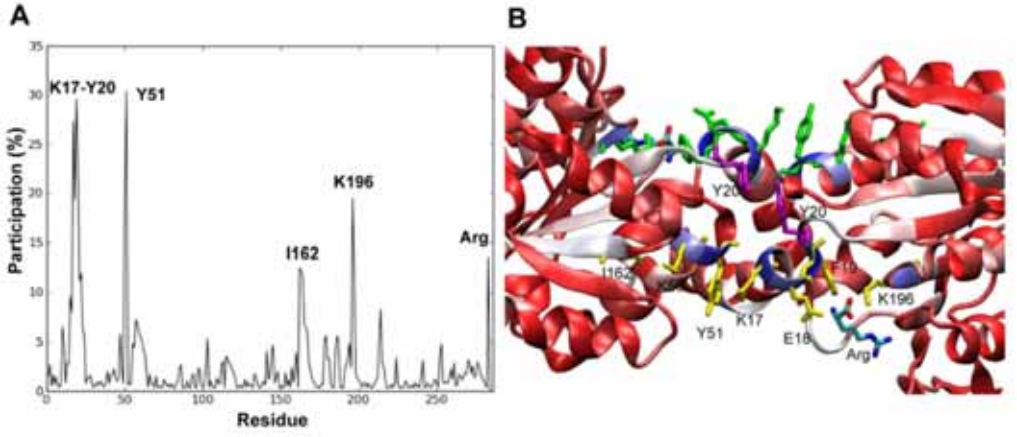


Figure 6. Communication pathways at the AF interface. (A) Percentage of communication pathways in which a given residue is on-pathway. (B) Color coded-ribbon diagram of the AF interface. The color code refers to the participation of the residues in the located communication pathways (the participation increases from red to blue).The main communication pathways across the interface are colored in green, yellow and violet, and the residues on-pathway are labeled.
doi:10.1371/journal.pcbi.1002201.g006

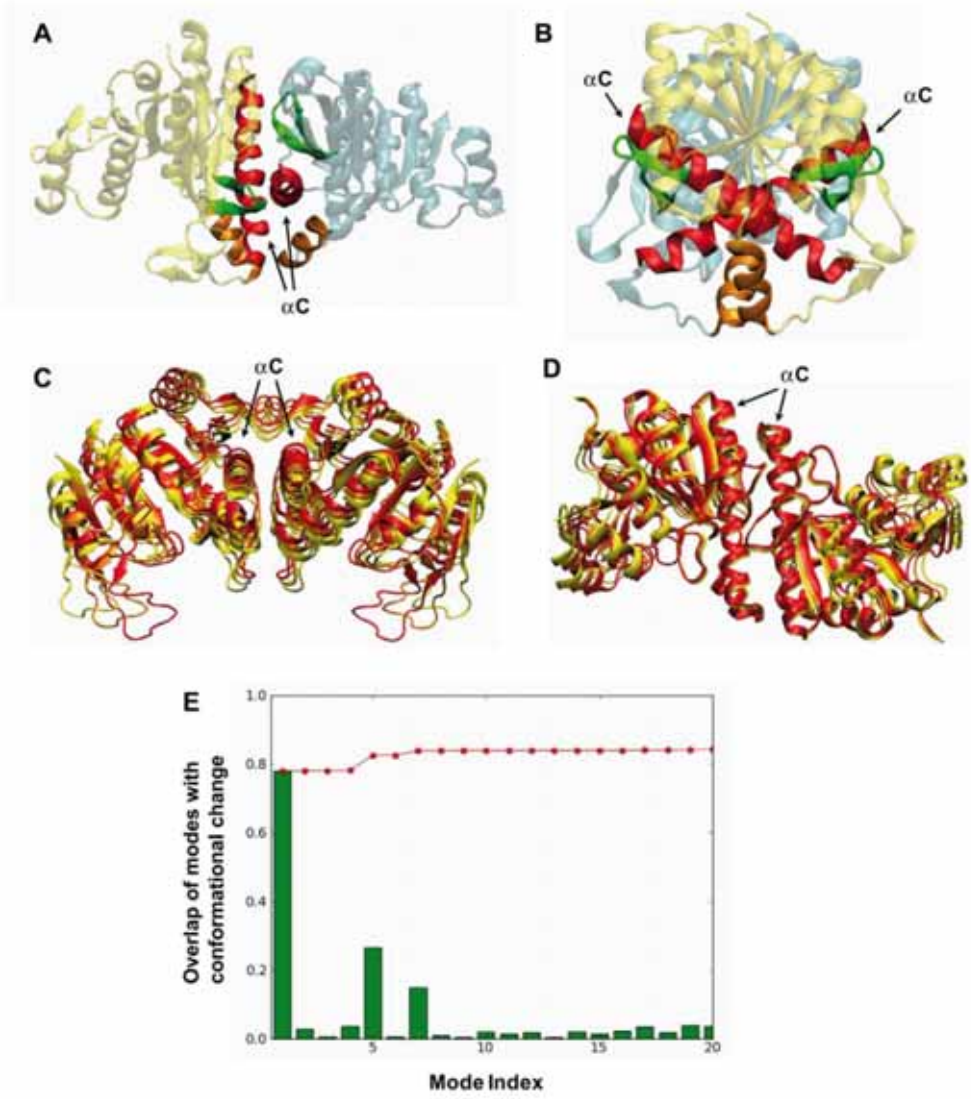


Figure 7. Comparison between *Ec*NAGK and *Ec*UMPK dimmers. (A) and (B) Ribbon representations of two perpendicular views of the *Ec*NAGK dimer. A different color is used for each subunit. Secondary structure elements building the intersubunit surface are colored differently (helices αC in red, β9-β10 hairpins in green and helices αB in orange). (C) and (D) Ribbon representations of two perpendicular views of the *Ec*UMPK dimer. Different conformations along the 1st ANM mode are generated with Eq. 3 (s = -20 in red , s = 0 in orange and s = 20 in yellow). See Video S8 for better visualization. (E) Comparison of *Ec*UMPK dimer modes with the allosteric conformational change observed in the GTP-bound form. *Red line*: cumulative overlaps CO(*m*) between ANM modes and the experimentally observed conformational transition between the UDP- and GTP-bound states (Eq. 4 in *Methods*), calculated for the 20 lowest-frequency modes. *Green bars*: overlap of each mode with the conformational change. This subset of 20 modes accounts for 84% of the conformational change, being predominantly contributed by the 1st mode.
doi:10.1371/journal.pcbi.1002201.g007

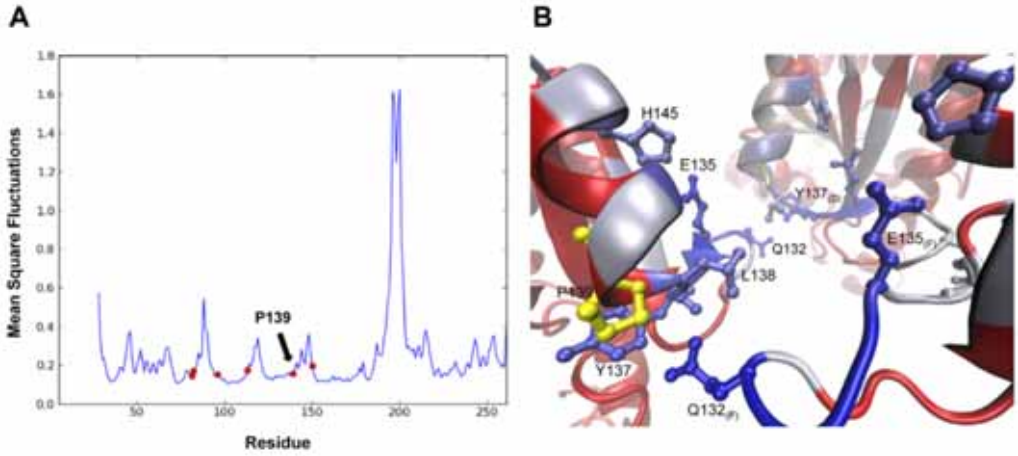


Figure 8. Collective dynamics and signal propagation in *Mt*UMPK. (A) Mobility profile obtained by ANM computed for all 3*N*-6 normal modes of the hexamer. The fluctuations of only one monomer are shown. Red dots correspond to those positions that were mutated in the study of Labesse and co-workers [60] (B) Color coded ribbon diagram of the interface. The color code (same as in Figure 6B) refers to the participation of the residues in the located communication pathways. P139 (shown in yellow) does not directly participate in intersubunit communication but highly constraints the neighboring residues Y137 and L138 that play a key role in allosteric signaling.
doi:10.1371/journal.pcbi.1002201.g008

and the presence of other intersubunit contacts (B-helices and β9–β10 hairpins) hinders a rigid-body rotation of the two subunits. This suggests that the unique dimeric assembly of UMPK gives rise to a particular soft mode not present in other AAK family members. This example further indicates that the design of the interfacial contact topology and oligomerization geometry is crucial in defining the functional mechanisms of oligomers.

Importance of spatial constraints in the allosteric regulation of UMPK

In some cases, a single residue may significantly affect the contact topology at the interface and, thus, the allosteric regulation. This has been explored in the context of the UMPK analogue from *Mycobacterium tuberculosis* (*Mt*UMPK), for which crystallographic and site-directed mutagenesis studies have been recently conducted [60]. The X-ray structure of *Mt*UMPK bound to GTP shows striking similarities to *Ec*UMPK structure. Notably, this similarity is extended to their global motions: the lowest frequency ANM modes of the two structures exhibit an overlap of 0.97. Given that the global modes of motion are fully determined by the overall shape of the protein, local perturbations are indeed unlikely to affect the low-frequency modes.

Site-directed mutagenesis studies, on the other hand, show the importance of some residues in both the activity and the cooperativity of the enzyme. Among them, P139 was pointed out to to be a key residue in the allosteric regulation of the enzyme. P139 is located close to the trimeric interface where three GTP molecules are bound. What is the dynamical role of this residue? The mean-square fluctuations profile obtained with the ANM shows that P139 occupies a position close to a local minimum (a rigid part of the protein) (Figure 8A). Such regions usually play a key mechanical role for mediating collective changes in structure, and mutations at such positions may potentially affect the allosteric dynamics of the protein.

We have analyzed the importance of P139 in mediating the allosteric communication among subunits A, D and F, which build one of the two trimeric interfaces where three GTP molecules are bound. We computed the communication pathways between GTP binding residues (starting from subunit A and ending at subunits D and F) and the percent contribution of each residue to MLPs, as done for *Tm*NAGK. Figure 8 shows the trimeric interface color-coded according to the percent contribution in the same way as in Figure 6B. We note that the participation of P139 (in yellow) to these pathways is minimal (note the red color in the backbone), but the adjacent residues Y137 and L138 are important mediators of inter-subunit communication via interactions with Q132.

This analysis suggests that the importance of P139 lies in constraining the orientation of nearby residues Y137 and L138 involved in inter-subunit signal propagation. The fact that this residue is highly restricted position in the global mode profile emphasizes its role in constraining the neighboring residues in a precise orientation pre-disposed to enable inter-subunit communication. The experimentally tested mutants (P139A, P139W and P139H) all showed a diminished allosteric regulation, but to different extents [60]. Further simulations at atomic scale might help explain the relative sizes of the effects induced by these mutations, but this is beyond the scope of the present work. It might be interesting to experimentally test the effect of mutations at L18, Y137 and Q132, since these residues emerge here as key elements enabling inter-subunit communication and they are distinctly restricted in the collective dynamics (Figure 8A) despite the relatively low packing density at the interface.

To summarize, the present study reveals several dynamic features of oligomeric proteins by means of an ENM analysis of family members with different degrees of oligomerization. A common dynamic feature of the oligomers presented here is the conservation of the inherent dynamics of their monomeric or dimeric building blocks. The way these blocks are assembled in different oligomers confers different types of collective mechanisms

unique to particular oligomerization geometries. Here are the main observations:

(1) The dimeric *Ec*NAGK and *Pf*CK conserve to a high extent those normal modes of the monomers which involve minimal conformational rearrangements at the intersubunit interface.

(2) The topology of the interface in *Pf*CK provides the protruding subdomains of the component subunits with remarkably high mobility, which apparently enhances the affinity for binding the carbamate substrate and for excluding other carbamate analogues that are more abundant, as suggested by recent experiments [37].

(3) The *Tm*NAGK hexamer has two different types of interfaces (AB and AF) that provide different dynamic properties to the hexamer. The AF interface provides the hexamer with the ability to perform *en bloc* motions that cooperatively engage all six subunits, in contrast to the AB interface that enjoys an internal flexibility relevant to the opening of the substrate binding site. The concerted movements of the six subunits coupled to the internal motion of the subunits give rise to a normal mode (mode 6, Figure 5) intimately linked to the allosteric transition of the hexameric enzyme.

(4) The ability of *Tm*NAGK to enable allosteric signaling has been studied by means of a Markov model of network communication. The MLPs connecting residues of chains A and F suggest that some residues of the interlaced N-terminal helices, which build the AF interface (e.g., K17, E18, F19 and Y20) are distinguished by their high allosteric potential. Notably, these residues coincide with the key mechanical sites (global hinges) that mediate the cooperative mode of motion.

(5) The different assembly of the subunits in the *Ec*UMPK dimer, with respect to *Ec*NAGK, gives rise to rigid-body movements of the subunits that are necessary for the allosteric regulation of *Ec*UMPK. The mutual disposition of the two long helices that build the interface in either enzyme proves to be crucial for favoring functional dynamics. Interestingly, the experimentally observed allosteric switch mechanism of UMPK is closely reproduced by a single mode (ANM mode 1; Figure 7E), in support of the functional significance of the collective motions uniquely defined by the dimeric architecture.

(6) In parallel with the observations made for *Tm*NAGK allosteric communication, a series of residues highly restricted in the collective dynamics of *Mt*UMPK play a key role in enabling intersubunit communication. P139 plays a structural role by introducing backbone constraints that precisely constrain nearby residues' side chains in orientations predisposed to optimal binding of GTP and inter-subunit communication. The significance of P139 in enabling allosteric communication is consistent with site-directed mutagenesis data [60].

In summary, the oligomers in the examined AAK family appear to selectively exploit the inherent dynamic abilities of its components, on the one hand, and favor coupled movements of intact subunits, on the other, to effectively sample cooperative movements (soft modes) that enable motions required for substrate binding and efficient allosteric responses. The architecture of the interfaces and the assembly geometry play an essential role in defining the most easily accessible (or softest) modes of motion, which in turn, are shown to be relevant to the functional mechanisms of the different oligomers, being presumably optimized by evolutionary pressure.

Methods

Anisotropic Network Model (ANM)

The low-frequency modes described by the NMA of different ENM variants [40,61–64] have proven to be robustly determined by the overall fold [7,65,66] and provide a consistent description of the conformational space most easily accessible to the protein [67]. Among them, we use here the most broadly used model, the anisotropic network model (ANM) [40,41]. In the ANM, the network nodes are located at the C^{α}-atoms' positions, and pairs of nodes within close proximity (a cutoff distance of 15 Å, including bonded or non-bonded pairs of amino acids [41]) are connected by springs of uniform force constant γ. The interaction potential of the molecule is given by

$$V_{ANM} = \frac{\gamma}{2}\sum_{i,j}^{M}\left(\left|\boldsymbol{R}_{ij}\right| - \left|\boldsymbol{R}_{ij}^{0}\right|\right)^{2} \quad (1)$$

where M is the number of springs, and $|\boldsymbol{R}_{ij}|$-$|\boldsymbol{R}_{ij}^{0}|$ is the inter-residue distance with respect to the equilibrium (crystal) structure. The second derivatives of V_{ANM} with respect to residue displacements yield the 3Nx3N Hessian matrix $\mathbf{H}$, the eigenvalue decomposition of which yields 3N-6 nonzero eigenvalues λ_k and eigenvectors $\boldsymbol{u}_k$ corresponding to the frequencies (squared) and shapes of the normal modes of motion accessible to the examined structure. Numbering of modes in this work starts from the first mode with a nonzero eigenvalue.

The cross-correlation between the displacements of residues i and j, contributed by mode k scales as

$$(\Delta\boldsymbol{R}_i \cdot \Delta\boldsymbol{R}_j)_k \propto \left(\boldsymbol{u}_k\boldsymbol{u}_k^{T}\right)_{ij}/\lambda_k \quad (2)$$

where the subscript ij designates the element of the matrix in square brackets. For $i = j$, equation (2) reduces to the square displacement of residue i in mode k. Clearly, lower-frequency modes (smaller λ_k) drive larger-amplitude motions.

Generation of large-amplitude conformational changes

Conformations sampled upon moving along mode k are generated using

$$[\boldsymbol{R}(\pm s)]_k = \boldsymbol{R}^{0} \pm s\lambda_k^{-1/2}\boldsymbol{u}_k \quad (3)$$

where $\boldsymbol{R}^{0}$ is the $3N$-dimensional vector representing the initial coordinates of all residues and s is a parameter that rescales the amplitude of the deformation induced by mode k. The movies S1-S8 in the Supporting Information are generated using this equation with a series of different s values for selected modes of examined proteins.

Comparison of experimental conformational changes with normal modes

The degree of overlap between a conformational change $\Delta\boldsymbol{r}$ observed by X-ray crystallography and the structural change predicted by the ANM to take place along mode k is quantified by $(\Delta\boldsymbol{r} \cdot \boldsymbol{u}_k)/|\Delta\boldsymbol{r}|$. Here $\Delta\boldsymbol{r}$ is the $3N$-dimensional difference vector between the α-carbon coordinates of two different forms resolved for the same protein under different conditions (e.g., substrate-

bound and -unbound forms of enzymes, or inward-facing or outward-facing forms of transporters). The cumulative overlap CO(m) between $\Delta \boldsymbol{r}$ and the directions spanned by a subset of m modes is calculated as

$$\mathrm{CO}(m)=\sqrt{\sum_{k=1}^{m}\left((\Delta \boldsymbol{r}\cdot \boldsymbol{u}_k)/|\Delta \boldsymbol{r}|\right)^2} \quad (4)$$

CO(m) sums up to unity for $m = 3N\text{-}6$, as the eigenvectors form a complete orthonormal set of basis vectors in the $3N$-6 dimensional space of internal conformational changes (see Figures 5D and 7E)

Subspace overlap

The similarity between the conformational spaces described by two subsets of *m and n* modes, $\boldsymbol{u}_k$ and $\boldsymbol{v}_l$, evaluated for two different systems can be quantified in terms of a double summation over squared overlaps as in Eq. 4, among all mxn pairs of modes (divided by m or n, depending on the reference set). The overlap $O(\boldsymbol{u}_k,\boldsymbol{v}_l)$ between the pairs of modes $\boldsymbol{u}_k$ and $\boldsymbol{v}_l$ calculated for different systems (e.g., Figure 3) is given by the inner product of the eigenvectors, i.e.,

$$O(\boldsymbol{u}_k,\boldsymbol{v}_l,)=\boldsymbol{u}_k\cdot\boldsymbol{v}_l \quad (5)$$

Note that $O(\boldsymbol{u}_k,\boldsymbol{v}_l)$ is equal to the correlation cosine between the two N-dimensional vectors, since the eigenvectors are normalized.

Distance variation maps

The change in a given inter-residue distance $|\boldsymbol{R}^0_{ij}|$ induced by a given mode k, $(\Delta R_{ij})_k$, is given by the projection of the deformation induced by the k^{th} mode onto the normalized distance vector, scaled by the inverse frequency,

$$(\Delta R_{ij})_k = s\,\lambda_k^{-1/2}\left[(\boldsymbol{u}_k)_j-(\boldsymbol{u}_k)_i\right]\cdot\frac{\boldsymbol{R}^0_{ij}}{\left|\boldsymbol{R}^0_{ij}\right|} \quad (6)$$

Here $(\boldsymbol{u}_k)_i$ designates the i^{th} super element (a 3D vector) of $\boldsymbol{u}_k$, and describes the relative displacement of the i^{th} residue (x-, y-, and z-components) along the k^{th} mode direction.

Communication pathways

Inter-residue communication has been suggested to play a key role in allosteric regulation and enzymatic catalysis [68,69], and has been the subject of many computational studies [48,70–72]. Here we use a Markov model of network communication [73,74] to identify communication pathways. The interactions between residue pairs connected in the ANM are defined by the affinity matrix **A**, whose elements are $a_{ij}=N_{ij}/(N_i\,N_j)^{1/2}$ where N_{ij} is the number of atom-atom contacts between residues i and j based on a cutoff distance of 4 Å, and N_i is the number of heavy atoms belonging to residue i. The density of contacts at each node i is given by $d_i=\sum_{j=1}^{N} a_{ij}$.The Markov transition matrix $\boldsymbol{M}=\{m_{ij}\}$, where $m_{ij}=a_{ij}/d_j$ determines the conditional probability of transmitting a signal from residue j to residue i in one time step [73]. We define $-\log(m_{ij})$ as the corresponding 'distance'. The maximum-likelihood paths (MLPs) for signal transfer between two end points are evaluated using the Dijkstra's algorithm [73]. In order to identify the residues that play a key role in establishing the communication between pairs of subunits, we considered the communication between all pairs of residues belonging to the two subunits of interest. In the application to the communication between the A and F subunits of *Tm*NAGK (Figure 6), an ensemble of $\mathcal{N}^2=282^2$ combinations of residue pairs (endpoints) have thus been considered (each chain consists of $\mathcal{N}=282$ residues). For each pair, we evaluated the MLP and thus determined the series of residues taking part in the MLP. To quantify the contribution of a given residue to intersubunit communication, we counted the occurrence of each residue in the complete ensemble of MLPs. Figure 6, panel A displays the resulting curve, peaks indicating the residues that make the largest contribution.

NMA of a subsystem coupled to a dynamic environment

In many applications the dynamics of a part of the protein (subsystem, S) may be of interest in the context of its environment (E). The Hessian of the whole system is conveniently partitioned into four submatrices [75,76]:

$$\mathbf{H}=\begin{pmatrix}\mathbf{H}_{SS} & \mathbf{H}_{SE}\\ \mathbf{H}_{ES} & \mathbf{H}_{EE}\end{pmatrix} \quad (7)$$

where $\mathbf{H}_{SS}$ is the Hessian submatrix for the subsystem, $\mathbf{H}_{EE}$ is that of the environment and $\mathbf{H}_{SE}$ (or $\mathbf{H}_{ES}$) refers to the coupling between the subsystem and the environment. Inasmuch as the environment responds to the subsystem structural changes by minimizing the total energy, the effective Hessian for the subsystem $\mathbf{H}^{eff}_{SS}$ coupled to the environment is

$$\mathbf{H}^{eff}_{SS}=\mathbf{H}_{SS}-\mathbf{H}_{SE}\mathbf{H}^{-1}_{EE}\mathbf{H}_{ES} \quad (8)$$

This approach has been advantageously employed in determining potential allosteric sites [77] and locating transition states of chemical reactions [78]. It will be used below in conjunction with the ANM for assessing the effect of oligomerization on the dynamics of monomeric and/or dimeric components (subsystem).

Structural data

We examined four enzymes belonging to the AAK family (Figure 1): *Ec*NAGK (dimer), *Tm*NAGK (hexamer), *Pf*CK (dimer) and *Ec*UMPK (hexamer). To this aim, we use the X-ray structures of *Ec*NAGK in the open state (PDB code: 2WXB), the arginine-bound *Tm*NAGK (PDB code: 2BTY), the ADP-bound *Pf*CK (PDB code: 1E19) and the UDP-bound *Ec*UMPK (PDB code: 2BND).

All diagrams of molecular structures have been generated using VMD [79].

Supporting Information

Figure S1 Distance variation maps of the 1st and 4th modes of the AF dimer. Blue positions indicate that the distance between two residues decreases, and a red position that it increases. If the inter-residue distances within a given subunit remain constant, this indicates a rigid-body motion of the subunit. See Videos S6 and S7 for better visualization of these two normal modes.
(TIF)

Figure S2 Comparison of the global dynamics of the dimeric component of *Ec*UMPK in the hexamer with that of the isolated dimeric component. Overlaps between

the 20 slowest modes of the dimer and hexamer are labelled in the heat map. The AB dimer is highlighted in the ribbon diagram of *Ec*UMPK and the rest of the hexamer (the environment) is shadowed. The structure is colored by chains. The first mode of the dimer is expressed by two modes within the hexamer (the overlap with hexameric modes 1 and 3 is 0.73 and 0.58, respectively). The dynamic properties of the dimer are remarkably well conserved in the hexamer as given by a subspace overlap of 0.95 of the 20 lowest-frequency modes.
(TIF)

Table S1 Overlap between the eight lowest frequency modes of the isolated *Ec*NAGK monomer and the *Ec*NAGK monomer within the dimmer.
(DOC)

Table S2 Overlap between the eight lowest frequency modes of the isolated *Pf*CK monomer and the *Pf*CK monomer within the dimmer.
(DOC)

Video S1 Symmetric substrate binding mode of motion of dimeric *Ec*NAGK (ANM mode 5).
(WMV)

Video S2 Asymmetric substrate binding mode of motion of dimeric *Ec*NAGK (ANM mode 4).
(WMV)

Video S3 Symmetric substrate binding mode of motion of dimeric *Pf*CK (ANM mode 3).
(WMV)

Video S4 Asymmetric substrate binding mode of motion of dimeric *Pf*CK (ANM mode 4).
(WMV)

Video S5 Allosteric mode of motion of hexameric *Tm*NAGK (ANM mode 6).
(WMV)

Video S6 Mode of motion of the isolated AF-type dimer of *Tm*NAGK (ANM mode 1).
(WMV)

Video S7 Mode of motion of the isolated AF-type dimer of *Tm*NAGK (ANM mode 4).
(WMV)

Video S8 Allosteric mode of motion of the dimeric component of *Ec*UMPK (ANM mode 1).
(WMV)

Acknowledgments

The authors thank V. Rubio, F. Gil-Ortiz and S. Ramon-Maiqués from the IBV-CSIC for useful crystallographic information. EM acknowledges fruitful discussions with members of Ivet Bahar's lab and thanks Chakra Chennubhotla for his assistance with software implementation.

Author Contributions

Conceived and designed the experiments: EM RC IB. Performed the experiments: EM. Analyzed the data: EM RC IB. Contributed reagents/materials/analysis tools: EM RC IB. Wrote the paper: EM RC IB.

References

1. Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. Nature 450: 964–972.
2. Goodman JL, Pagel MD, Stone MJ (2000) Relationships between protein structure and dynamics from a database of NMR-derived backbone order parameters. J Mol Biol 295: 963–978.
3. Zhang FL, Brüschweiler R (2002) Contact model for the prediction of NMR N-H order parameters in globular proteins. J Am Chem Soc 124: 12654–12655.
4. Tobi D, Bahar I (2005) Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. Proc Natl Acad Sci U S A 102: 18908–18913.
5. Bakan A, Bahar I (2009) The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. Proc Natl Acad Sci U S A 106: 14349–14354.
6. Bahar I, Lezon TR, Yang LW, Eyal E (2010) Global Dynamics of Proteins: Bridging Between Structure and Function. Annu Rev Biophys 39: 23–42.
7. Tama F, Brooks CL (2006) Symmetry, form, and shape: Guiding principles for robustness in macromolecular machines. Annu Rev Biophys Biomol Struct 35: 115–133.
8. Gorfe AA, Lu BZ, Yu ZY, McCammon JA (2009) Enzymatic Activity versus Structural Dynamics: The Case of Acetylcholinesterase Tetramer. Biophys J 97: 897–905.
9. Halle B (2002) Flexibility and packing in proteins. Proc Natl Acad Sci U S A 99: 1274–1279.
10. Tama F, Sanejouand YH (2001) Conformational change of proteins arising from normal mode calculations. Protein Eng 14: 1–6.
11. Wolf-Watz M, Thai V, Henzler-Wildman K, Hadjipavlou G, Eisenmesser EZ, et al. (2004) Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. Nat Struct Mol Biol 11: 945–949.
12. Eisenmesser EZ, Bosco DA, Akke M, Kern D (2002) Enzyme dynamics during catalysis. Science 295: 1520–1523.
13. Jimenez A, Clapes P, Crehuet R (2009) Protein Flexibility and Metal Coordination Changes in DHAP-Dependent Aldolases. Chemistry-a European Journal 15: 1422–1428.
14. Grueninger D, Schulz GE (2008) Antenna domain mobility and enzymatic reaction of L-rhamnulose-1-phosphate aldolase. Biochemistry 47: 607–614.
15. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, et al. (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature 450: 913–U927.
16. Perutz MF (1989) Mechanisms of cooperativty and allosteric regulation in proteins. Q Rev Biophys 22: 139–236.
17. Eaton WA, Henry ER, Hofrichter J (1991) Application of linear free-energy relations to protein conformational changes: the quaternary structural change of hemoglobin. Proc Natl Acad Sci U S A 88: 4472–4475.
18. Xu CY, Tobi D, Bahar I (2003) Allosteric changes in protein structure computed by a simple mechanical model: Hemoglobin T <-> R2 transition. J Mol Biol 333: 153–168.
19. Changeux JP, Edelstein SJ (2005) Allosteric mechanisms of signal transduction. Science 308: 1424–1428.
20. Mouawad L, Perahia D (1996) Motions in hemoglobin studied by normal mode analysis and energy minimization: Evidence for the existence of tertiary T-like, quaternary R-like intermediate structures. J Mol Biol 258: 393–410.
21. Koshland DE, Nemethy G, Filmer D (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits. Biochemistry 5: 365–385.
22. Monod J, Wyman J, Changeux JP (1965) On nature of allosteric transitions - A plausible model. J Mol Biol 12: 88–118.
23. Sinko W, Oliveira C, Williams S, Van Wynsberghe AW, Durrant JD, et al. (2011) Applying molecular dynamics simulations to identify rarely sampled ligand-bound conformational states of undecaprenyl pyrophosphate synthase, an antibacterial target. Chem Biol Drug Des 77: 412–420.
24. Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc Natl Acad Sci U S A 93: 13–20.
25. Stites WE (1997) Protein-protein interactions: Interface structure, binding thermodynamics, and mutational analysis. Chem Rev 97: 1233–1250.
26. Marina A, Alzari PM, Bravo J, Uriarte M, Barcelona B, et al. (1999) Carbamate kinase: New structural machinery for making carbamoyl phosphate, the common precursor of pyrimidines and arginine. Protein Sci 8: 934–940.
27. Ramon-Maiques S, Marina A, Uriarte M, Fita I, Rubio V (2000) The 1.5 angstrom resolution crystal structure of the carbamate kinase-like carbamoyl phosphate synthetase from the hyperthermophilic archaeon Pyrococcus furiosus, bound to ADP, confirms that this thermostable enzyme is a carbamate kinase, and provides insight into substrate binding and stability in carbamate kinases. J Mol Biol 299: 463–476.
28. Ramon-Maiques S, Marina A, Gil-Ortiz F, Fita I, Rubio V (2002) Structure of acetylglutamate kinase, a key enzyme for arginine biosynthesis and a prototype for the amino acid kinase enzyme family, during catalysis. Structure 10: 329–342.
29. Marco-Marin C, Gil-Ortiz F, Rubio V (2005) The crystal structure of Pyrococcus furiosus UMP kinase provides insight into catalysis and regulation in microbial pyrimidine nucleotide biosynthesis. J Mol Biol 352: 438–454.
30. Ramon-Maiques S, Fernandez-Murga ML, Gil-Ortiz F, Vagin A, Fita I, et al. (2006) Structural bases of feed-back control of arginine biosynthesis, revealed by

the structures of two hexameric N-acetylglutamate kinases, from Thermotoga maritima and Pseudomonas aeruginosa. J Mol Biol 356: 695–713.
31. Marco-Marin C, Gil-Ortiz F, Perez-Arellano I, Cervera J, Fita I, et al. (2007) A novel two-domain architecture within the amino acid kinase enzyme family revealed by the crystal structure of Escherichia coli glutamate 5-kinase. J Mol Biol 367: 1431–1446.
32. Gil-Ortiz F, Ramon-Maiques S, Fernandez-Murga ML, Fita I, Rubio V (2010) Two Crystal Structures of Escherichia coli N-Acetyl-L-Glutamate Kinase Demonstrate the Cycling between Open and Closed Conformations. J Mol Biol 399: 476–490.
33. Gil-Ortiz F, Ramon-Maiques S, Fita I, Rubio V (2003) The course of phosphorus in the reaction of N-acetyl-L-glutamate kinase, determined from the structures of crystalline complexes, including a complex with an AlF4- transition state mimic. J Mol Biol 331: 231–244.
34. Marcos E, Crehuet R, Bahar I (2010) On the Conservation of the Slow Conformational Dynamics within the Amino Acid Kinase Family: NAGK the Paradigm. PLoS Comput Biol 6: e1000738.
35. Llacer JL, Contreras A, Forchhammer K, Marco-Marin C, Gil-Ortiz F, et al. (2007) The crystal structure of the complex of P-II and acetylglutamate kinase reveals how P-II controls the storage of nitrogen as arginine. Proc Natl Acad Sci U S A 104: 17644–17649.
36. Fernandez-Murga ML, Rubio V (2008) Basis of arginine sensitivity of microbial N-acetyl-L-glutamate kinases: Mutagenesis and protein engineering study with the Pseudomonas aeruginosa and Escherichia coli enzymes. J Bacteriol 190: 3018–3025.
37. Ramon-Maiques S, Marina A, Guinot A, Gil-Ortiz F, Uriarte M, et al. (2010) Substrate Binding and Catalysis in Carbamate Kinase Ascertained by Crystallographic and Site-Directed Mutagenesis Studies: Movements and Significance of a Unique Globular Subdomain of This Key Enzyme for Fermentative ATP Production in Bacteria. J Mol Biol 397: 1261–1275.
38. Ali MH, Imperiali B (2005) Protein oligomerization: How and why. Bioorg Med Chem 13: 5013–5020.
39. Grueninger D, Treiber N, Ziegler MOP, Koetter JWA, Schulze MS, et al. (2008) Designed protein-protein association. Science 319: 206–209.
40. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, et al. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80: 505–515.
41. Eyal E, Yang LW, Bahar I (2006) Anisotropic network model: systematic evaluation and a new web interface. Bioinformatics 22: 2619–2627.
42. Krebs WG, Alexandrov V, Wilson CA, Echols N, Yu HY, et al. (2002) Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. Proteins 48: 682–695.
43. Yang L, Song G, Carriquiry A, Jernigan RL (2008) Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. Structure 16: 321–330.
44. Yang LW, Eyal E, Bahar I, Kitao A (2009) Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics. Bioinformatics 25: 606–614.
45. Thomas A, Field MJ, Mouawad L, Perahia D (1996) Analysis of the low frequency normal modes of the T-state of aspartate transcarbamylase. J Mol Biol 257: 1070–1087.
46. Zheng WJ, Brooks BR, Thirumalai D (2009) Allosteric Transitions in Biological Nanomachines are Described by Robust Normal Modes of Elastic Networks. Current Protein & Peptide Science 10: 128–132.
47. Rueda M, Chacon P, Orozco M (2007) Thorough validation of protein normal mode analysis: A comparative study with essential dynamics. Structure 15: 565–575.
48. Liu L, Koharudin LMI, Gronenborn AM, Bahar I (2009) A comparative analysis of the equilibrium dynamics of a designed protein inferred from NMR, X-ray, and computations. Proteins 77: 927–939.
49. Romo TD, Grossfield A (2011) Validating and improving elastic network models with molecular dynamics simulations. Proteins 79: 23–34.
50. Niv MY, Filizola M (2008) Influence of oligomerization on the dynamics of G-protein coupled receptors as assessed by normal mode analysis. Proteins 71: 575–586.
51. Wang YM, Jernigan RL (2005) Comparison of tRNA motions in the free and ribosomal bound structures. Biophys J 89: 3399–3409.
52. Kantarci N, Doruker P, Haliloglu T (2006) Cooperative fluctuations point to the dimerization interface of p53 core domain. Biophys J 91: 421–432.
53. Ishida H, Jochi Y, Kidera A (1998) Dynamic structure of subtilisin-eglin c complex studied by normal mode analysis. Proteins 32: 324–333.
54. Yang Z, Majek P, Bahar I (2009) Allosteric Transitions of Supramolecular Systems Explored by Network Models: Application to Chaperonin GroEL. PLoS Comput Biol 5: e1000360.
55. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372: 774–797.
56. Holm L, Kaariainen S, Rosenstrom P, Schenkel A (2008) Searching protein structure databases with DaliLite v.3. Bioinformatics 24: 2780–2781.
57. Hubbard SJ, Argos P (1994) Cavities and packing at protein interfaces. Protein Sci 3: 2194–2206.
58. Briozzo P, Evrin C, Meyer P, Assairi L, Joly N, et al. (2005) Structure of Escherichia coli UMP kinase differs from that of other nucleoside monophosphate kinases and sheds new light on enzyme regulation. J Biol Chem 280: 25533–25540.
59. Meyer P, Evrin C, Briozzo P, Joly N, Barzu O, et al. (2008) Structural and Functional Characterization of Escherichia coli UMP Kinase in Complex with Its Allosteric Regulator GTP. J Biol Chem 283: 36011–36018.
60. Labesse G, Benkali K, Salard-Arnaud I, Gilles AM, Munier-Lehmann H (2011) Structural and functional characterization of the Mycobacterium tuberculosis uridine monophosphate kinase: insights into the allosteric regulation. Nucleic Acids Res 39: 3458–3472.
61. Hinsen K, Petrescu AJ, Dellerue S, Bellissent-Funel MC, Kneller GR (2000) Harmonicity in slow protein dynamics. Chem Phys 261: 25–37.
62. Li GH, Cui Q (2002) A coarse-grained normal mode approach for macromolecules: An efficient implementation and application to Ca2+-ATPase. Biophys J 83: 2457–2474.
63. Suhre K, Sanejouand YH (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res 32: W610–W614.
64. Bahar I, Atilgan AR, Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 2: 173–181.
65. Bahar I, Rader AJ (2005) Coarse-grained normal mode analysis in structural biology. Curr Opin Struct Biol 15: 586–592.
66. Nicolay S, Sanejouand YH (2006) Functional modes of proteins are among the most robust. Phys Rev Lett 96: 078104.
67. Kundu S, Melton JS, Sorensen DC, Phillips GN (2002) Dynamics of proteins in crystals: Comparison of experiment with simple models. Biophys J 83: 723–732.
68. Clarkson MW, Gilmore SA, Edgell MH, Lee AL (2006) Dynamic coupling and allosteric behavior in a nonallosteric protein. Biochemistry 45: 7693–7699.
69. Goodey NM, Benkovic SJ (2008) Allosteric regulation and catalysis emerge via a common route. Nat Chem Biol 4: 474–482.
70. Agarwal PK, Billeter SR, Rajagopalan PTR, Benkovic SJ, Hammes-Schiffer S (2002) Network of coupled promoting motions in enzyme catalysis. Proc Natl Acad Sci U S A 99: 2794–2799.
71. Suel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 10: 59–69.
72. Bode C, Kovacs IA, Szalay MS, Palotai R, Korcsmaros T, et al. (2007) Network analysis of protein dynamics. FEBS Lett 581: 2776–2782.
73. Chennubhotla C, Bahar I (2006) Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. Mol Syst Biol 2: 36.
74. Chennubhotla C, Bahar I (2007) Signal propagation in proteins and relation to equilibrium fluctuations. PLoS Comput Biol 3: 1716–1726.
75. Zheng WJ, Brooks BR (2005) Probing the local dynamics of nucleotide-binding pocket coupled to the global dynamics: Myosin versus kinesin. Biophys J 89: 167–178.
76. Ming D, Wall ME (2005) Allostery in a coarse-grained model of protein dynamics. Phys Rev Lett 95: 198103.
77. Ming DM, Wall ME (2006) Interactions in native binding sites cause a large change in protein dynamics. J Mol Biol 358: 213–223.
78. Anglada JM, Besalu E, Bofill JM, Crehuet R (2001) On the quadratic reaction path evaluated in a reduced potential energy surface model and the problem to locate transition states. J Comput Chem 22: 387–406.
79. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. J Mol Graph 14: 33–38.

4.3. Thermal stability of enzymes

So far we have considered the reactive and dynamical properties of enzymes. However, as we have pointed out in the introduction, an important aspect of enzymes is that they have evolved to perform their function under very specific environmental conditions. Of great interest are the enzymes working at high temperatures (thermophilic) for their potential biotechnological applications. In this section, we have studied how thermal adaptation can determine the dynamical properties of an enzyme. Put it in a broader context, this study on conformational and diffusive motions at short time scales complements our view on enzyme dynamics extracted from the previous section on slow conformational motions.

4.3.1. Flexibility and diffusion observed by neutron scattering

To better understand the stability mechanisms of thermophilic proteins, we have focused on a neutron scattering experiment[2] that set a new paradigm on the relationship between thermostability and flexibility. From the experiment, the authors suggested that the adaptation of thermophilic proteins to high temperatures lies in the lower sensitivity of their internal flexibility (at short time scales) to temperature changes (Figure 5). Amazingly, they also found that, at low temperature, the thermophilic enzyme is more flexible and less active, which is in sharp contrast to the traditional view of thermostable proteins as more rigid entities.

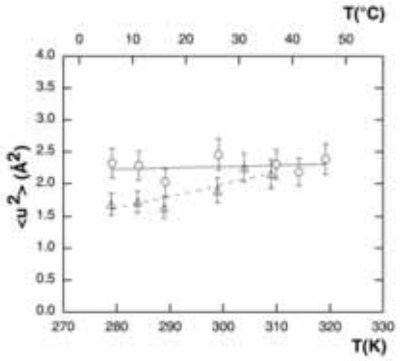


Figure 5. Mean-square-displacements (<u²>) measured by elastic incoherent neutron scattering[2]. Data for thermophilic malate dehydrogenase (circles) and mesophilic lactate dehydrogenase (triangles) enzymes.

[2]Tehei M. et al. (2005) *J Biol Chem* 280: 40974

The authors compared the dynamical properties of two tetrameric enzymes adapted to very different temperatures, i.e. hyperthermophilic malate dehydrogenase (100ºC) and a mesophilic lactate dehydrogenase (30ºC) that is very similar in terms of sequence and structure. Our aim was to elucidate the rationale behind that finding. Since the mean-square-displacement measured by neutron scattering can, in principle, include contributions from intramolecular dynamics and diffusion, we performed long molecular dynamics simulations (~200 ns) for describing the intramolecular motions of both enzymes and Brownian dynamics for their diffusion with a 1000-molecule box to account for crowding effects like in the experiment. Simulations were in agreement with the experimental observations, but only when taking into account the diffusion contribution. This study thus has implications on the interpretation of past and future neutron scattering experiments performed in solution. In addition, these results showed evidence that under crowding conditions, like *in vivo*, thermophilic and homologue mesophilic proteins have diffusional properties with different thermal behavior due to their different surface composition. The more intense electrostatic potential at the surface of the thermophilic protein, due to the larger number of charged residues, entails stronger electrostatic inter-protein interactions in solution that affect the diffusional behavior. This study opens up opportunities to further studies to ascertain whether this is a general trend and, if so, which biological implications might have.

A detailed presentation of the results and methodologies used in this study can be found in the article: *Crowding induces differences in the diffusion of thermophilic and mesophilic proteins: a new look at neutron scattering results* (2011) Biophys. J. 101: 2782-2789. A "New and Notable" comment on this article has been published in the same issue.

2782 Biophysical Journal Volume 101 December 2011 2782–2789

Crowding Induces Differences in the Diffusion of Thermophilic and Mesophilic Proteins: A New Look at Neutron Scattering Results

Enrique Marcos, Pau Mestres, and Ramon Crehuet*
Department of Biological Chemistry and Molecular Modeling, Institute of Advanced Chemistry of Catalonia (IQAC – CSIC), Barcelona, Spain

ABSTRACT The dynamical basis underlying the increased thermal stability of thermophilic proteins remains uncertain. Here, we challenge the new paradigm established by neutron scattering experiments in solution, in which the adaptation of thermophilic proteins to high temperatures lies in the lower sensitivity of their flexibility to temperature changes. By means of a combination of molecular dynamics and Brownian dynamics simulations, we report a reinterpretation of those experiments and show evidence that under crowding conditions, such as in vivo, thermophilic and homolog mesophilic proteins have diffusional properties with different thermal behavior.

INTRODUCTION

Thermophilic organisms require very high temperatures for surviving, and elucidating the adaptation strategies of their proteins can provide clues for designing new proteins with enhanced thermostability (1). This also provides novel opportunities to understand how protein sequence and structure are related with dynamics and function. To this aim, comparative studies of thermophilic proteins and their corresponding homologs working at low temperature, i.e., mesophilic, find great usefulness. There has been a long-standing controversy on the dynamical requirements for thermostability because different techniques are sensitive to dynamical processes at vastly different timescales (2–7). A few years ago, an innovative mechanism of protein thermostability was invoked by Zaccai and co-workers (8) based on elastic incoherent neutron scattering (EINS) experiments in solution. They suggested that the key dynamical feature required for protein thermostability was an enhanced resilience ($\langle k' \rangle$), which is defined as the inverse of the variation of the mean-square fluctuation ($\langle u^2 \rangle$) with temperature (9): $\langle k' \rangle = 1/(d\langle u^2 \rangle/dT)$. The smaller the temperature dependence of $\langle u^2 \rangle$ the higher the resilience. These experiments showed that, at the ~100 ps timescale, a thermophilic enzyme (Malate Dehydrogenase from *Methanococcus jannaschii*; *Mj*MalDH) was ~10 times more resilient than a mesophilic homolog (Lactate Dehydrogenase from *Oryctolagus cuniculus*; *Oc*LDH) (see *open symbols* in Fig. 1 *A*). In other words, the flexibility ($\langle u^2 \rangle$) of the thermophilic enzyme is less sensitive to temperature than that of the mesophilic one. They also observed that $\langle u^2 \rangle$ values of *Mj*MalDH were higher than those of *Oc*LDH in the temperature range studied (280–320 K), which ultimately means that higher resilience need not imply less flexibility. On the contrary, this new paradigm claims that thermoenzymes are more flexible even at low temperatures.

The correlation between resilience and thermostability has also been observed by neutron scattering experiments on whole cells from organisms adapted to different temperatures (10). Given that resilience has also been observed to be a property very sensitive to the solvent conditions (11), we question whether the higher resilience observed for thermophilic proteins arises only from protein internal dynamics, an issue that has not been addressed hitherto. In this work, we investigate the contributions from intramolecular motions and protein diffusion to the global dynamics of *Mj*MalDH and *Oc*LDH, as measured by EINS (8). We use molecular dynamics (MD) and Brownian dynamics (BD) simulations of the crowded solution (200 mg/mL) used in the experiment, which also mimics in vivo conditions. To our knowledge, this combination of MD and BD represents a new approach to incorporate crowding effects in the simulation of protein global dynamics in solution as explored by neutron scattering.

MD simulations have proven to be a valuable computational technique for exploring protein internal dynamics and are very suitable for examining neutron scattering data (12–17). MD can describe dynamical events at the short timescales (from picoseconds to hundreds of nanoseconds) that are usually explored by neutron scattering instruments. However, describing protein diffusion in solution with MD, taking into account interactions among diffusing proteins, is computationally unattainable. In contrast, BD simulations are well suited to explore translational and rotational diffusion and the interactions among hundreds of protein molecules at timescales from nanoseconds to milliseconds (18–20). Because one of the main assumptions of BD is ignoring protein internal dynamics, a combination of MD and BD simulations is promising in giving a global picture of protein dynamics in solution.

EINS experiments probe atomic motions within a space-time window defined by the characteristics of the

Submitted August 12, 2011, and accepted for publication September 23, 2011.

*Correspondence: ramon.crehuet@iqac.csic.es

Editor: Martin Blackledge.

0006-3495/11/12/2782/8 $2.00

doi: 10.1016/j.bpj.2011.09.033

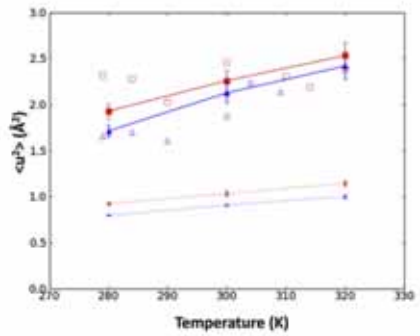


FIGURE 1 Experiment versus simulation. Comparison between experimental (8) and simulated $\langle u^2 \rangle$ values of the thermophilic (*red squares*) and mesophilic (*blue triangles*) proteins at the three temperatures studied. Simulated $\langle u^2 \rangle$ from intramolecular dynamics are represented with dashed lines, and $\langle u^2 \rangle$ including translational and rotational diffusion are represented with solid lines.

spectrometer and, in general, can cover motions of a few angstroms in hundreds of picoseconds (21). These experiments allow obtaining a measure of the average dynamics of a protein expressed as the atomic mean-square displacement (MSD) ($\langle u^2 \rangle$). To study the dynamics of a protein close to physiological conditions, these experiments are carried out in solution. This is a challenge when examining protein internal dynamics because contamination of the elastic scattering due to protein diffusion can be present (22–24) and, thus, specific sample conditions and instrumental settings are required to minimize this contribution. Because EINS experiments require high protein concentrations to be conducted satisfactorily (typically ~100–200 mg/mL), protein diffusion is hindered by such a crowded media, so that in many cases the diffusion contribution is assumed to be negligible. Furthermore, when comparing the dynamics of different proteins with similar size, it is usual practice to consider that their small diffusion is very similar and, hence, differences obtained in $\langle u^2 \rangle$ must account for differences in their intramolecular dynamics (25). Indeed, this was one of the implicit assumptions of Zaccai and co-workers (8). In this investigation we challenge this view by means of MD and BD simulations and show that a) the internal dynamics and diffusion both have similar contributions to the global flexibility measured for *Mj*MalDH and *Oc*LDH and that b) the diffusional properties of both proteins are strikingly different under crowding conditions explaining the distinct thermal behavior observed in the experiment.

MD and BD simulations were conducted at 280, 300, and 320 K spanning the range of temperatures experimentally studied. The crystal structures of both tetrameric proteins (PDB codes 1HYG (26) and 9LDT (27)) were used as inputs of the simulations. MD simulations were performed with Gromacs (28). Each MD simulation was equilibrated for 40 ns and a production run of 160 ns followed. To have an estimate of our error and to probe potentially different conformations, we have used five points of these long trajectories to simulate the EINS data, as further explained in the Supporting Material. BD simulations were carried out with the code developed by Elcock and co-workers (18,19). BD simulations of each protein were conducted under periodic boundary conditions in a 1000-molecule box with the same experimental concentration (200 mg/mL). The sensitivity of the results to BD parameters has also been considered (see the Supporting Material). We have generated time-trajectories (hereafter called MD+BD) composed of intramolecular motions from MD simulations and both translational and rotational diffusive motions from BD simulations. Comparison with experiment is possible by calculating the MSD ($\langle u^2 \rangle$) from computational results using the same data treatment methods as in the experiment (29). This implies calculating from the MD and MD+BD trajectories the basic quantity measured in neutron scattering, the scattering function $S(Q, \omega = 0)$, and including the effects of the energy resolution and Q-range of the experiment. The $\langle u^2 \rangle$ is obtained using the Gaussian approximation (see Materials and Methods and Supporting Material for details of the simulation protocols and analysis).

MATERIALS AND METHODS

Elastic incoherent neutron scattering

Neutron scattering experiments obtain information of the atomic motions in a sample by measuring the exchange of momentum (Q) and energy (ω) of the incident neutrons in a scattering process (21). The basic quantity obtained from these experiments that contains information on the dynamics of the sample is the dynamic structure factor $S(Q, \omega)$. The contribution from nonhydrogen atoms to incoherent scattering is negligible, given their much smaller scattering length. Because hydrogens are homogeneously distributed throughout a protein structure, the observed structure factor gives an average measure of the dynamics of the sample.

The energies of incident neutrons are described by the energy resolution function, $R(\omega)$, which depends on the neutron spectrometer. The width of $R(\omega)$ determines the timescale of motions accessible by the instrument, with narrower widths corresponding to longer timescales. The momentum transfer, on the other hand, defines the spatial scale of motions accessible by the instrument. Therefore, $S(Q, \omega)$ gives information on the dynamics of the sample within a well-defined space-time window of observation. The EINS experiments in (8) were done on the IN13 instrument at ILL (Grenoble, France), which has an energy resolution of 8 μeV (full width at half-maximum). The Q-range used was 1.2–2.2 $Å^{-1}$. These experimental conditions have access to motions of ~1 Å in ~100 ps.

The elastic peak corresponds to the structure factor measured without exchange of energy, $S(Q, \omega = 0)$. We employed the Gaussian approximation for extracting the MSD of the sample. The elastic intensity is Q dependent and for confined motions takes the following form:

$$S(Q, \omega = 0) = e^{-\frac{1}{6}\langle u^2 \rangle Q^2}, \quad (1)$$

where $\langle u^2 \rangle$ is the MSD averaged over the atoms in the protein. Linearization of Eq. 1 by taking the slope of a natural log plot of $S(Q, \omega = 0)$ vs. Q^2 (Guinier plot) allows obtaining $\langle u^2 \rangle$.

Following Hayward and Smith (29), for a quantitative comparison between simulated trajectories and the experimental MSDs we computed the dynamic structure factor and then extracted the MSD with Eq.1. To compute S(Q, $\omega t = 0$), we first need to calculate $F_{inc}(Q, t)$ in the Q range studied experimentally using Eq. 2:

$$F_{inc}(Q, t) = \frac{1}{N}\sum_{\alpha} b^2_{\alpha,\,inc}\left\langle e^{i\mathbf{q}\cdot(\mathbf{R}(t)-\mathbf{R}(0))}\right\rangle. \qquad (2)$$

For each experimental Q value, $F_{inc}(Q, t)$ is an average over a number of q vectors with random orientations and the same modulus $Q = |\mathbf{q}|$. In this work, we have averaged $F_{inc}(Q, t)$ with 50 $\mathbf{q}$ vectors to guarantee an isotropic distribution of $\mathbf{q}$ vectors. We finally compute the convoluted structure factor by Fourier transforming the product $F_{inc}(Q, t)\cdot R(t)$, where R(t) is the Fourier transform of the energy resolution function $R(\omega)$, described as a Lorentzian. The analysis of simulated trajectories to obtain neutron scattering properties has been performed with the nMoldyn program (30).

Molecular dynamics

All MD simulations were performed with the GROMACS 4.0.5 package (28). Each simulated system consisted of a single tetrameric protein immersed in a rhombic dodecahedral water box with Na^+ and Cl^- ions that were added for neutrality. Standard protonation states were assigned to all protein residues. In agreement with the experiment, substrates bound to the crystallographic structures were removed. The total numbers of atoms in the simulations of the thermophilic and mesophilic proteins are 89,229 and 90,393, respectively.

The two systems under study were subjected to ~200 ns simulations at 280, 300, and 320 K at constant temperature and pressure. The temperature was kept constant with the Berendsen thermostat (31) with a coupling time constant of 0.1 ps. The pressure was controlled with the Berendsen barostat (31) with a coupling constant of 0.5 ps and an isotropic compressibility of $4.5\cdot10^{-5}$ bar^{-1}. The OPLS all-atom (32) force field was used in combination with the TIP3P model (33) for water molecules. Periodic boundary conditions were used. Short-range electrostatic interactions were calculated explicitly with a 10 Å cutoff and long-range electrostatic interactions were calculated with the particle mesh Ewald method (34) with a grid spacing of 1.2 Å and a fourth-order spline interpolation. Lennard-Jones interactions were calculated using a switch function between 0.8 and 0.9 nm. All bonds were constrained using the LINCS algorithm (35), which allowed using an integration time step of 2 fs.

Each MD simulation was setup as follows. The structure of the solvated protein was energy minimized with the steepest descent algorithm. Next, to equilibrate the solvent surrounding the protein an MD simulation at the target temperature was performed with harmonic position restraints on the heavy atoms of the protein with a force constant of 1000 kJ mol^{-1} nm^{-2}. Subsequently, a 200 ns trajectory was carried out and the last 160 ns were used for production. We have performed simulations with two aims. First, we ran long trajectories (~200 ns and saved each 20 ps) to have access to different conformations. Second, we have run five short trajectories (2 ns length and saved each 1 ps) starting from different regions of the potential energy surface explored by the long MD trajectories (at 10, 50, 100, and 160 ns after the 40 ns of equilibration). The aim of these short trajectories is to assess the conformational dependence of intramolecular motions at the 100 ps timescale probed by neutron scattering. To analyze the contribution of intramolecular dynamics to $\langle u^2\rangle$ we have subtracted translations and rotations by superimposing each frame to a reference structure.

Brownian dynamics

We have followed the same setup as Elcock and co-workers (19) (see the Supporting Material for details). BD simulations for each protein were conducted for 10 μs at 280, 300, and 320 K under periodic boundary conditions in a cubic cell of 1000 molecules. The simulations were equilibrated for 2 μs and the last 8 μs were used for production. A time step of 2.5 ps was used for all simulations. The production runs of 8 μs were used for computing radial distribution functions. In addition, after the 2 μs of equilibration another production run of 2 ns with a shorter time step (1 ps) was followed for each protein at each temperature to describe the diffusive motion at the experimental timescale of 100 ps with better resolution. These 2 ns trajectories were used to generate MD+BD trajectories (see below) for computing neutron scattering properties.

The translational diffusion coefficient (self-diffusion) was calculated from the Einstein relation $D_{transl}(\Delta t) = MSD/(6\Delta t)$, where MSD is computed as

$$MSD(\Delta t) = \frac{1}{N(T-\Delta t+1)}\sum_{i}^{N}\sum_{t=0}^{T-\Delta t}\left[R_i(t+\Delta t) - R_i(t)\right]^2, \qquad (3)$$

where MSD is averaged over N atoms, T is the time length of the trajectory, and Δt the time separation between saved frames.

Generation of MD+BD trajectories

To describe the global dynamics of a protein molecule we have treated the intramolecular protein motions decoupled from rigid-body diffusive motions (translations and rotations), a realistic assumption as shown in (36). We have generated a trajectory composed by intramolecular motions described by MD simulations (after subtraction of translations and rotations) and diffusive motions obtained from BD simulations. Herein this trajectory will be referred as MD+BD trajectory and is generated as

$$\mathbf{r}^m(t) = \mathbf{r}_{MD}(t)\cdot\mathbf{R}^m_{rot}(t) + \mathbf{R}^m_{cm}(t) \qquad (4)$$

where $\mathbf{r}^m(t)$ and $\mathbf{r}_{MD}(t)$ are the 3N-dimensional vectors of atomic coordinates of the MD+BD trajectory for molecule m and the MD trajectory respectively, $\mathbf{R}^m_{rot}(t)$ is the 3×3 rotational matrix of molecule m at time t and $\mathbf{R}^m_{cm}(t)$ is the time-dependent position of the center of mass of molecule m.

We have generated 60 MD+BD trajectories from 60 randomly chosen molecules from the 1000-molecule box to compute $\langle u^2\rangle$ at each temperature. We have proven that the obtained $\langle u^2\rangle$ is well converged with 60 random molecules. Because of the negligible conformational dependence of the short time MD trajectories, we have used the same MD trajectory to build each MD+BD trajectory. Therefore, the difference between these 60 trajectories will come from differences in the rotational and translational diffusive motions of each molecule. Subsequently, S(Q, $\omega = 0$) was calculated for each molecule and averaged for all molecules at each Q value. A Guinier plot of these averaged S(Q) values was done for extracting the $\langle u^2\rangle$ of the crowded solution.

Figures including molecular structures and electrostatic potentials have been generated using VMD (37) and APBS (38,39). The structural alignments have been performed with DALI (40).

RESULTS AND DISCUSSION

Intramolecular dynamics

The simulated $\langle u^2\rangle$ obtained from the MD simulations (*dashed lines* in Fig. 1) are qualitatively consistent with the experiment (8) in that the thermophilic protein is more flexible than the mesophilic one, but they are underestimated by ~1 $Å^2$ with respect to the experiment (8). The thermophilic protein has higher flexibility both in the backbone

(Fig. 2) and in all residue atoms (Fig. S1 in the Supporting Material). We have quantified the temperature dependence of $\langle u^2 \rangle$ with the slope of a linear fit of $\langle u^2 \rangle$ versus temperature, $d\langle u^2 \rangle/dT$, and noted that it is the same for both proteins (0.005 $Å^2\ K^{-1}$). This is also in contrast to the experiment (8) where $d\langle u^2 \rangle/dT$ was observed to be much lower for the thermophilic protein (0.002 vs. 0.020 $Å^2\ K^{-1}$). We have checked that the conformational dependence of intramolecular $\langle u^2 \rangle$ is negligible (see *small error bars in dashed lines* from Fig. 1). The significant difference between simulated and experimental $\langle u^2 \rangle$ values indicate that, in the experiment, the apparent $\langle u^2 \rangle$ values not only correspond to internal protein dynamics and that an important contribution from translational and rotational diffusion must be present. This is also supported by former MD studies that have been successful in reproducing neutron scattering data from a wide range of biological systems where protein diffusion was absent, i.e., protein powders (12–15) and membrane proteins (16,17).

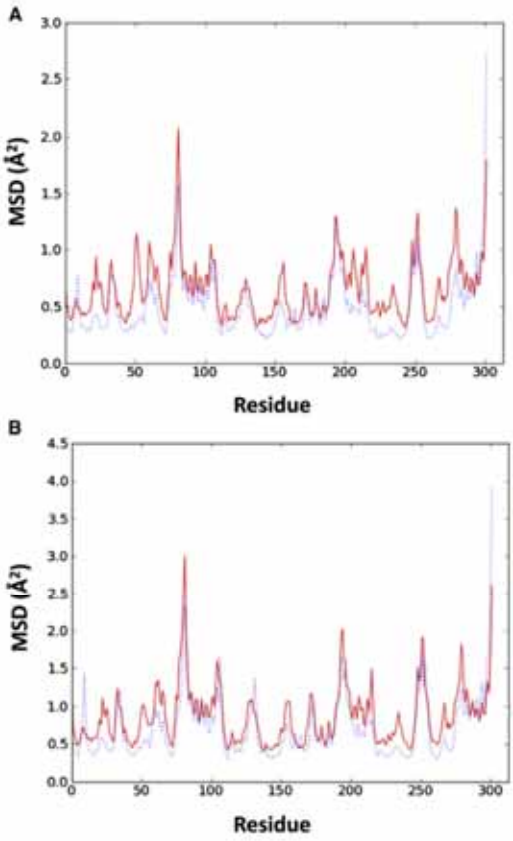


FIGURE 2 Thermophilic enzyme is more flexible in almost all regions. Representation of the MSD of the backbone atoms of the thermophilic (*red*) and mesophilic (*blue*) proteins residues at (*A*) 280 K and (*B*) 320 K, and is averaged over the four chains. For better comparison, the residues of both proteins have been structurally aligned. Remark that flexible regions are correlated between both proteins. The higher mobility of the thermophile is distributed throughout all the residues, and is not the result of a particularly flexible region. When all the atoms of the residue are included, higher fluctuations are observed but the trend does not change (see Fig. S1).

Macromolecular diffusion in the crowded media

Now we turn our attention to the contribution from diffusion to the measured $\langle u^2 \rangle$. A snapshot of a BD simulation of the crowded solution (thermophilic enzyme at 280 K) is shown in Fig. 3 *A*. Fig. 1 shows the $\langle u^2 \rangle$ values obtained from MD+BD trajectories (see *solid lines*). The agreement between the simulated and experimental $\langle u^2 \rangle$ values has been significantly improved. Both translational and

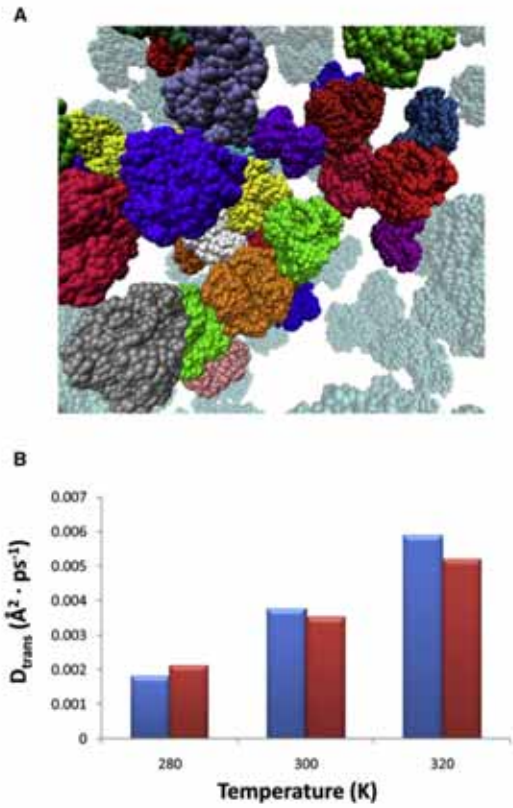


FIGURE 3 Diffusion in the crowded media. (*A*) Snapshot of the thermophilic protein solution simulated with Brownian dynamics at 200 mg/mL and at 280 K. (*B*) Translational diffusion coefficients at 100 ps. Red and blue bars correspond to the thermophilic and mesophilic proteins, respectively, at the three temperatures studied.

rotational diffusion contribute to $\langle u^2 \rangle$ and their separated contributions are shown in Fig. S2. This is of the utmost importance for a correct interpretation of the experiment. From our results, both internal protein dynamics and diffusion have similar contributions to the experimental $\langle u^2 \rangle$ and, therefore, underestimating the diffusion contribution can lead to erroneous interpretation of EINS data, as noted by several authors (22–24). In that respect, we would like to underscore the importance of measurements on self-diffusion coefficients as an ideal complement of EINS experiments in solution.

The agreement between the experimental and simulated $d\langle u^2 \rangle/dT$ of the mesophilic protein is striking, being 0.020 and 0.018 $\text{Å}^2 \cdot \text{K}^{-1}$, respectively. For the thermophilic protein, the simulated $d\langle u^2 \rangle dT$ (0.015 $\text{Å}^2\ \text{K}^{-1}$) is lower (see also the Supporting Material) than that of the mesophilic protein, in accord with the experiment. For the thermophile we get a slightly lower $d\langle u^2 \rangle/dT$ than the mesophile. This value is still much higher than the experimental one, which suggests that we might be underestimating the difference in diffusional properties of the mesophile and the thermophile. In particular, the aggregation effects of the thermophile (vide infra) could need a better treatment than the Brownian model used. This fact, instead of invalidating our conclusions, underscores that assuming similar diffusional properties for both molecules is not correct, despite their similar size and shape.

Fig. 3 *B* shows the translational diffusion coefficient (at the experimental timescale of 100 ps) obtained from BD simulations of thermophilic and mesophilic proteins: the diffusion of the mesophilic protein is more temperature dependent than that of the thermophilic one. Because of this, the global dynamics and, thus, the $\langle u^2 \rangle$ of the thermophilic protein become less temperature-dependent than that of the mesophilic one, as pointed out above.

Interparticle interactions in the crowded media

Despite the similar weight and size of the two proteins, the different amino acid composition of their surfaces must be responsible for important differences in their diffusional properties within the crowded environment, where excluded-volume effects enhance protein-protein interactions (41). Indeed, thermophilic proteins are characterized by their higher proportion of charged residues (Asp, Glu, Lys, Arg) (42) in the surface relative to mesophilic homologs and thus electrostatic and hydrophobic protein-protein interactions are comparatively different in both proteins. Here, we explore how the amino acid composition of protein surfaces can affect the diffusional properties. Fig. 4 *A* illustrates the differences in the electrostatic potential of both proteins. The higher intensity of colors and the longer field lines in the thermophilic protein represent the stronger electrostatic interactions that are present among this type of molecules. Fig S3 plots the electrostatic energy of the six BD simulations and clearly shows that electrostatic interactions in the thermophilic protein system are notoriously more stabilizing than in the mesophilic one at the three temperatures studied. Despite the higher net charges of the thermophilic protein (−24 vs. 6 electronic units), electrostatic interactions are not necessarily repulsive, because the heterogeneous charge distribution at the surface (see Fig. 4 *A*) allows protein molecules to interact through oppositely charged protein areas by changing their relative orientations. It turns out that the increased proportion of charged residues in the thermophilic protein system results in more intense and favorable electrostatic interactions.

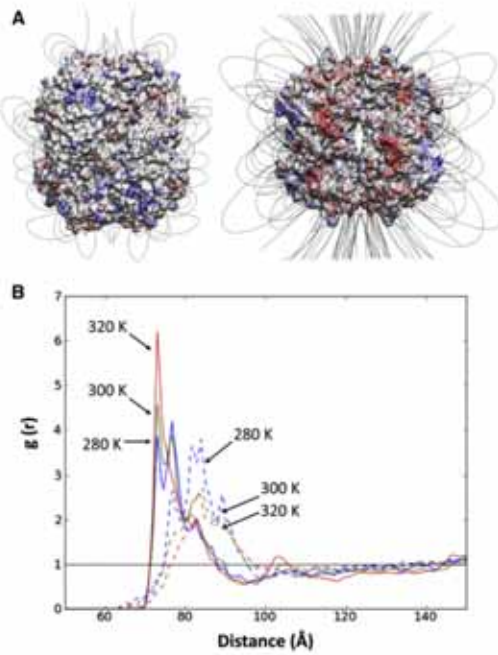


FIGURE 4 Electrostatic interactions. (*A*) Molecular structures of the mesophilic (*left*) and thermophilic (*right*) proteins colored according to the electrostatic potential at 300 K. Field lines are also represented. Blue and red colors represent areas of the protein where positive and negative charge density is accumulated. (*B*) Radial distribution functions, g(r), obtained from simulations of the thermophilic (*solid lines*) and mesophilic (*dashed lines*) proteins at 280 K (*blue*), 300 K (*green*), and 320 K (*red*).

The difference in the mutual interactions between thermophilic and mesophilic proteins is well illustrated by the radial distribution functions, g(r), obtained from BD simulations (see Fig. 4 *B*). Two remarkable differences between both proteins are observed. First, the thermophilic protein displays stronger short-range attractive interactions than the mesophilic one, as shown by higher peaks at shorter interprotein distances (both proteins have a radius of ~35 Å). Second, these short-range attractive interactions in the

thermophilic protein are enhanced upon increasing temperature, whereas in the mesophilic one they are weakened. Both features are in line with the observation in Fig. S3 that the electrostatic energy in the thermophilic protein system is lower, as pointed out previously, and that it decreases with temperature as opposed to the mesophilic protein system. Indeed, it is widely accepted that electrostatic interactions between oppositely charged protein residues increase with temperature due to a reduction of the dielectric constant of water, which ultimately leads to a reduction in the desolvation penalty required for salt bridge formation. Such enhancement of salt bridge interactions within a protein has been quantified in (43–47) and has been suggested as a key mechanism underlying the increased thermostability of thermophilic proteins, which have an increased number of charged residues (42). The thermal stabilization of electrostatic interactions in the thermophilic protein system implies that, upon increasing temperature, the interacting molecules will be closer to each other enhancing excluded volume effects and forming transient complexes that, due to their larger size, tend to diffuse at a lower rate. In view of this, the natural increase in diffusion with temperature is partially compensated by enhanced attractive interactions that increase the population of slower diffusive clusters. We suggest that this effect accounts for the differences we observe from BD in the temperature dependence of g(r) and, ultimately, in the diffusion of both proteins. Indeed, the smaller temperature dependence of the self-diffusion coefficient of the thermophilic protein can be viewed as an extension of the corresponding states model in that the thermophilic and mesophilic proteins have similar diffusional properties at their optimum temperature for activity. The implications of the enhancement of electrostatic interactions with temperature on diffusion had not been described so far. Although we already captured the main trends observed in the experiment, the different surface composition can result in other effects not taken into account by the present model. The description of the solvent as a continuum in the BD simulations oversights hydrodynamic interactions (48) and the effects of hydration water (49–51), which has been shown to play a key role in the association kinetics between proteins (52) and in protein hydrodynamics (53). In this regard, former studies (54,55) already showed that hydration water molecules of thermophilic proteins are more densely packed and exhibit less mobility than those of mesophilic homologs due to the aforementioned differences in surface composition. Therefore, hydration effects are likely to modulate the association events already observed from BD and, thus, protein diffusion. Overall, differences in interprotein interactions, hydration, and hydrodynamic interactions will largely account for the observed differences in the macromolecular motion of both enzymes.

Earlier experimental studies showed that variations in the effective charge of the bovine serum albumin, by changing the pH (56) and ionic strength (57), tune interprotein electrostatic interactions and, as a consequence, affect the diffusional behavior. This sensitivity of interprotein interactions to charge variations is fully consistent with our observation that the differences in the electrostatic properties of the thermo-mesophilic pair lead to different diffusional properties.

CONCLUSIONS

We can ultimately argue that the EINS experiment performed by Zaccai and co-workers (8), instead of revealing specific features of internal flexibility linked to thermostability, shows how a thermo-mesophilic pair of proteins can have very different diffusional properties within a crowded media, despite having strikingly similar shape and molecular weight. This illustrates the important implications of protein diffusion on the interpretation of EINS data and opens new perspectives on other previous studies in solution (10,11,25). Ultimately, this supports the idea that the concept of resilience (9) must be handled with care to characterize protein internal dynamics in solution.

Yet, no comparative studies on the diffusion of thermo-mesophilic pairs have been reported. To our knowledge, this is the first study pointing to important differences in the diffusional properties of a thermo-mesophilic pair of proteins based on theoretical and experimental data. We anticipate that this might be a general difference between thermophilic and mesophilic proteins opening up opportunities to new experimental studies. We wonder whether such difference in the thermal behavior of diffusion is linked to a requirement for maximum activity at the corresponding physiological conditions where crowding effects dominate. We leave as future work experimental studies on the diffusion of both proteins. Given the vast diversity of dynamical events at the intramolecular and diffusion level, we will also explore the dynamics of this thermo-mesophilic pair at longer timescales. This may help us to understand why the thermophilic enzyme is less active at low temperatures, though being more flexible at the picosecond timescale here studied.

SUPPORTING MATERIAL

Details of the simulation protocols and analysis, four figures, and references (58–67) are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(11)01124-6.

The authors thank Adrian H. Elcock for providing his Brownian Dynamics code and useful suggestions on simulations. We are grateful to Joe Zaccai, Jeremy C. Smith, Frank Gabel, and Pau Bernadó for valuable discussions. We thank the Galicia Supercomputing Center for computational resources.

This work was supported by grants from the Junta de Ampliación de Estudios (JAE) programme of Consejo Superior de Investigaciones Científicas, the Spanish Ministerio de Educación Ciencia (MEC) (CTQ2009-08223), and the Catalan Agency for Management of University and Research Grants (AGAUR) (2005SGR00111).

REFERENCES

1. Vieille, C., and G. J. Zeikus. 2001. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* 65:1–43.
2. Závodszky, P., J. Kardos, …, G. A. Petsko. 1998. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl. Acad. Sci. USA.* 95:7406–7411.
3. Hernandez, G., F. E. Jenney, Jr., …, D. M. LeMaster. 2000. Millisecond time scale conformational flexibility in a hyperthermophile protein at ambient temperature. *Proc. Natl. Acad. Sci. USA.* 97:3166–3170.
4. Fitter, J., and J. Heberle. 2000. Structural equilibrium fluctuations in mesophilic and thermophilic alpha-amylase. *Biophys. J.* 79:1629–1636.
5. Wolf-Watz, M., V. Thai, …, D. Kern. 2004. Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat. Struct. Mol. Biol.* 11:945–949.
6. Butterwick, J. A., J. Patrick Loria, …, A. G. Palmer, 3rd. 2004. Multiple time scale backbone dynamics of homologous thermophilic and mesophilic ribonuclease HI enzymes. *J. Mol. Biol.* 339:855–871.
7. Salmon, L., G. Bouvignies, …, M. Blackledge. 2011. Nuclear magnetic resonance provides a quantitative description of protein conformational flexibility on physiologically important time scales. *Biochemistry.* 50:2735–2747.
8. Tehei, M., D. Madern, …, G. Zaccai. 2005. Neutron scattering reveals the dynamic basis of protein adaptation to extreme temperature. *J. Biol. Chem.* 280:40974–40979.
9. Zaccai, G. 2000. How soft is a protein? A protein dynamics force constant measured by neutron scattering. *Science.* 288:1604–1607.
10. Tehei, M., B. Franzetti, …, G. Zaccai. 2004. Adaptation to extreme environments: macromolecular dynamics in bacteria compared in vivo by neutron scattering. *EMBO Rep.* 5:66–70.
11. Tehei, M., D. Madern, …, G. Zaccai. 2001. Fast dynamics of halophilic malate dehydrogenase and BSA measured by neutron scattering under various solvent conditions influencing protein stability. *Proc. Natl. Acad. Sci. USA.* 98:14356–14361.
12. Tarek, M., and D. J. Tobias. 1999. Environmental dependence of the dynamics of protein hydration water. *J. Am. Chem. Soc.* 121:9740–9741.
13. Tarek, M., G. J. Martyna, and D. J. Tobias. 2000. Amplitudes and frequencies of protein dynamics: analysis of discrepancies between neutron scattering and molecular dynamics simulations. *J. Am. Chem. Soc.* 122:10450–10451.
14. Tarek, M., and D. J. Tobias. 2000. The dynamics of protein hydration water: a quantitative comparison of molecular dynamics simulations and neutron-scattering experiments. *Biophys. J.* 79:3244–3257.
15. Wood, K., A. Frölich, …, M. Weik. 2008. Coincidence of dynamical transitions in a soluble protein and its hydration water: direct measurements by neutron scattering and MD simulations. *J. Am. Chem. Soc.* 130:4586–4587.
16. Wood, K., S. Grudinin, …, G. Zaccai. 2008. Dynamical heterogeneity of specific amino acids in bacteriorhodopsin. *J. Mol. Biol.* 380: 581–591.
17. Wood, K., D. J. Tobias, …, M. Weik. 2010. The low-temperature inflection observed in neutron scattering measurements of proteins is due to methyl rotation: direct evidence using isotope labeling and molecular dynamics simulations. *J. Am. Chem. Soc.* 132:4990–4991.
18. McGuffee, S. R., and A. H. Elcock. 2006. Atomically detailed simulations of concentrated protein solutions: the effects of salt, pH, point mutations, and protein concentration in simulations of 1000-molecule systems. *J. Am. Chem. Soc.* 128:12098–12110.
19. McGuffee, S. R., and A. H. Elcock. 2010. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLOS Comput. Biol.* 6:e1000694.
20. Mereghetti, P., R. R. Gabdoulline, and R. C. Wade. 2010. Brownian dynamics simulation of protein solutions: structural and dynamical properties. *Biophys. J.* 99:3782–3791.
21. Gabel, F., D. Bicout, …, G. Zaccai. 2002. Protein dynamics studied by neutron scattering. *Q. Rev. Biophys.* 35:327–367.
22. Pérez, J., J. M. Zanotti, and D. Durand. 1999. Evolution of the internal dynamics of two globular proteins from dry powder to solution. *Biophys. J.* 77:454–469.
23. Hayward, J. A., J. L. Finney, …, J. C. Smith. 2003. Molecular dynamics decomposition of temperature-dependent elastic neutron scattering by a protein solution. *Biophys. J.* 85:679–685.
24. Gabel, F. 2005. Protein dynamics in solution and powder measured by incoherent elastic neutron scattering: the influence of Q-range and energy resolution. *Eur. Biophys. J.* 34:1–12.
25. Meinhold, L., D. Clement, …, J. C. Smith. 2008. Protein dynamics and stability: the distribution of atomic fluctuations in thermophilic and mesophilic dihydrofolate reductase derived using elastic incoherent neutron scattering. *Biophys. J.* 94:4812–4818.
26. Lee, B. I., C. Chang, …, S. W. Suh. 2001. Crystal structure of the MJ0490 gene product of the hyperthermophilic archaebacterium *Methanococcus jannaschii*, a novel member of the lactate/malate family of dehydrogenases. *J. Mol. Biol.* 307:1351–1362.
27. Dunn, C. R., H. M. Wilks, …, J. J. Holbrook. 1991. Design and synthesis of new enzymes based on the lactate dehydrogenase framework. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 332:177–184.
28. Hess, B., C. Kutzner, …, E. Lindahl. 2008. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* 4:435–447.
29. Hayward, J. A., and J. C. Smith. 2002. Temperature dependence of protein dynamics: computer simulation analysis of neutron scattering properties. *Biophys. J.* 82:1216–1225.
30. Róg, T., K. Murzyn, …, G. R. Kneller. 2003. nMoldyn: a program package for a neutron scattering oriented analysis of molecular dynamics simulations. *J. Comput. Chem.* 24:657–667.
31. Berendsen, H. J. C., J. P. M. Postma, …, J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.
32. Jorgensen, W. L., D. S. Maxwell, and J. Tirado-Rives. 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* 118:11225–11236.
33. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.
34. Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald - an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089–10092.
35. Hess, B., H. Bekker, …, J. Fraaije. 1997. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* 18:1463–1472.
36. Hamaneh, M. B., L. Zhang, and M. Buck. 2011. A direct coupling between global and internal motions in a single domain protein? MD Investigation of extreme scenarios. *Biophys. J.* 101:196–204.
37. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38, 27–28.
38. Baker, N. A., D. Sept, …, J. A. McCammon. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA.* 98:10037–10041.
39. Fogolari, F., A. Brigo, and H. Molinari. 2002. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recognit.* 15:377–392.
40. Holm, L., S. Kääriäinen, …, A. Schenkel. 2008. Searching protein structure databases with DaliLite v.3. *Bioinformatics.* 24:2780–2781.
41. Ellis, R. J. 2001. Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.* 26:597–604.

42. Kumar, S., and R. Nussinov. 2001. How do thermophilic proteins deal with heat? *Cell. Mol. Life Sci.* 58:1216–1233.
43. de Bakker, P. I. W., P. H. Hünenberger, and J. A. McCammon. 1999. Molecular dynamics simulations of the hyperthermophilic protein sac7d from *Sulfolobus acidocaldarius*: contribution of salt bridges to thermostability. *J. Mol. Biol.* 285:1811–1830.
44. Thomas, A. S., and A. H. Elcock. 2004. Molecular simulations suggest protein salt bridges are uniquely suited to life at high temperatures. *J. Am. Chem. Soc.* 126:2208–2214.
45. Danciulescu, C., R. Ladenstein, and L. Nilsson. 2007. Dynamic arrangement of ion pairs and individual contributions to the thermal stability of the cofactor-binding domain of glutamate dehydrogenase from *Thermotoga maritima*. *Biochemistry.* 46:8537–8549.
46. Vinther, J. M., S. M. Kristensen, and J. J. Led. 2010. Enhanced stability of a protein with increasing temperature. *J. Am. Chem. Soc.* 133: 271–278.
47. Karshikoff, A., and R. Ladenstein. 2001. Ion pairs and the thermotolerance of proteins from hyperthermophiles: a "traffic rule" for hot roads. *Trends Biochem. Sci.* 26:550–556.
48. Ando, T., and J. Skolnick. 2010. Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion. *Proc. Natl. Acad. Sci. USA.* 107:18457–18462.
49. Nucci, N. V., M. S. Pometun, and A. J. Wand. 2011. Site-resolved measurement of water-protein interactions by solution NMR. *Nat. Struct. Mol. Biol.* 18:245–249.
50. Halle, B. 2004. Protein hydration dynamics in solution: a critical survey. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 359:1207–1223, discussion 1223–1224, 1323–1328.
51. Modig, K., E. Liepinsh, …, B. Halle. 2004. Dynamics of protein and peptide hydration. *J. Am. Chem. Soc.* 126:102–114.
52. Ahmad, M., W. Gu, …, V. Helms. 2011. Adhesive water networks facilitate binding of protein interfaces. *Nat. Commun.* 2:261.
53. Halle, B., and M. Davidovic. 2003. Biomolecular hydration: from water dynamics to hydrodynamics. *Proc. Natl. Acad. Sci. USA.* 100: 12135–12140.
54. Melchionna, S., R. Sinibaldi, and G. Briganti. 2006. Explanation of the stability of thermophilic proteins based on unique micromorphology. *Biophys. J.* 90:4204–4212.
55. Sterpone, F., C. Bertonati, …, S. Melchionna. 2009. Key role of proximal water in regulating thermostable proteins. *J. Phys. Chem. B.* 113:131–137.
56. Meechai, N., A. M. Jamieson, and J. Blackwell. 1999. Translational diffusion coefficients of bovine serum albumin in aqueous solution at high ionic strength. *J. Colloid Interface Sci.* 218:167–175.
57. Roosen-Runge, F., M. Hennig, …, F. Schreiber. 2010. Protein diffusion in crowded electrolyte solutions. *Biochim. Biophys. Acta.* 1804:68–75.
58. Dolinsky, T. J., J. E. Nielsen, …, N. A. Baker. 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* 32(Web Server issue):W665–W667.
59. Li, H., A. D. Robertson, and J. H. Jensen. 2005. Very fast empirical prediction and rationalization of protein pK(a) values. *Proteins.* 61:704–721.
60. Gabdoulline, R. R., and R. C. Wade. 1996. Effective charges for macromolecules in solvent. *J. Phys. Chem.* 100:3868–3878.
61. Gabdoulline, R. R., and R. C. Wade. 1997. Simulation of the diffusional association of barnase and barstar. *Biophys. J.* 72:1917–1929.
62. García De La Torre, J., M. L. Huertas, and B. Carrasco. 2000. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophys. J.* 78:719–730.
63. Hinsen, K. 2000. The molecular modeling toolkit: A new approach to molecular simulations. *J. Comput. Chem.* 21:79–85.
64. Haynes, W. M., editor. 2010–2011. CRC Handbook of Chemistry and Physics, 91st Ed. CRC Press, Boca Raton, FL.
65. Madern, D., C. Ebel, …, G. Zaccai. 2001. Differences in the oligomeric states of the LDH-like L-MalDH from the hyperthermophilic archaea *Methanococcus jannaschii* and *Archaeoglobus fulgidus*. *Biochemistry.* 40:10310–10316.
66. Velev, O. D., E. W. Kaler, and A. M. Lenhoff. 1998. Protein interactions in solution characterized by light and neutron scattering: comparison of lysozyme and chymotrypsinogen. *Biophys. J.* 75:2682–2697.
67. Svergun, D. I., S. Richard, …, G. Zaccai. 1998. Protein hydration in solution: experimental observation by x-ray and neutron scattering. *Proc. Natl. Acad. Sci. USA.* 95:2267–2272.

SUPPORTING MATERIAL

Crowding Induces Differences in the Diffusion of Thermophilic and Mesophilic proteins: a New Look at Neutron Scattering Results

Enrique Marcos, Pau Mestres and Ramon Crehuet*

Department of Biological Chemistry and Molecular Modeling, Catalan Institute of Advanced Chemistry (IQAC - CSIC), E-08034 Barcelona, Spain

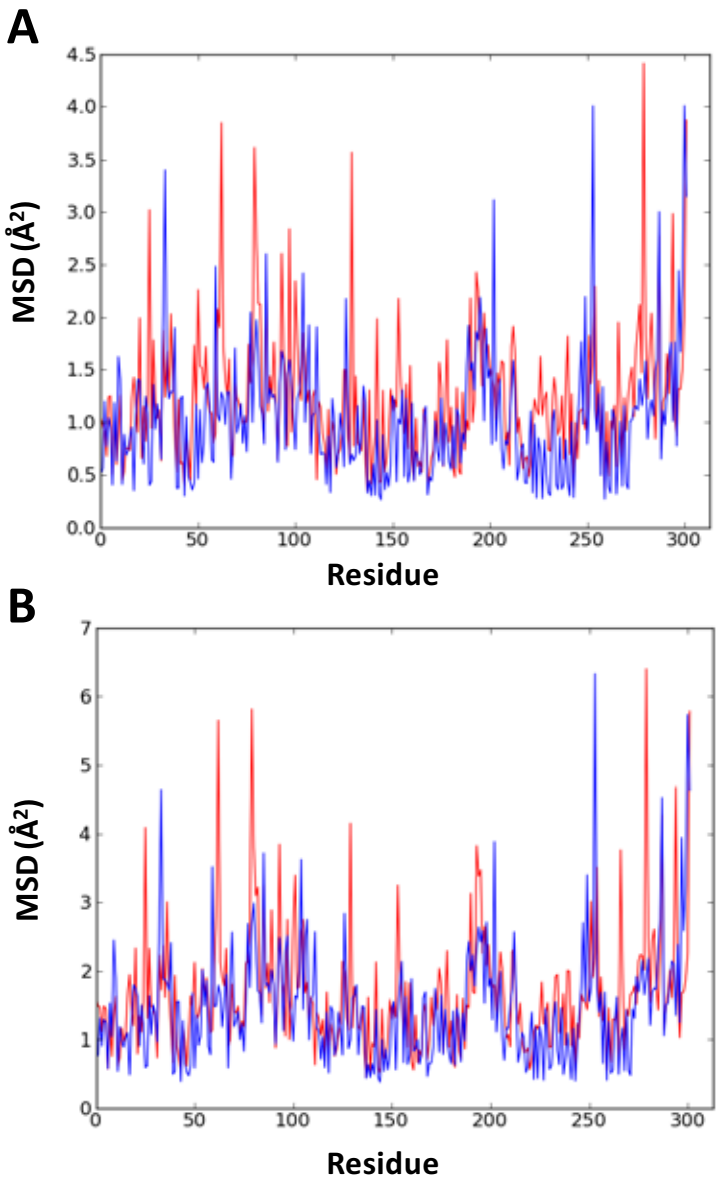


Figure S1. Representation of the mean-square-displacement (MSD) of all the atoms of the thermophilic (red) and mesophilic (blue) proteins residues at (A) 280 K and (B) 320 K. The MSD is computed with Eq. 3 (see *Materials and Methods*) at the time scale of 100 ps and is averaged over the 4 chains. For better comparison, the residues of both proteins have been structurally aligned with DALI [1].

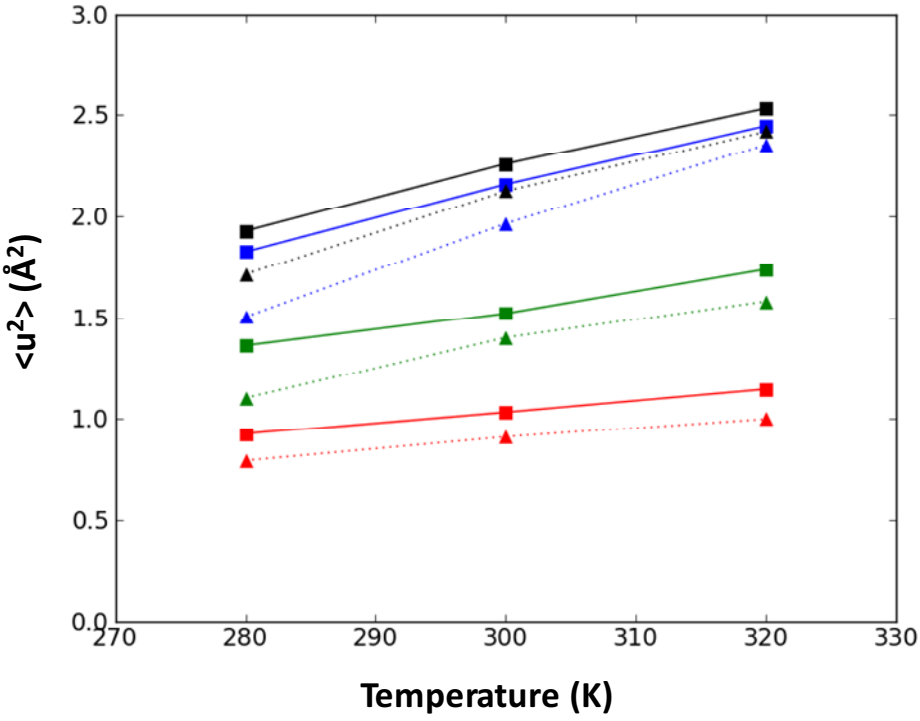


Figure S2. Representation of the simulated $\langle u^2 \rangle$ for the thermophilic (squares and solid lines) and mesophilic (triangles and dashed lines) proteins taking into account different contributions: internal dynamics (red), internal dynamics + rotational diffusion (green), internal dynamics + translational diffusion (blue) and internal dynamics + translational + rotational diffusion (black). Note that the different dynamical contributions to $\langle u^2 \rangle$ are not additive, since there is not a linear relationship between the atomic displacement and the scattering function from which $\langle u^2 \rangle$ is calculated (see *Materials and Methods*)

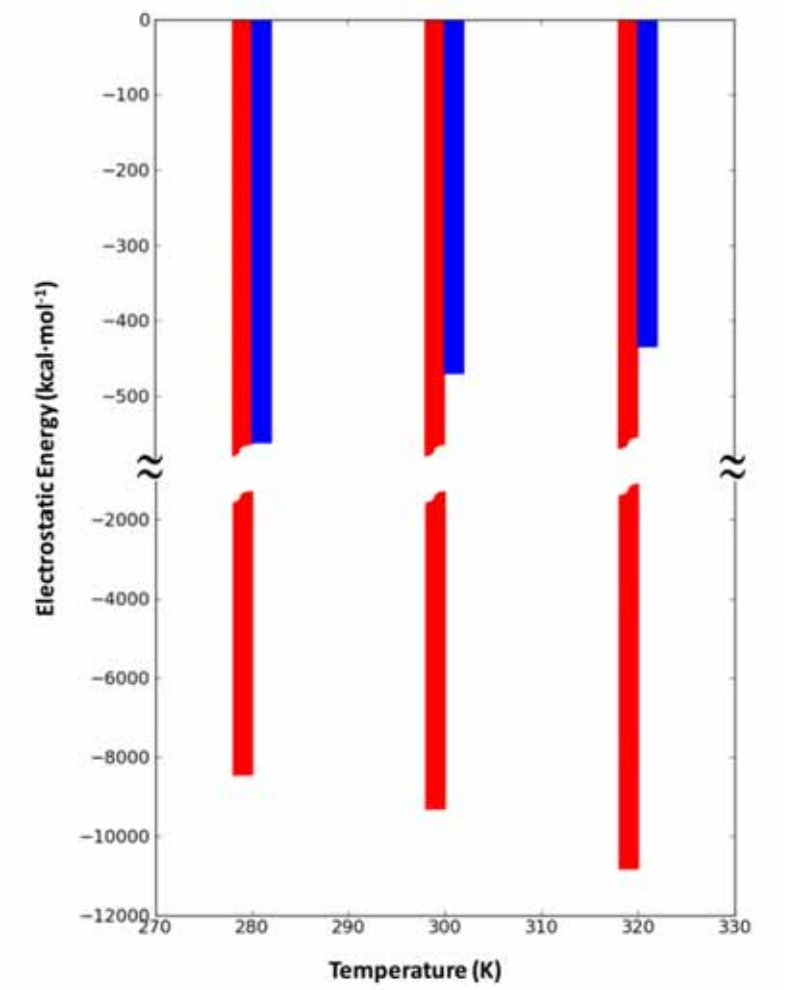


Figure S3. Average electrostatic energy over the 8 μs production-phase of BD simulations for the thermophilic (red bars) and mesophilic (blue bars) protein systems at 280, 300 and 320 K. The y-axis is broken for better visualization of the different trends obtained for both protein systems.

Table S1. $\langle u^2 \rangle$ values (with uncertainties) computed for the thermophilic and mesophilic proteins considering internal dynamics and the diffusion contribution.

	Mesophilic		Thermophilic	
Temperature	Internal $\langle u^2 \rangle$	Internal+Diffusion $\langle u^2 \rangle$	Internal $\langle u^2 \rangle$	Internal+Diffusion $\langle u^2 \rangle$
280	0.795±0.003	1.714±0.068	0.925±0.015	1.929±0.080
300	0.910±0.011	2.127±0.111	1.033±0.030	2.261±0.117
320	1.000±0.020	2.420±0.143	1.145±0.032	2.533±0.145
$d\langle u^2 \rangle/dT$	0.005±0.000	0.018±0.004	0.005±0.000	0.015±0.004

We have done the one-tailed Student's test on the hypothesis that $d\langle u^2 \rangle/dT$ (mesophilic) > $d\langle u^2 \rangle/dT$ (thermophilic). This hypothesis turns out to be true with a confidence of 70%. Despite such statistical confidence is not very large, the $d\langle u^2 \rangle/dT$ values already reflect a different thermal behavior that, when analysed in terms of the actual MSD as computed with Eq. 3, becomes more accentuated (to be published). This is due to the fact that instrumental resolution effects and approximations in the neutron scattering analysis reduce the actual MSD, as shown by Hayward and Smith [2]

Structural Data

The structures of the thermophilic (Malate Dehydrogenase from *Methanococcus Jannaschii*) and mesophilic (Lactate Dehydrogenase from pig muscle) proteins were obtained from the 1HYG [3] and 9LDT [4] entries in the Protein Data Bank, respectively. Both structures are tetrameric and consist of 1252 and 1324 residues respectively. In Zaccai's study [5], instead, the source of the mesophilic lactate dehydrogenase investigated in the experiments was *Oryctolagus cunniculus*, but as there is no crystallographic structure available for this protein they used the homologue (90% sequence identity) from pig-muscle to explain the structural basis of their finding. Because we need a crystallographic structure as a starting point of the simulations we have used the latter structure.

Brownian Dynamics simulations

We performed the Brownian Dynamics simulation (BD) with a code developed by Elcock and co-workers [6] in a previous work for the simulation of concentrated protein solutions. We have followed the same setup as Elcock and co-workers [7]. In short, atomic partial charges and protonation states of protein residues were determined with the PDB2PQR server [8] using PropKa [9]. This gives a total charge of -24e and 6e for the thermophilic and mesophilic proteins at pH=7.5, the same pH as in the neutron scattering experiment. For the calculation of electrostatic interactions we first computed protein electrostatic potentials. To this aim, APBS calculations [10,11] have been performed with a focussing grid of 4 Å and 2 Å. The ion concentration was set to 35 mM, the same as in the neutron scattering experiment [5] (considering KCl at 20 mM and the ion concentration from the added Tris-HCl solution at pH=7.5), with a radius of 2.0 Å. Subsequently, we used the SDA software to derive *effective charges* [12,13] from these electrostatic potentials to compute electrostatic interactions. BD simulations need infinite-dilution translational and rotational diffusion coefficients as input parameters and, as they have not been measured experimentally, these were calculated with the Hydropro software [14] using default parameters. The strength of hydrophobic interactions is described with the ε_{LJ} parameter of a Lennard-Jones potential and was set to 0.185 and 0.285 kcal/mol for the thermophilic and mesophilic proteins (see below for calibration procedure). The protein model used for BD simulations includes, in addition

to effective charges, all non-hydrogen atoms exposed to the solvent, which were determined with a 4 Å solvent probe using MMTK [15]. The APBS, SDA and Hydropro calculations have been performed independently at the three temperatures studied (280, 300 and 320 K). The dielectric constant of water has been set to 85.05, 77.70 and 70.94 at 280, 300 and 320 K respectively [16]. The water viscosity has been set to 1.47, 0.89 and 0.58 cP at 280, 300 and 320 K respectively [16].

Calibration of ε_{LJ} for Brownian dynamics simulations

The strength of hydrophobic interactions, ε_{LJ}, is the only adjustable parameter of the energy model. It is usually calibrated by fitting experimental data such as the second virial coefficient or the self-diffusion coefficient. Therefore, it implicitly corrects for deficiencies of the model in describing intermolecular interactions.

For the thermophilic protein, the only available experimental data suitable for ε_{LJ} calibration is the concentration dependence of the radius of gyration obtained by small-angle neutron scattering (SANS) [17]. Because the radius of gyration, Rg, measured by SANS arises from the radial distribution function, g(r), (see section below on SANS), which is an equilibrium property, this calibration guarantees that the energy model describes properly the system thermodynamics. BD simulations were run at two concentrations (3.3 and 10 mg/mL) and under the same experimental conditions (T=298.15K, pH=8.0, 100 mM KCl). For computing the SANS spectra of these simulations, it is worth pointing out that the integral in Eq. 2 to compute the structure factor, S(Q), is very sensitive to small deviations of g(r) at long distances due to the r dependence. For this reason, we have run very long simulations (50 μs) to achieve as much convergence as possible of g(r) at long distances and the integral was computed up to 350 Å. We used the form factor, P(Q), of a sphere with a radius fitting the experimental value of Rg extrapolated to zero concentration. The Q-range used was the same as in the experiment: (Rg_{app}·Q = 0.3-1.3). We tested different values of the ε_{LJ} parameter (0.150, 0.185 and 0.200 kcal/mol) and the corresponding results are plotted in Figure S4. The ε_{LJ} value of 0.185 kcal/mol was found to have the best agreement with the experimental value, which is close to previously reported ε_{LJ} values used for other proteins [6,7].

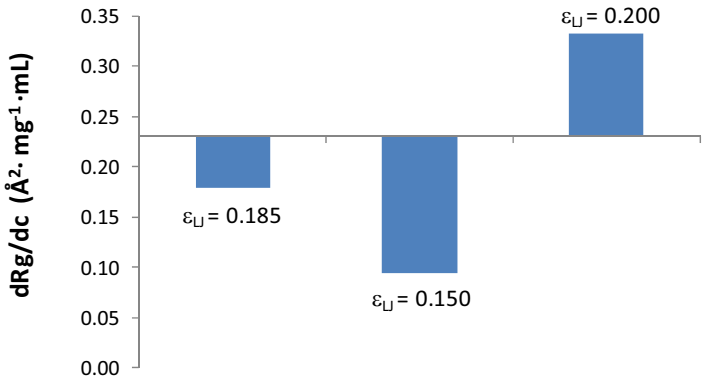


Figure S4. Representation of the dRg/dc values obtained from BD simulations testing different ε_{LJ} parameters. The horizontal line sets the experimental value.

We have also calibrated ε_{LJ} with the same experimental data at 343.15 K to test the temperature-dependence of this parameter. By running BD simulations as before, we found that $\varepsilon_{LJ,343}$ = 0.275 kcal/mol matches the experimental data pointing to a strong temperature-dependence of the parameter. To have an estimation of the values of ε_{LJ} that should be adopted at the three temperatures under study we have interpolated ε_{LJ} at 280, 300 and 320 K ($\varepsilon_{LJ,280K}$ = 0.135 kcal/mol; $\varepsilon_{LJ,300K}$ = 0.180 kcal/mol; $\varepsilon_{LJ,320K}$ = 0.225 kcal/mol). By running BD simulations of the crowded solution with these parameters, we observe that these new parameter values have mild changes in the diffusion coefficient and, thus, $\langle u^2 \rangle$. Thus we have decided to adopt the initial value of ε_{LJ} = 0.185 kcal/mol for the simulations at the three studied temperatures.

For the mesophilic protein, there is no experimental data available for calibration so that we have used the value of ε_{LJ} = 0.285 kcal/mol as in ref [7], which was shown to be adequate to reproduce the experimental self-diffusion coefficient of the Green Fluorescent Protein in *E. coli* cells. We have adopted the same parameter value at the three temperatures studied and show that fits remarkably well the temperature-dependence of $\langle u^2 \rangle$, as shown in the main text.

Small-Angle Neutron Scattering (SANS)

SANS provides information on the molecular weight, shape and intermolecular interactions in solution. The scattering intensity from small-angle neutron scattering experiments, I(Q), is given by the product of the structure factor, S(Q), the form factor, P(Q), and the protein concentration, ρ [18]:

$$I(Q) = \rho P(Q) S(Q) \tag{1}$$

The form factor depends on the shape of a single molecule. It can be obtained from analytical functions that have been developed for some geometrical shapes or from computer programs like CRYSON [19]. The structure factor, on the other hand, provides information on intermolecular interactions as is related to the radial distribution function, g(r), of intermolecular distances [18]:

$$S(Q) = 1 + 4\pi\rho \int_0^\infty (g(r) - 1) \frac{\sin(Qr)}{Q} r \, dr \tag{2}$$

where r is the inter-protein distance and Q the modulus of the scattering vector.

From the scattering intensity, I(Q), one can extract the effective radius of gyration (Rg) under the Guinier approximation provided that $Rg \cdot Q \leq 1$:

$$I(Q) = I(0) e^{-\frac{1}{3} R_g^2 Q^2} \tag{3}$$

The slope of a natural log plot of I(Q) *vs* Q^2 gives Rg. The radius of gyration obtained in this way provides an average measure of intermolecular interactions. When Rg increases with sample concentration, this is indicative of attractive intermolecular interactions.

The radial distribution function, g(r), can be easily obtained from Brownian dynamics simulations by averaging intermolecular distances over the total simulation length and number of molecules. For the calculation of S(Q) from g(r) the integration over long distances is troublesome because of the limited size of the simulation box and poorer statistics at long distances.

SUPPORTING REFERENCES

1. Holm L, Kaariainen S, Rosenstrom P, Schenkel A (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics* 24: 2780-2781.
2. Hayward JA, Smith JC (2002) Temperature dependence of protein dynamics: Computer simulation analysis of neutron scattering properties. *Biophys J* 82: 1216-1225.
3. Lee BI, Chang C, Cho SJ, Eom SH, Kim KK, Yu YG, Suh SW (2001) Crystal structure of the MJ0490 gene product of the hyperthermophilic archaebacterium Methanococcus jannaschii, a novel member of the lactate/malate family of dehydrogenases. *J Mol Biol* 307: 1351-1362.
4. Dunn CR, Wilks HM, Halsall DJ, Atkinson T, Clarke AR, Muirhead H, Holbrook JJ (1991) Design and synthesis of new enzymes based on the lactate-dehydrogenase framework. *Philos Trans R Soc Lond B Biol Sci* 332: 177-184.
5. Tehei M, Madern D, Franzetti B, Zaccai G (2005) Neutron scattering reveals the dynamic basis of protein adaptation to extreme temperature. *J Biol Chem* 280: 40974-40979.
6. McGuffee SR, Elcock AH (2006) Atomically detailed simulations of concentrated protein solutions: The effects of salt, pH, point mutations, and protein concentration in simulations of 1000-molecule systems. *J Am Chem Soc* 128: 12098-12110.
7. McGuffee SR, Elcock AH (2010) Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Comput Biol* 6: e1000694.
8. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res* 32: 665-667.
9. Li H, Robertson AD, Jensen JH (2005) Very fast empirical prediction and rationalization of protein pK(a) values. *Proteins: Struct, Func Bioinf* 61: 704-721.
10. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* 98: 10037-10041.
11. Fogolari F, Brigo A, Molinari H (2002) The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J Mol Recognit* 15: 377-392.
12. Gabdoulline RR, Wade RC (1996) Effective charges for macromolecules in solvent. *J Phys Chem* 100: 3868-3878.
13. Gabdoulline RR, Wade RC (1997) Simulation of the diffusional association of Barnase and Barstar. *Biophys J* 72: 1917-1929.

14. de la Torre JG, Huertas ML, Carrasco B (2000) Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophys J* 78: 719-730.
15. Hinsen K (2000) The molecular modeling toolkit: A new approach to molecular simulations. *J Comput Chem* 21: 79-85.
16. CRC Hanbook of Chemistry and Physics, 91st ed. Boca Raton: CRC Press.
17. Madern D, Ebel C, Dale HA, Lien T, Steen IH, Birkeland NK, Zaccai G (2001) Differences in the oligomeric states of the LDH-like L-MalDH from the hyperthermophilic archaea Methanococcus jannaschii and Archaeoglobus fulgidus. *Biochemistry* 40: 10310-10316.
18. Velev OD, Kaler EW, Lenhoff AM (1998) Protein interactions in solution characterized by light and neutron scattering: Comparison of lysozyme and chymotrypsinogen. *Biophys J* 75: 2682-2697.
19. Svergun DI, Richard S, Koch MHJ, Sayers Z, Kuprin S, Zaccai G (1998) Protein hydration in solution: Experimental observation by x-ray and neutron scattering. *Proc Natl Acad Sci U S A* 95: 2267-2272.

4.3.2. Intramolecular dynamics of the thermo-mesophilic pair of enzymes

After having analyzed the contributions to the overall flexibility as measured by neutron scattering, we aimed to identify differences in the internal dynamics linked to thermal adaptation. In the previous study we showed that the internal flexibility (MSD) of the thermophilic was higher than that of the mesophilic homologue, but the temperature-dependence of MSD was strikingly similar in the two proteins. Given that this analysis probed atomic motions at the 100 ps time scale, as in the experiment, here we explored a broader range of time scales.

Our first observation was that the higher flexibility of the thermophilic enzyme extends to the ns time scale at the three temperatures studied (280, 300 and 320 K). Of particular importance, however, is the fact that the temperature-dependence of MSD (dMSD/dT) increases with the time scale. Interestingly, such sensitivity of MSD to temperature changes is very similar in the two proteins at time scales below 300 ps, but it is at longer time scales that the dMSD/dT of the mesophilic increases more rapidly than that of the thermophilic homologue. This supports the fundamental idea originally proposed by Zaccai and co-workers regarding the higher temperature-dependence of the internal flexibility of the mesophilic enzyme, but at longer time scales.

We studied two possible origins for the different thermal behavior of the two proteins: structure and salt-bridge interactions. First, we note that the flexibility of residues located in coil regions (loops, turns or random coils) tends to be more temperature-dependent. Therefore, the less compact structure of the mesophilic enzyme and its higher content on loop residues makes the flexibility of the protein structure more temperature-dependent than in the thermophilic homologue. Second, salt-bridge interactions between oppositely charged residues strengthen with temperature due to a reduction of the desolvation penalty required for salt-bridge formation. We show that the larger proportion of charged residues (especially at the surface) of the thermophilic protein enables the formation of a larger number of interactions that are stabilized with temperature and thus provide more robustness to the structure at high temperature. This mechanism enables some regions of the thermophilic protein to reduce their flexibility upon increasing temperature.

Besides the different temperature-dependence of MSD between the two proteins, we note that the distribution of residue mobilities is significantly different. The mesophilic enzyme exhibits a broader distribution of mobilities relative to the thermophilic homologue, which means a higher dynamical heterogeneity. The reason for this lies also in the lower packing of the structure. In relation to this, we show that the main conformational fluctuations of the mesophilic protein have a more local character, whereas the motions in the thermophilic homologue tend to be more collective. As a result of this, we suggest that the motions of the thermophilic protein emerge from broader energy landscapes.

A more detailed presentation of these results is given in the following article: *Dynamic fingerprints of protein thermostability revealed by long molecular dynamics*. Submitted to *J. Chem. Theory Comput.*

Dynamic fingerprints of protein thermostability revealed by long molecular dynamics

Enrique Marcos, Aurora Jiménez and Ramon Crehuet*

Department of Biological Chemistry and Molecular Modelling, Institute of Advanced Chemistry of Catalonia (IQAC - CSIC), E-08034 Barcelona, Spain

INTRODUCTION

Enzymes achieve their outstanding efficiency under very specific environmental conditions. Factors like temperature, pressure or salt concentration have a tremendous impact on enzyme activity, but the broad versatility of evolution allows finding enzymes adapted to extremely diverse environments. Of particular interest are thermophilic and hyperthermophilic proteins from organisms that grow at very high temperatures ranging from 50 to 120ºC. Such adaptation to high temperatures requires very resistant proteins to heat denaturation and elucidating the origin of this resistance will provide clues for designing proteins with enhanced thermal stability for a wide range of applications.

There has been a surge of experimental and computational studies comparing thermophilic proteins with their corresponding homologs working at room temperature, known as mesophilic. From studies in the last 20 years, it seems that there is not a unique strategy adopted by evolution to thermostabilize proteins. Structural and amino acid sequence comparisons between a vast range of thermophilic and mesophilic proteins point to some features that correlate with increased thermostability, namely larger proportion of charged residues, more compact hydrophobic core and shortened loops at the surface [1-3]. The dynamical requirements for protein thermostability, however, are more controversial. Since thermophilic proteins unfold at higher temperature and are less active than their mesophilic homologs at lower temperature, thermophilic proteins have been traditionally considered more rigid. According to the *corresponding states* hypothesis,

thermophilic and mesophilic proteins achieve similar flexibility at their respective temperatures for maximum activity. Nevertheless, experimental and simulation techniques able to explore atomic motions at different time scales have indicated that the panorama is more complex and that there is not a unique strategy for protein thermostability. Some of these studies found thermophilic proteins to be more rigid than their mesophilic homologs [4-8], whereas others showed the opposite [9-14]. This lack of consensus stems from the absence of a unique mechanism of thermostability and, on the other hand, from the fact that these techniques explore different aspects of protein dynamics among the vast diversity of dynamic events that occur in a broad range of time scales. Enhanced flexibility in thermophilic proteins [9, 11] can entail an increase in conformational entropy of the native state that lowers the entropy change upon unfolding providing more stability [15]. On the other hand, NMR relaxation experiments [8] show that large-amplitude motions in a thermophilic adenylate kinase do not occur as frequently as in a mesophilic homolog at low temperature, which supports the corresponding states hypothesis. This is a clear indication that a proper definition of flexibility requires the specification of time scale and type of motion. For this reason, the flexibility regarding different dynamic events is not directly comparable and, ultimately, their linkage to stability and function does not have to be necessarily the same.

One of the most intriguing dynamical mechanisms of protein thermostability was put forward few years ago by Zaccai and co-workers based on elastic neutron scattering experiments [11]. They observed that the mean-square-displacement (MSD) at short time scales (~100 ps) of a thermophilic enzyme was less sensitive to temperature changes than the corresponding mesophilic homolog. These results motivated the authors to suggest that this can be a plausible mechanism for thermophilic proteins to control the structural fluctuations at high temperature to avoid unfolding. In a recent work [16], however, we have shown that such amazing difference in the thermal behavior of MSD between the two proteins, indeed, arises to a large extent from significant differences in diffusion under crowding conditions. After having rationalized the mechanism behind the experimental observation, in this work we aim to explore a broader range of time scales than in the experiment and characterize fundamental differences in the intramolecular dynamics of both proteins by means of molecular dynamics simulations (MD). In particular, we have focused on understanding the factors that determine a different sensitivity of intramolecular protein motions to temperature changes, which is strongly

linked to thermal adaptation. We have performed 200 ns simulations of each protein at three different temperatures (280, 300 and 320 K), adding to a total simulation time of 1.2 μs. This exceeds the time scales explored in previous MD studies comparing other themo-mesophilic pairs of proteins.

METHODS

Structural data

The homotetrameric structures of the thermophilic enzyme (malate dehydrogenase from the *Methanococcus Jannaschii* archaea) and its mesophilic homolog (lactate dehydrogenase from pig muscle) were obtained from the 1HYG [17] and 9LDT [18] entries in the Protein Data Bank, respectively. Both structures have 313 and 331 residues per chain, respectively, and have a 28% sequence identity.

Molecular Dynamics (MD)

All MD simulations were performed with the Gromacs simulation package (version 4.0.5) [19]. Each protein was simulated under periodic boundary conditions using a rhombic dodecahedral water box. Standard protonation states were assigned to all protein residues, being -16 and 8 electronic units the total charge of the thermophilic and mesophilic proteins respectively. The electric neutrality of the systems was achieved by adding Na^+ counterions to the thermophilic protein system and Cl^- counterions to the mesophilic one. Substrates bound to the crystallographic structures were removed. The total numbers of atoms in the simulations of the thermophilic and mesophilic proteins is 89229 and 90393 respectively.

The two solvated systems under study were subjected to 200 ns simulations at temperatures set to 280, 300 and 320 K. We used the OPLS all-atom force-field [20] for describing protein interactions and TIP3P [21] to model water molecules. Short-range electrostatic interactions were calculated explicitly with a 10 Å cutoff and long-range electrostatic interactions were computed via the Particle Mesh Ewald method [22] using a grid spacing of 1.2 Å and a 4^{th} order spline interpolation. Lennard-Jones interactions were calculated using a switch function between 0.8 and 0.9 nm. An integration time step of 2

fs was used by constraining all bonds with LINCS [23]. The temperature was controlled with the Berendsen thermostat [24] using a coupling time constant of 0.1 ps. The pressure was kept constant at 1 bar via the Berendsen barostat [24] using an isotropic compressibility of $4.5 \cdot 10^{-5}$ bar^{-1} and a coupling constant of 0.5 ps. Before production runs were started, the structure of the solvated protein was energy minimized with the steepest descent algorithm. Next, solvent surrounding the protein was equilibrated by running a MD simulation at the target temperature using harmonic position restraints on the heavy atoms of the protein with a force constant of 1000 kJ mol^{-1} nm^{-2}. Finally, a 200 ns trajectory was carried out and the last 160 ns were used for production. Snapshots were saved every 20 ps for subsequent analysis. In this work, we have only considered intramolecular protein motions. Translational and rotational diffusive motions were removed by superimposing every frame into the first frame.

Time scale dependence of protein dynamics

The time-dependent mean-square-displacement, MSD (Δt), was calculated from time-trajectories as:

$$\mathrm{MSD}(\Delta t) = \frac{1}{N(T-\Delta t+1)} \sum_{i}^{N} \sum_{t=0}^{T-\Delta t} \left[\mathbf{R}_i(t+\Delta t) - \mathbf{R}_i(t) \right]^2 \qquad (1)$$

where $\mathbf{R}_i$ is the position vector of atom i, N the number of atoms, T the time length of the trajectory and Δt the time separation between saved frames.

Principal Component Analysis (PCA)

It is possible to describe the conformational fluctuations in terms of collective variables that concentrate the most important dynamic information from the trajectory and filter out the noise from irrelevant local motions. This can be done by doing a principal component analysis (PCA) of MD trajectories, also known as *essential dynamics* [25]. First, the covariance matrix from the fluctuations of atomic positions is built as:

$$C_{ij} = \left\langle \left(q_i - \langle q_i \rangle \right) \left(q_j - \langle q_j \rangle \right) \right\rangle \qquad (2)$$

where q_k is the *k* component of vector $\mathbf{q} = \{q_1, ..., q_{3N}\}$ which defines the coordinates of the system of *N* atoms. $\mathbf{C}$ is a symmetric 3*N* x 3*N* matrix, whose diagonal elements represent the atomic mean-square-fluctuations and the off-diagonal elements the correlation between two variables. The eigenvectors of $\mathbf{C}$ are 3*N*-dimensional vectors that indicate the direction of motion of the principal components or *essential modes* and the corresponding eigenvalues correspond to the mean-square-fluctuations associated to the mode. Each eigenvector defines the direction of motion as a displacement from the average structure. The trajectory can be projected into a principal mode *k* ($\mathbf{v}_k$) as: $p_k(t) = \mathbf{v}_k^t \cdot (\mathbf{q}(t) - \langle \mathbf{q} \rangle)$. Such analysis is particularly useful for characterizing conformational transitions with different amplitudes occurring in the course of the simulation.

The projection $p_k(t)$ has short-term fluctuations around the global trend determined by the direction of principal component *k*. To quantify the amplitude of these fluctuations, we have first determined the trend of the projection at each time t by calculating a moving average $\overline{p}_k(t)$, which have been calculated as:

$$\overline{p}_k(t) = \frac{1}{2n} \sum_{m=t-n}^{t+n} p_k(m) \qquad (3)$$

where $\overline{p}_k(t)$ is an average over all frames between t+*n* and t-*n*. We have set *n* to 227 points, which implies taking the average over 4.54 ns. Subsequently we have calculated the deviation of $p_k(t)$ from the trend $\overline{p}_k(t)$ as a variance (σ^2):

$$\sigma^2 = \frac{1}{T-2n} \sum_{t=n}^{T-n} [p_k(t) - \overline{p}_k(t)]^2 \qquad (4)$$

where *T* is the total number of frames

Calculation of degree of packing

To describe the packing density at each protein residue, we have adopted the definition of Chennubhotla and Bahar [26] originally introduced in a Markov model of

communication among residues. The interaction between pairs of residues is defined with an affinity matrix, **A**, whose A_{ij} elements are calculated as:

$$a_{ij} = \frac{N_{ij}}{\sqrt{N_i N_j}} \tag{5}$$

where N_{ij} is the number of atom-atom contacts between residues i and j within a cutoff distance of 4 Å and N_i,N_j are the number of heavy atoms of both residues. From the affinity matrix it is defined the degree diagonal matrix $\mathbf{D} = \{d_i\}$ where $d_i = \sum_{j=1}^{N} a_{ij}$ and reflects the packing density at each residue.

Collectivity of modes of motion

The collectivity index [27] of a given mode of motion, e.g. normal mode or principal component, gives a measure of how many atoms contribute to this mode. We have used this index to characterize the dynamical heterogeneity associated to the principal components. The collectivity (k_i) of a given mode $\mathbf{r}_i$ is defined as:

$$k_i = \frac{1}{N} \exp\left\{ -\sum_{n=1}^{N} \mathrm{u}_{i,n}^2 \log \mathrm{u}_{i,n}^2 \right\} \tag{6}$$

where i is the mode index, n run over N atoms and

$$\mathrm{u}_{i,n}^2 = \alpha \frac{\left|\mathbf{r}_{i,n}\right|^2}{m_n} \tag{7}$$

where m_n is the mass of atom n and α is a normalization constant to give $\sum_{n=1}^{N} \mathrm{u}_{i,n}^2 = 1$. The collectivity tends to decrease for local motions (high-frequency modes), which means that a lower number of atoms contribute to the motion described by the mode. Highly collective or *cooperative* motions, on the other hand, are those in which a large proportion of atoms participate.

RESULTS AND DISCUSSION

Overall flexibility at different time scales and temperatures

First, we have explored how the thermophilic and mesophilic proteins differ in both the time scale- and temperature-dependence of their conformational fluctuations. In the present investigation, this has been determined with the mean-square-displacement (MSD), as given by Eq. 1, at time scales ranging from 20 ps to 160 ns at temperatures set to 280, 300 and 320 K. The results are summarized in Figure 1. Panel a shows the MSDs taking into account the full trajectories, whereas panel b focuses on time scales up to 50 ns including an approximation to the error. The errors of each curve have been determined with the bootstrapping method by calculating the MSD of 100 randomly generated 50 ns-chunks of each 160 ns trajectory.

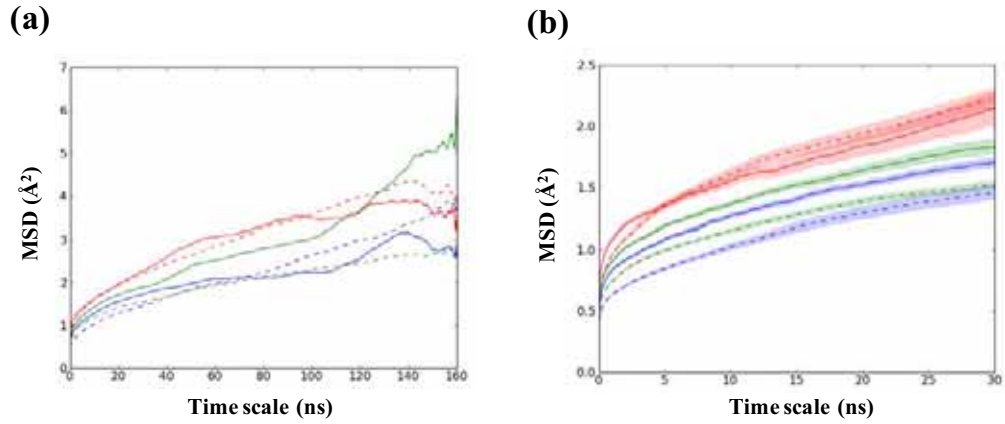


Figure 1. Mean square displacements of the thermophilic (solid lines) and mesophilic (dashed lines) over the 0-160 ns time scales range at three different temperatures: 280 K (blue), 300 K (green), 320 K (red). Panel b shows the 0-30 ns range of panel A. MSD values are averaged over all backbone atoms. We have computed the errors bars in panel b with the bootstrapping statistical method, which implies computing the MSD from randomly generated chunks of the trajectory.

At time scales below 5 ns (see Figure 1b), the MSD values of the thermophilic protein are above the mesophilic one at the three temperatures studied. In other words, the thermophilic protein is more flexible than the mesophilic one at short time scales. This shows that the observation in our previous theoretical study [16] that the thermophilic

protein is more mobile than the mesophilic one at the 100 ps time scale also applies to a broader range of time scales. Interestingly, at the highest temperature (320 K) and beyond the 5 ns time scale, the MSD of the mesophilic protein approaches that of the thermophilic protein. Thus the increase in flexibility with temperature and time scale differs between the two proteins. Nevertheless, for longer time scales it is not possible to ascertain if the mesophilic protein turns out to be even more flexible than the thermophilic one, since a 160 ns trajectory does not provide enough statistics to characterize atomic motions beyond the 10-20 ns time scale. This could be done by lengthening the simulation an order of magnitude at least.

The fact that at 320 K the mesophilic protein achieves similar flexibility to that of the thermophilic protein at the ~5 ns time scale is indicative of a different sensitivity to temperature changes in the two proteins. Figure 2 shows the temperature-dependence of MSD (dMSD/dT) at different time scales. The dMSD/dT quantity has been determined by taking the slope of a linear fit of MSD versus temperature at each time scale. Figure 2 clearly illustrates how the temperature-dependence of MSD increases with the time scale in both proteins, being this increase of dMSD/dT more accentuated in the mesophilic protein. At time scales ranging from 20 ps to 100-200 ps, dMSD/dT is very similar in the two proteins, but it is beyond the 300 ps time scale that the difference in dMSD/dT is manifested. The key message conveyed by this figure is that experimental techniques probing atomic motions at different time scales would show a different thermal behavior of intramolecular dynamics, which is a key aspect for studying the basis of protein thermostability. It is worth mentioning that dMSD/dT is a quantity routinely determined by elastic neutron scattering experiments at the picoseconds time scale and thus information on the dynamics at longer time scales would be an ideal complement.

The lower tendency of the thermophilic enzyme of increasing its mobility with temperature can have important implications for thermostability, since this can protect the structure from undergoing conformational motions related to unfolding events. Zaccai and co-workers based on neutron scattering experiments [11] at the 100 ps time scale had already proposed this idea as the dynamical basis underlying the higher thermal stability of thermophilic proteins. Although in our previous study [16] we showed that a contribution of diffusion to the observed overall dynamics plays a role in interpreting their data, here we support the fundamental essence of their idea. Below we address some

of the reasons of such difference in dMSD/dT between the two proteins. It is intriguing the fact that the different dMSD/dT observed for both proteins at the ps-ns time scales apparently indicates a fingerprint of thermostability, taking into account that the unfolding process, which is a rare event, takes place at much longer time scales.

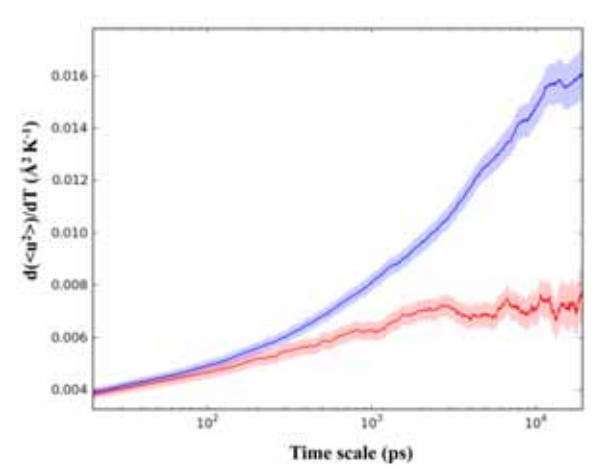


Figure 2. Representation of the temperature-dependence of MSD from backbone atoms of the thermophilic (red) and mesophilic proteins (blue) at different time scales (from 20 ps to 10 ns). We have computed the errors bars with the bootstrapping statistical method, which implies computing the MSD from randomly generated chunks of the trajectory.

Dynamical heterogeneity

Does the higher flexibility of the thermophilic protein apply to all residues or some specific regions? Of course, the protein surface tends to be more mobile than those regions buried at the protein core, which is subjected to larger spatial constraints. This makes the protein structure dynamically heterogeneous. Is the dynamical heterogeneity of the thermophilic protein different to that of the mesophilic homolog? To gain insights into this matter, now we turn our attention to the contribution of residues to the overall flexibility.

Figure 3 shows the MSD of each residue (averaged over the four chains) of the thermophilic and mesophilic proteins (panel a). The most flexible parts of the two proteins correspond to surface loops, turns connecting secondary structure elements and

chain extremes, whereas the most rigid residues are located at the protein core. In particular, the N and C termini exhibit much higher mobilities in the mesophilic protein (see arrows in Figure 3a) in support to the general observation that chain extremes are better anchored in thermophilic homologs [1]. Figures 3c and 3d illustrate this heterogeneous distribution of residue mobilities with a color-coded diagram of both protein structures. Despite the lower overall flexibility of the mesophilic protein, some of its regions of the former display very high mobilities not accessible by the thermophilic homolog, which is indicative of different distributions of mobilities as will be shown below.

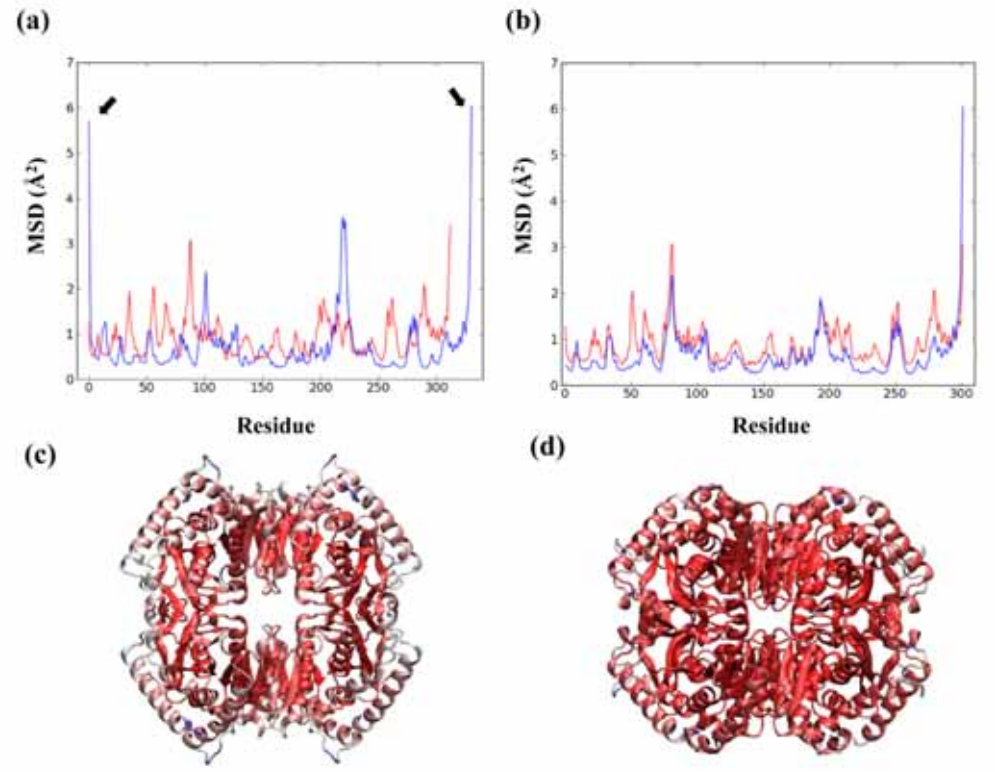


Figure 3. Residue mobilities at 280 K and at the 1 ns time scale. (a) MSD of each residue of the thermophilic and mesophilic proteins respectively. (b) MSD of structurally aligned residues. (c,d) ribbon diagrams that are color-coded according to the MSD value of each residue of the thermophilic and mesophilic proteins respectively. From red to blue the MSD increases. The MSD of each residue has been computed as an average over the MSD of all backbone atoms of the same residue.

To compare the flexibility of structurally analogous regions of the two proteins, we have structurally aligned both protein structures with DALI [28]. Figure 3b shows the MSD of aligned residues. The mobility of the thermophilic protein is higher in all aligned residues except for the C terminal extreme. Note that the whole N terminal part of the mesophilic protein and the loop defined by residues 217-225, which are very flexible (see Figure 3a), have not been aligned and, thus, do not appear as high maxima in Figure 3b. It is remarkable that most rigid regions of the thermophilic protein (local minima in the MSD profile) exhibit a higher mobility than the most rigid residues of the mesophilic homolog. Figure 4a clearly illustrates the broader distribution of MSD values in the mesophilic protein (blue curve) with respect to the thermophilic one (red curve). The inset of Figure 4a shows that very high MSD values are only adopted by the mesophilic protein.

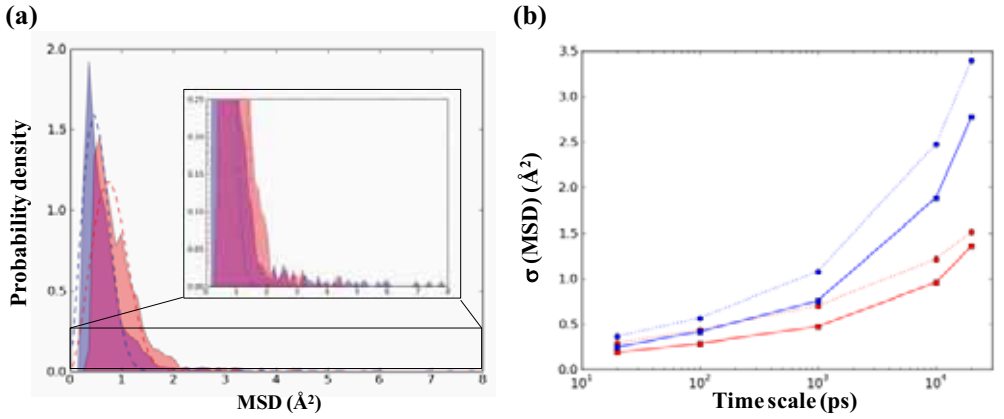


Figure 4. Dynamical heterogeneity of the thermophilic (red) and mesophilic (blue) proteins. (a) Normalized histograms of residue mobility at 1 ns and 280 K. The area common to both proteins is colored in violet to highlight the differences in their distributions. The inset shows the population of the highest MSD values. Dashed lines correspond to Weibull fits. (b) Standard deviation (σ) of MSD residues at different time scales and temperatures (280 K, solid lines; 320 K, dotted lines).

The different dynamical heterogeneity of both proteins has consequences in the interpretation of neutron scattering data. In particular, elastic incoherent neutron scattering (EINS) provides a measure of the MSD of the sample by using the gaussian approximation, which assumes that all atomic displacements are equal, ignoring dynamical heterogeneity [29]. Therefore, when comparing the dynamics of two proteins,

their different dynamical heterogeneity will introduce different errors in the MSD obtained from the gaussian approximation. Smith and co-workers [30] suggested a solution to this problem by modeling a Weibull distribution of atomic displacements when comparing the dynamics of a thermo-mesophilic pair of dihydrofolate reductases. By performing EINS experiments they found a broader distribution of MSD mobilities in the thermophilic protein, in contrast to our results. To gain insight into this discrepancy, we have fitted a Weibull function to each distribution (see dashed lines in Figure 4a) and, subsequently, calculated the standard deviation (σ) associated to the fitted functions (see Supporting Information for details in this calculation). This analysis reveals that the σ of the fit to the mesophile's distribution is now lower to that of the thermophile (0.24 vs 0.32 $Å^2$), which is in contrast to the results obtained with the original data (0.75 vs 0.47 $Å^2$). Note in Figure 4a the larger width of the fitted function of the thermophile. This apparent contradiction is reconciled by the fact that the tail of the Weibull fit to the mesophile's distribution does not include the most flexible residues, which contribute to increase σ substantially. The inability of the Weibull function to include the most flexible residues therefore masks the heterogeneous character of the distribution. Thus we can argue that our observation of a higher dynamical heterogeneity in the mesophilic protein is fully consistent with the results obtained by Smith and co-workers [29].

We have quantified this dispersion of MSD values with the standard deviation (σ) over all residues at different time scales and temperatures (see Figure 4b). The figure shows that σ increases with time scale and temperature in both proteins, but this increase turns out to be more accentuated for the mesophilic protein. Figure 4b also shows that σ of the mesophilic protein is larger than that of the thermophilic one in the range of time scales studied and, in addition, increases with temperature to a higher extent. We wonder whether the dispersion in residue mobilities can be regarded as a dynamic fingerprint distinguishing thermophilic proteins from mesophilic ones. If so, a MD simulation at a single temperature would find wide use in distinguishing proteins with different thermal adaptation.

How correlated is the dynamics at different time scales and temperatures?

Given that some experimental techniques only have access to short time scales, such as neutron scattering, one wonders whether the conformational dynamics at short time

scales (ps) correlate with that at longer time scales (ns or beyond), which involve different types of motion. We have already seen that the overall flexibility increases with time scale and temperature. Do all residues increase their mobilities with time scale and temperature at the same rate? How do residue mobilities correlate between different time scales and temperatures? In Figure 5 we have analyzed these issues. Figures 5a and S1 show that the shape of the MSD profile is well conserved among different time scales. Figure 5b illustrates this good correlation by comparing the MSDs between two different time scales differing in three orders of magnitude (20 ps and 10 ns). Such smooth increase in flexibility with time-scale, however, hinders the fact that the increase in flexibility is not the same for all residues. For example, at 20ps regions 'a' and 'b' (see arrows in Figure 5a) have similar flexibilities, whereas at much longer time scales (10 ns) region 'b' becomes far more flexible than 'a'. This can be clearly seen when normalizing the flexibilities at different time-scales (see Figure S2). Note that the thermophilic protein has a more conserved pattern than the mesophilic one. This implies that, in the mesophilic protein, the rates of MSD increase with time scale are more residue-dependent or *heterogeneous* throughout the structure. This can be viewed as an alternative manifestation of the higher dynamical heterogeneity of the mesophilic protein. It is thus expected that these more heterogeneous rates of MSD increase in the mesophilic protein result in distributions of residues flexibilities with a steeper increase of σ upon raising the time scale (Figure 4b). The conserved pattern of the thermophile can be explained by its higher compactness and collectivity of protein motions, as will be described below. Therefore, although the general trend is that flexible regions remain flexible and fixed regions remain fixed, the relative flexibilities among residues at different time scales does change. This has also been reported from NMR experiments [7] that compared two different time-scales. Again, this emphasizes the crux of this paper: that talking about dynamics without specifying a time scale can result in contradictory conclusions.

Figure 5b also illustrates how the MSD of the mesophilic protein (in blue) at 320 K and a time scale of 10 ns exhibits a sharper increase, with respect to 20 ps, than the thermophilic homolog. The sharper increase in MSD of the most flexible residues of the mesophilic protein compensates the lower mobility of its more rigid residues resulting in an overall flexibility fairly similar to that of the thermophilic protein. This is a complementary perspective of the observation from Figure 1 that the overall MSD of the mesophilic protein at this temperature increases more rapidly with time scale to the extent

that at ~5 ns reaches the MSD values of the thermophilic one. The corresponding representations at the other temperatures are shown in Figure S3.

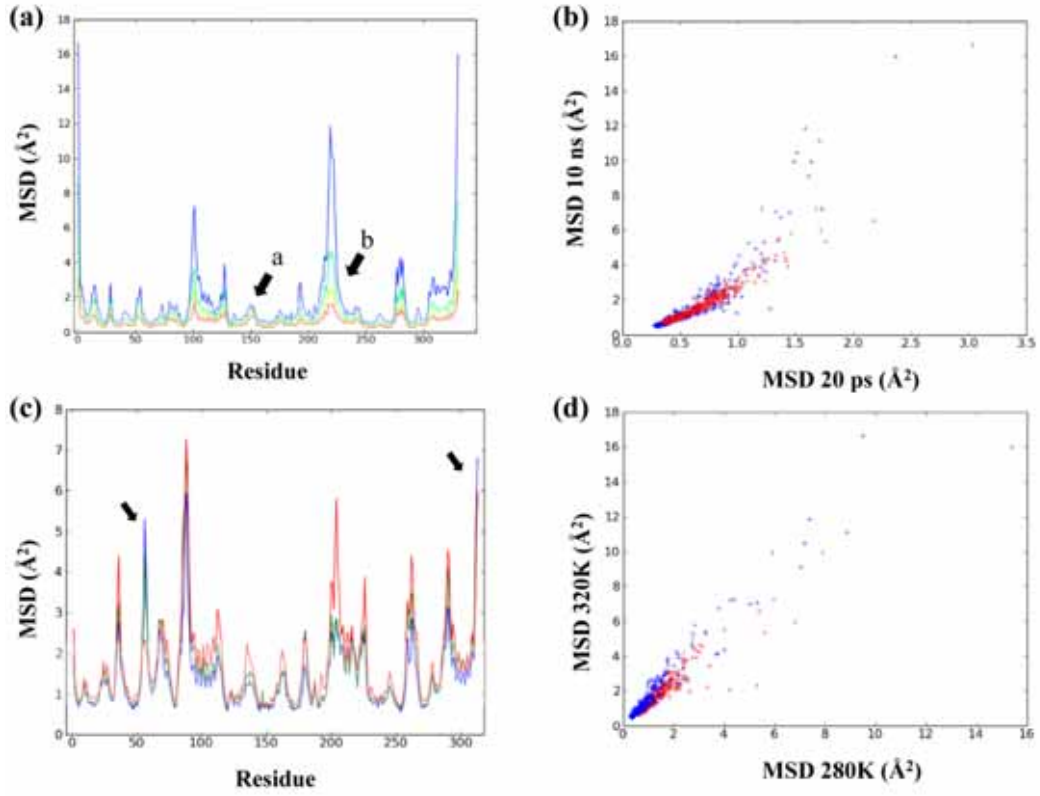


Figure 5. Flexibility at different time scales and temperatures. (a) MSD of the mesophilic protein at 320 K and time scales: 20 ps (red), 100 ps (yellow), 1 ns (green) and 10 ns (blue). (b) Correlation between the MSDs at 20 ps and 10 ns of the thermophilic (red crosses) and mesophilic (blue crosses) proteins at 320 K. (c) MSD (10ns) of the thermophilic protein at 280 (blue), 300 (green) and 320 K (red). (d) Correlation between the MSDs (10 ns) at 280 and 320 K for the thermophilic (red crosses) and mesophilic (blue crosses) proteins.

In Figure 2 we have given support to the thermostability mechanism of thermophilic proteins suggested by Zaccai and co-workers [11], which is based on a reduced temperature-dependence of the overall flexibility. To gain insights into the origin of our results, here we aim to deconvolute the contribution of residues to this overall behavior of the protein. With regard to the MSD variation with temperature, we have compared the

residue flexibilities between different temperatures at the time scale of 10 ns. We note in Figures 5C and S4 that all residues of both proteins, except for residues 50-58 and the C-terminal tail of the thermophilic protein (see arrows in Fig. 5C), tend to increase the MSD with temperature. Figure 5d shows that this rate of increase is fairly similar among all residues and that the slope of this linear trend is higher for the mesophilic protein in line with Figure 2, which provides evidence of its larger dMSD/dT. This figure shows that all residues contribute to increase dMSD/dT. To better illustrate the temperature-dependence of the MSD of each residue we have represented this quantity in Figure 6 as the slope of a linear fit of MSD (at 10 ns) *vs* temperature. Now we turn our attention to the factors that determine differences in the dMSD/dT of residues in the two proteins.

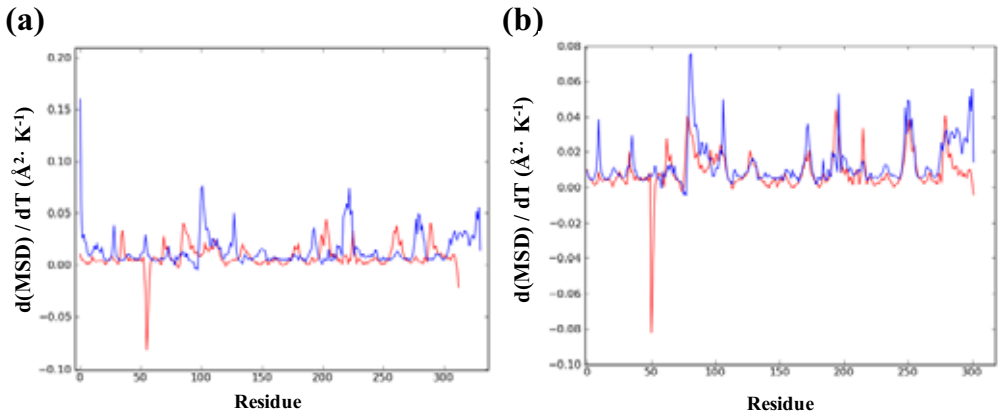


Figure 6. (a) Representation of the dMSD/dT values computed for the backbone atoms of each residue of the thermophilic (red) and mesophilic (blue) proteins at the 10 ns time scales. (b) Same representation as panel a, but for structurally aligned residues.

Structural differences between the mesophilic and thermophilic homologs

Figure 6b shows the dMSD/dT values shown in Figure 6a, but only for structurally aligned residues. Figure 6b shows that the dMSD/dT values of the mesophilic protein are, in general, above those of the thermophilic homolog. As noted above, even some residues of the thermophilic protein (50-58 and the C-terminal tail) undergo a decrease in MSD with temperature, in contrast to the mesophilic protein, the residues of which all have a positive dMSD/dT. However, it is worth emphasizing that some regions display very close similarity between the two proteins. To gain insights into the origin of these

similarities and differences in the dMSD/dT of both proteins, it is instructive to evaluate different contributions to the average dMSD/dT as shown in Table 1. Herein those residues that do not form helices or β-sheets are termed coils, including loops, turns or random coils.

Table 1. Contributions to the temperature-dependence of MSD at the 10 ns time scale

	Thermophilic	Mesophilic
$\langle dMSD/dT \rangle^{a}$	0.0081	0.0147
$\langle dMSD/dT \rangle_{align}{}^{b}$	0.0082	0.0130
$\langle dMSD/dT \rangle_{notalign}{}^{c}$	0.0069	0.0248
$\langle dMSD/dT \rangle_{(L)}{}^{d}$	0.0082	0.0202
$\langle dMSD/dT \rangle_{(H/E)}{}^{e}$	0.0081	0.0114
% of aligned residues forming coils	34.8	39.4
% of aligned residues forming helices/sheets	65.2	60.6
% of not aligned residues forming coils	63.6	72.4
% of not aligned residues forming helices/sheets	36.4	27.6

[a] Average over all residues

[b] Average over structurally aligned residues

[c] Average over not structurally aligned residues

[d] Average over residues forming coils (L)

[e] Average over residues forming helices (H) or sheets (E)

The average of dMSD/dT over structurally aligned residues, $\langle dMSD/dT \rangle_{align}$, is still higher for the mesophilic protein (0.0130 *vs* 0.0082 $Å^2 \cdot K^{-1}$). However, it is interesting to note that the difference between both proteins is enlarged when comparing the average over not aligned residues, $\langle dMSD/dT \rangle_{notalign}$, (0.0248 *vs* 0.0069 $Å^2 \cdot K^{-1}$). This separation of the contributions from aligned and not aligned residues reveals that the flexibility of structural features unique to the mesophilic protein is much more sensitive to changes in temperature. In particular, a 72.4% of not aligned residues of the mesophilic protein form

coils. In line with this, we show in Table 1 that the dMSD/dT of all coil regions, $\langle dMSD/dT \rangle_{(L)}$, of the mesophilic protein is notably larger than that of helices and β-sheets, $\langle dMSD/dT \rangle_{(H/E)}$, (0.0202 *vs* 0.0114 $Å^2 \cdot K^{-1}$). In contrast to this, the dMSD/dT of the thermophilic protein proves to be practically insensitive to the secondary structure. Thus the higher dMSD/dT of the mesophilic protein has an important structural reason. Indeed, it is widely accepted that thermophilic proteins tend to exhibit shortened surface loops to lower the entropy change of unfolding [31], which provides higher stability. Here we show that this structural feature has implications on the sensitivity of protein flexibility to temperature changes.

A related structural feature that has been associated to enhanced thermostability is an improved packing [32, 33]. To explore further this idea we have quantified the compactness of both proteins with a measure of packing, i.e. density at residue (d_i), introduced by Chennubhotla and Bahar (see Methods) that averages the number of contacts at each residue. We have averaged the density of all residues ($1/N \sum_{i=1}^{N} d_i$) in the crystal structures of both proteins. It turns out that the degree of packing is higher in the thermophilic protein (6.8 *vs* 6.3). To characterize the distribution of packing across the protein structure, Figure 7 displays the density averaged over residues at different shells of the protein structure. We have divided the protein structure in five shells (s_i, i=1-5) as a function of distance to the center of mass with a 10 Å-width. For instance, s_1 refers to residues at a distance ranging from 0 to 10 Å from the center of mass. It is worth reminding the reader that in the first 10 Å shell of the thermophilic protein there are no residues due to the large central hole (see Figure 3c).

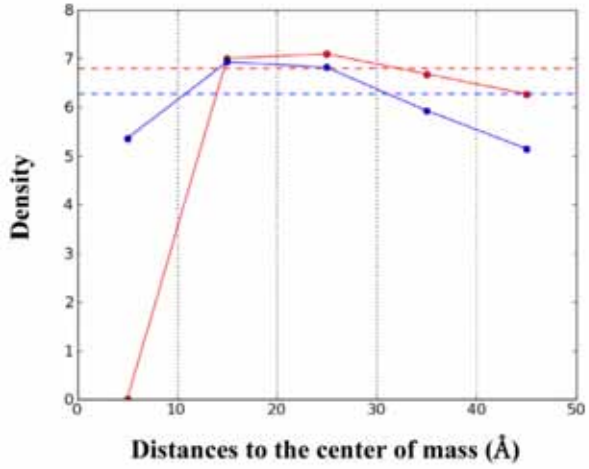


Figure 7. Density of residues at different shells of the crystal structures of the thermophilic (red) and mesophilic (blue) proteins. Horizontal dashed lines represent the average of each data series over the five shells considered. The last shell corresponds to distances greater than 45 Å.

Figure 7 illustrates that the structure of the thermophilic protein is more compact than the mesophilic homolog near the surface. This is consistent with the larger content of residues forming coils, which have limited intra-protein contacts, at the surface of the mesophilic protein. In addition, it is pertinent to note that the higher packing of the thermophilic protein does not necessarily mean less flexibility. On the contrary, it turns out to be more flexible, at least at the time scales here explored. What is even more interesting is the fact that the higher density of the thermophilic protein structure correlates with its higher dynamical homogeneity. Indeed, the more densely packed is the structure the more efficient is expected to be the propagation of thermal fluctuations. The low number of intra-protein contacts of longer coils at the surface of the mesophilic protein enables large conformational fluctuations in these regions, but at the same time is expected to hamper the propagation of such fluctuations to the rest of the structure. This ultimately increases the gap between the most flexible and rigid residues of the protein leading to a higher dynamical heterogeneity.

A structural index that is consistent with the higher flexibility of the thermophilic protein is the Lindemann coefficient (Δ_L) [34], which estimates the solid-liquid behaviour of a

protein. The higher Δ_L is, the more *liquid-like* the protein structure is. In this case, Δ_L has been calculated with FlexServ [35] and values of 0.131 and 0.118 have been obtained for the thermophilic and mesophilic proteins respectively. Thus the fact that the thermophilic protein turns out to be more *liquid-like* is consistent with its higher flexibility.

Salt bridge interactions

It is important to remark that the flexibility of structurally analogous regions is still significantly more temperature-dependent in the mesophilic protein. Thus there has to be another factor that determines the dMSD/dT in these regions. We now examine how the larger proportion of charged residues generally observed in thermophilic proteins can contribute to lowering dMSD/dT by forming salt bridge interactions. Previously we have shown in Figure 6b that a striking difference in the thermal behavior of both proteins lies in residues 50-58 and the C-terminal tail of the thermophilic protein, which undergo a decrease in MSD with temperature. To rationalize the origin of this amazing behavior, we have examined all salt bridges formed by charged residues of these two regions. In particular, we determined the probability density of the distance of all salt bridges formed between a given residue and oppositely charged residues within 20 Å. Figure 8 shows the corresponding probability densities for residues R56 and K311, which is located at the C-terminal tail. It is evident from the figure that both residues exhibit an increase in the probability of forming salt bridges at short distances upon increasing temperature. In particular, the probability of finding a short salt bridge (with distance ≤ 4.5 Å), or *occupancy*, is found to increase a 25% and a 23 % in R56 and K311, respectively, when raising the temperature from 280 to 320 K. It is noteworthy that, in the case of residue R56, increasing the temperature from 280 to 320 K entails the formation of new salt bridges with residue R56, since at low temperature none short salt bridges are formed as shown in Figure 8a (blue curve). Therefore, either the formation of new salt bridges or the strengthening of already present salt bridges with temperature is likely to constrain the motion of the specific residue and that of nearby residues. Taking into account that these effects are induced by increases in temperature, the negative dMSD/dT observed in Figure 6 for these regions is consistent with this idea. Indeed, such tightening of salt bridges upon raising temperature has already been pointed out by others as a mechanism for thermal stability [36-39]. As the temperature increase, the dielectric constant of water decreases leading to a reduction of the desolvation penalty of salt bridge formation that

ultimately strengthens this interaction and the robustness of the structure. Our present observation is fully consistent with recent NMR studies by Vinther et al. on another thermostable protein [39]. They revealed that, at the pico-nanoseconds time scale, the backbone flexibility in different areas of the protein decreases with temperature as a result of salt bridge tightening, which ultimately contributes to increase protein stability.

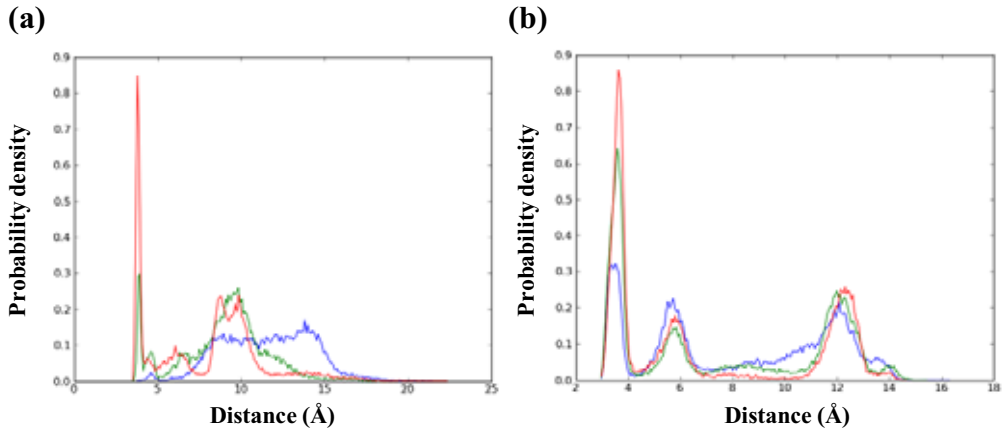


Figure 8. Probability distributions of all salt bridges integrating residues R56 (panel a) and K311 (panel b) of the thermophilic protein at 280 (blue), 300 (green) and 320 K (red). The occupancy of the salt bridge is the integral of these distributions from 0-4.5 Å. The salt bridge distance is defined between the CZ atom of Arg (and NZ atom of Lys) and the CG atom of Asp (and CD atom of Glu).

In Figure 9a we show that the content in charged residues follows a similar trend as that of the packing at different shells of the structure. The main difference in the content of charged residues lies again near the surface, as commonly accepted [3]. While the mesophilic protein presents a rather constant proportion of charged residues throughout the structure, the thermophilic one steadily increases this proportion as the distance to the surface is reduced. For instance, charged residues at the surface of the thermophilic protein represent a 55%, whereas only a 32% in the mesophilic homolog. Regarding the content on charged residues, it is interesting to note that the thermophilic protein presents another signature of enhanced thermostability, that is the higher arginine-to-lysine ratio [1, 40, 41] relative to the mesophilic homolog (0.66 *vs* 0.42).

A clear manifestation of salt bridge tightening with temperature at the surface of the thermophilic protein is the thermal behaviour of the radius of gyration (Rg), which is sensitive to fluctuations of the surface. Figure 9b shows the time evolution of Rg at different temperatures. While the radius of gyration of the mesophilic protein displays modest variations with temperature, the thermophilic protein shows a significant decrease of Rg upon increasing temperature. In particular, the central hole of the thermophilic protein undergoes a contraction that entails a significant decrease in the solvent accessible surface (SAS) of residues within 20 Å from the center of mass (second shell), as shown in Figure S5 (panel a). In contrast, the SAS of the mesophilic protein (within this 20 Å shell) is much less sensitive to temperature variations (see panel b in Figure S5). This implies that upon increasing temperature, the thermophilic protein has a tendency to disrupt protein-water interactions to enhance intra-protein interactions, as opposed to the mesophilic homolog. Figure 9c shows that the thermophilic protein systematically experiences salt bridge tightening in each shell, but that the major tightening occurs in the second shell. In particular, we note that salt bridges involving R56, which is located within this shell and has been highlighted in Figure 8a, are those that contribute the most to the tightening of salt bridges in this shell. On the contrary, Figure 9d shows that the salt bridge tightening with temperature undergone by the mesophilic protein is less systematic in the different shells. In some cases an increase in temperature does not necessarily involve an increase in the occupancy of salt bridges. The larger proportion of charged residues (see Figure 9a) underscores the importance of the more efficient tightening of salt bridges upon increasing temperature throughout the thermophilic structure. Although we have identified two regions with a negative dMSD/dT, this does not exclude the possibility that the motion of other regions is constrained due to temperature increase. We suggest that the effect of salt bridge tightening in other regions would be to partially counterbalance the natural increase of mobility with temperature.

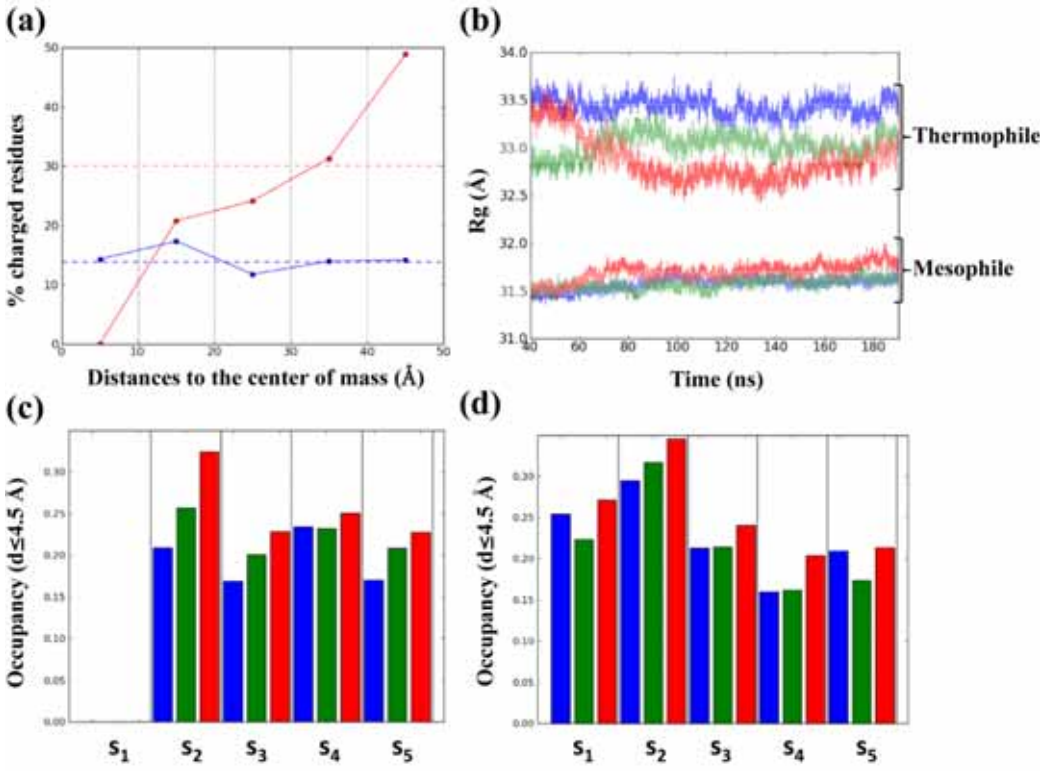


Figure 9. (a) Proportion of charged residues at different shells of the protein structure. Vertical dotted lines denote the range of distances defining a shell. Horizontal dashed lines represent the average of each data series over the five shells considered. The last shell corresponds to distances greater than 45 Å. (b) Radius of gyration as a function of time at the three temperatures studied (280 K, blue; 300 K, green; 320 K, red). (c) Average occupancy of all salt bridges of the thermophilic protein within each shell s_i at the three temperatures (same color code as in panel b). The protein structure is divided in 5 shells of a 10 Å-width as in Figure 7. (d) Same as panel c for the mesophilic protein.

It is pertinent to note that the radius of gyration undergoes slow fluctuations, which is indicative of a slowly convergent property. This makes necessary monitoring this property at long time scales to allow meaningful interpretation. For instance, from the 40-50 ns interval it would seem that the radius of gyration is converged (see Figure S6), whereas the full 200 ns run reveals that the fluctuating Rg is unlikely to be converged and, thus, much longer trajectories are required.

In this work we provide evidence of two factors determining the thermal behavior of MSD: structure and salt bridge interactions. Although both factors play a primary role at the ps-ns time scales here explored, this does not rule out the contribution of other factors underlying the difference in dMSD/dT between both proteins. Future work examining longer time scales is expected to provide new insight into their different sensitivity of protein motions to temperature changes.

Principal component analysis

We have carried out a principal component analysis of the trajectories at each temperature to give insight into differences in collective motions between both proteins. We have projected the trajectories obtained at each temperature into the first two PCs, which describe a larger amount of variance. Figure 10 represents the contour plots of the probability density of projections. It is evident from the figure that the conformational fluctuations of the two proteins are distributed in a strikingly different fashion along the first two PCs. At a given temperature, the thermophilic protein fluctuates in a higher extent than the mesophilic one, when either moving within the same minimum or between minima of the conformational space defined by the first two PCs. Of course, the larger fluctuations observed in the thermophilic protein system are consistent with the higher flexibility at short time scales as pointed out above. It is appealing the idea that the largest conformational transitions occur through *tight* pathways in the mesophilic protein, while in the thermophilic homolog *broader* routes are allowed. This ultimately provides insight into a significant difference in topology of the energy landscape of both types of proteins.

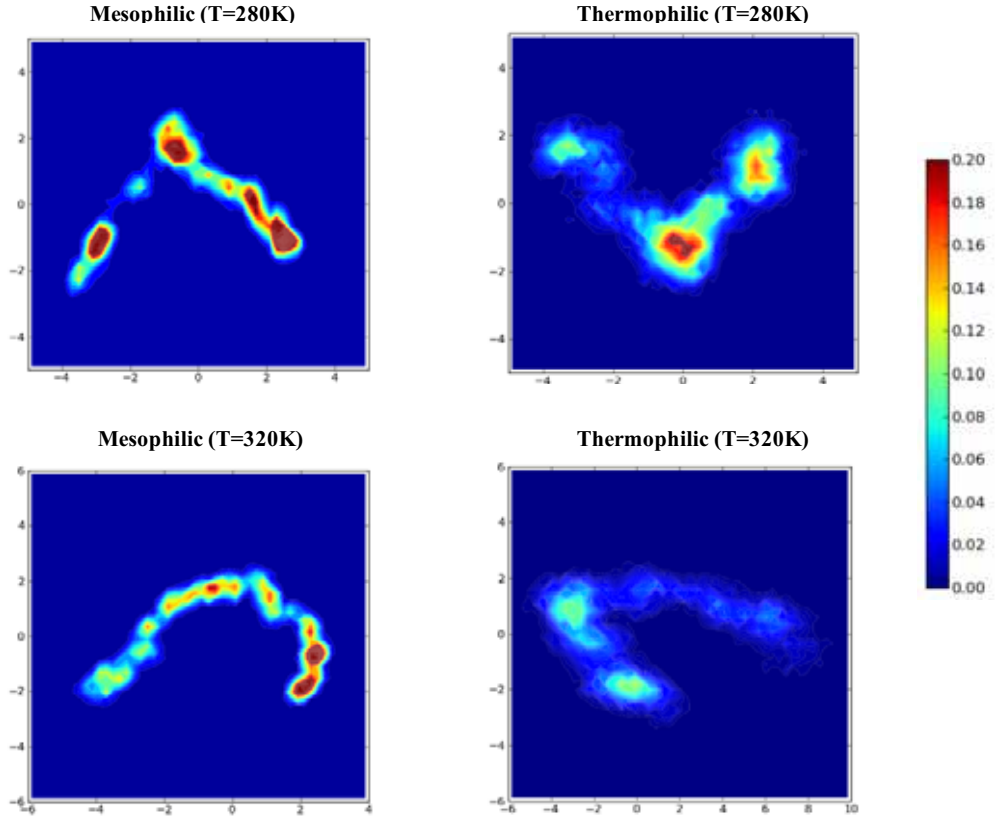


Figure 10. Contour plots of the projection of trajectories into the first two principal components at 280 and 320 K.

Taking into account such difference in the dispersion of conformational fluctuations of the two systems along the first two PCs (Figure 11), we examined how this might be linked to the difference in dynamical heterogeneity highlighted above. We scrutinized this possibility with the collectivity [27] of each principal component (see Methods) to give a measure of how many residues contribute to a given mode and thus characterize the dynamical heterogeneity associated to each mode.

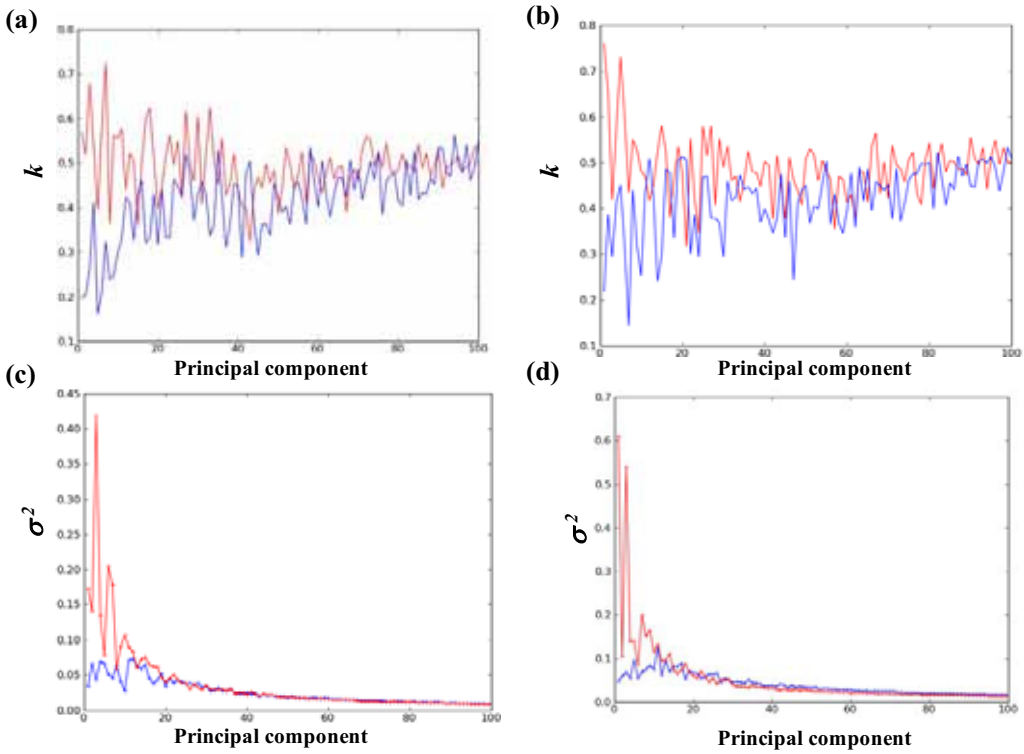


Figure 11. Collectivity of the first 100 principal components at (a) 280 K and (b) 320 K. Variance of the projection of trajectories into the first 100 PCs at (c) 280 K and (d) 320 K. This has been computed with Eq. 4. Data series for the mesophilic and thermophilic proteins are depicted in blue and red respectively.

Figure 11 displays the collectivity indices of the first 100 principal components at 280 (panel a) and 320 K (panel b). Interestingly, in the first PCs, the collectivity of the mesophilic protein is notably lower than that of the thermophilic counterpart, being remarkable the difference in collectivity of the first two PCs addressed above. These results indicate that the main conformational fluctuations of the mesophilic protein have a more local character than in the thermophilic one, which presents more cooperative motions. Put it more succinctly, the higher dynamical heterogeneity of the mesophilic protein suggested above is clearly manifested here by the low collectivity of the first principal components. Given the marked difference in the cooperative nature of motions in the two proteins, we wonder whether this might be related to the different topology of

their energy landscapes as illustrated in Figure 10. Does the broader energy landscape of the thermophilic protein describe more collective motions? There is an apparent correlation when considering the first two PCs, but to explore further this idea we examined the variance of the projection of trajectories into each of the first 100 principal components (see Methods for details on the calculation of this variance). Figure 11 (panels c and d) shows the variance of projections of each PC for the two proteins. It reveals that the variance in the first PCs is systematically higher in the thermophilic protein. Interestingly, the modes in which the thermophilic protein has a higher collectivity than the mesophilic homolog coincide well with those in which the variance is also higher. This strongly suggests that the more cooperative motions of the thermophilic protein emerge from broader energy landscapes. It is of interest to assess whether this feature of the energy landscape can be regarded as a hallmark distinguishing thermophilic proteins from their mesophilic counterparts in general. For a more detailed characterization of the landscape of both proteins, we leave as future work the use of clustering analysis methods in conjunction with Markov models for studying the kinetics associated to the conformational transitions of each protein.

CONCLUSIONS

The present study focuses on differences in the dynamical properties of a thermophilic protein and its mesophilic homolog that can be associated to thermal stability. The analysis of molecular dynamics trajectories at three different temperatures reveals that the overall flexibility (MSD) of the thermophilic protein is higher than that of the mesophilic homolog at the picosecond and few nanoseconds time scales. However, the flexibility of the mesophilic protein exhibits a higher sensitivity to temperature changes. The lower temperature dependence of the MSD of the thermophilic protein lies in the improved packing of the structure and the larger content on charged residues which mediate salt bridge interactions that tend to strengthen with temperature. Moreover, the distribution of mobilities among protein residues differs in the two proteins. The mesophilic protein exhibits a broader distribution which implies a higher dynamical heterogeneity. In this case, loops and N and C termini are very mobile but the core of the protein is more rigid than in the thermophilic protein. Such difference in the dynamical heterogeneity is reflected in the principal components describing the main conformational fluctuations of the proteins. The modes of the thermophilic protein are characterized by a higher

collectivity entailing more cooperative motions, whereas the first modes of the mesophilic homolog tend to have a more local character. Overall, we suggest that the more cooperative motions of the thermophilic protein emerge from broader energy landscapes.

On balance, our findings from simulations on the lower dMSD/dT of the thermophilic protein support the fundamental idea originally suggested by Zaccai and co-workers. Given the high correlation of the flexibility between time scales differing in three orders of magnitude (from 20 ps to 10 ns), it is tempting to regard a low dMSD/dT as a dynamic fingerprint of enhanced thermostability. Here we have extended the time scales studied in the experiment and, as future work, we aim to study this contrasting thermal behavior at much longer time scales, which are more relevant to unfolding events. This will be done with coarse-grained models that will be calibrated by the all-atom simulations analyzed here. In addition to this, the present paper suggests two additional dynamic fingerprints distinguishing thermophilic and mesophilic proteins. First, the higher dynamical hetereogeneity of the mesophilic protein and, second, the broader energy landscape of the thermophilic protein related to more cooperative motions. It is certainly true that these predictions need further testing in other thermo-mesophilic pairs of proteins to ascertain the extent to which these properties can be generalized.

ACKNOWLEDGMENT

We are grateful to Miriam Reverter for her assistance in the analysis of trajectories. The authors thank the Galicia Supercomputing Center for computational resources. This work was supported by grants from the JAE programme of CSIC, the Spanish MEC (CTQ2009-08223) and the Catalan AGAUR (2005SGR00111).

REFERENCES

1. Vieille C, Zeikus GJ (2001) Hyperthermophilic enzymes: Sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev* 65:1-43.

2. Razvi A, Scholtz JM (2006) Lessons in stability from thermophilic proteins. *Prot Sci* 15:1569-1578.

3. Kumar S, Nussinov R (2001) How do thermophilic proteins deal with heat? *Cell Mol Life Sci* 58:1216-1233.

4. Wrba A, Schweiger A, Schultes V, Jaenicke R, Zavodszky P (1990) Extremely thermostable D-glyceraldehyde-3-phosphate dehydrogenase from the Eubacterium Thermotoga-Maritima. *Biochemistry* 29:7584-7592.

5. Lazaridis T, Lee I, Karplus M (1997) Dynamics and unfolding pathways of a hyperthermophilic and a mesophilic rubredoxin. *Protein Sci* 6:2589-2605.

6. Zavodszky P, Kardos J, Svingor A, Petsko GA (1998) Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc Natl Acad Sci U S A* 95:7406-7411.

7. Butterwick JA, Loria JP, Astrof NS, Kroenke CD, Cole R, Rance M, Palmer AG (2004) Multiple time scale backbone dynamics of homologous thermophilic and mesophilic ribonuclease HI enzymes. *J Mol Biol* 339:855-871.

8. Wolf-Watz M, Thai V, Henzler-Wildman K, Hadjipavlou G, Eisenmesser EZ, Kern D (2004) Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat Struct Mol Biol* 11:945-949.

9. Fitter J, Heberle J (2000) Structural equilibrium fluctuations in mesophilic and thermophilic alpha-amylase. *Biophys J* 79:1629-1636.

10. Hernandez G, Jenney FE, Adams MWW, LeMaster DM (2000) Millisecond time scale conformational flexibility in a hyperthermophile protein at ambient temperature. *Proc Natl Acad Sci U S A* 97:3166-3170.

11. Tehei M, Madern D, Franzetti B, Zaccai G (2005) Neutron scattering reveals the dynamic basis of protein adaptation to extreme temperature. *J Biol Chem* 280:40974-40979.

12. Grottesi A, Ceruso MA, Colosimo A, Di Nola A (2002) Molecular dynamics study of a hyperthermophilic and a mesophilic rubredoxin. *Proteins: Struct, Funct, Genet* 46:287-294.

13. Wintrode PL, Zhang DQ, Vaidehi N, Arnold FH, Goddard WA (2003) Protein dynamics in a family of laboratory evolved thermophilic enzymes. *J Mol Biol* 327:745-757.

14. Colombo G, Merz KM (1999) Stability and activity of mesophilic subtilisin E and its thermophilic homolog: Insights from molecular dynamics simulations. *J Am Chem Soc* 121:6895-6903.

15. Stone MJ (2001) NMR relaxation studies of the role of conformational entropy in protein stability and ligand binding. *Acc Chem Res* 34:379-388.

16. Marcos E, Mestres P, Crehuet R (2011) Crowding Induces Differences in the Diffusion of Thermophilic and Mesophilic Proteins: A New Look at Neutron Scattering Results. *Biophys J* 101:2782-2789.

17. Lee BI, Chang C, Cho SJ, Eom SH, Kim KK, Yu YG, Suh SW (2001) Crystal structure of the MJ0490 gene product of the hyperthermophilic archaebacterium Methanococcus jannaschii, a novel member of the lactate/malate family of dehydrogenases. *J Mol Biol* 307:1351-1362.

18. Dunn CR, Wilks HM, Halsall DJ, Atkinson T, Clarke AR, Muirhead H, Holbrook JJ (1991) Design and synthesis of new enzymes based on the lactate-dehydrogenase framework. *Philos Trans R Soc Lond B Biol Sci* 332:177-184.

19. Hess B, Kutzner C, Van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput* 4:435-447.

20. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* 118:11225-11236.

21. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* 79:926-935.

22. Darden T, York D, Pedersen L (1993) Particle Mesh Ewald - An N·log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* 98:10089-10092.

23. Hess B, Bekker H, Berendsen HJC, Fraaije J (1997) LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* 18:1463-1472.

24. Berendsen HJC, Postma JPM, Van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81:3684-3690.

25. Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. *Proteins: Struct, Funct, Genet* 17:412-425.

26. Chennubhotla C, Bahar I (2006) Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. *Mol Syst Biol* 2:36.

27. Bruschweiler R (1995) Collective protein dynamics and nuclear-spin relaxation. *J Chem Phys* 102:3396-3403.

28. Holm L, Kaariainen S, Rosenstrom P, Schenkel A (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics* 24:2780-2781.

29. Hayward JA, Smith JC (2002) Temperature dependence of protein dynamics: Computer simulation analysis of neutron scattering properties. *Biophys J* 82:1216-1225.

30. Meinhold L, Clement D, Tehei M, Daniel R, Finney JL, Smith JC (2008) Protein dynamics and stability: The distribution of atomic fluctuations in thermophilic and mesophilic dihydrofolate reductase derived using elastic incoherent neutron scattering. *Biophys J* 94:4812-4818.

31. Thompson MJ, Eisenberg D (1999) Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. *J Mol Biol* 290:595-604.

32. Chan MK, Mukund S, Kletzin A, Adams MWW, Rees DC (1995) Structure of a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase. *Science* 267:1463-1469.

33. Britton KL, Yip KSP, Sedelnikova SE, Stillman TJ, Adams MWW, Ma K, Maeder DL, Robb FT, Tolliday N, Vetriani C, Rice DW, Baker PJ (1999) Structure determination of the glutamate dehydrogenase from the hyperthermophile Thermococcus litoralis and its comparison with that from Pyrococcus furiosus. *J Mol Biol* 293:1121-1132.

34. Zhou YQ, Vitkup D, Karplus M (1999) Native proteins are surface-molten solids: Application of the Lindemann criterion for the solid versus liquid state. *J Mol Biol* 285:1371-1375.

35. Camps J, Carrillo O, Emperador A, Orellana L, Hospital A, Rueda M, Cicin-Sain D, D'Abramo M, Gelpi JL, Orozco M (2009) FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics* 25:1709-1710.

36. de Bakker PIW, Hunenberger PH, McCammon JA (1999) Molecular dynamics simulations of the hyperthermophilic protein Sac7d from Sulfolobus acidocaldarius: Contribution of salt bridges to thermostability. *J Mol Biol* 285:1811-1830.

37. Thomas AS, Elcock AH (2004) Molecular simulations suggest protein salt bridges are uniquely suited to life at high temperatures. *J Am Chem Soc* 126:2208-2214.

38. Danciulescu C, Ladenstein R, Nilsson L (2007) Dynamic arrangement of ion pairs and individual contributions to the thermal stability of the cofactor-binding domain of glutamate dehydrogenase from Thermotoga maritima. *Biochemistry* 46:8537-8549.

39. Vinther JM, Kristensen SM, Led JJ (2011) Enhanced Stability of a Protein with Increasing Temperature. *J Am Chem Soc* 133:271-278.

40. Mrabet NT, Vandenbroeck A, Vandenbrande I, Stanssens P, Laroche Y, Lambeir AM, Matthijssens G, Jenkins J, Chiadmi M, Vantilbeurgh H, Rey F, Janin J, Quax WJ, Lasters I, Demaeyer M, Wodak SJ (1992) Arginine residues as stabilizing elements in proteins. *Biochemistry* 31:2239-2253.

41. Siddiqui KS, Poljak A, Guilhaus M, De Francisci D, Curmi PMG, Feller G, D'Amico S, Gerday C, Uversky VN, Cavicchioli R (2006) Role of lysine versus arginine in enzyme cold-adaptation: Modifying lysine to homo-arginine stabilizes the cold-adapted alpha-amylase from Pseudoalteramonas haloplanktis. *Proteins: Struct, Func, Bioinf* 64:486-501.

Supporting Information

Weibull function

A Weibull function takes the form:

$$\rho(r) = \frac{\alpha}{\beta}\left(\frac{r}{\beta}\right)^{\alpha-1} e^{-\left(\frac{r}{\beta}\right)^{\alpha}} \quad (1)$$

where α and β are the two parameters that determine the shape of the function. The standard deviation associated to a distribution function is given by:

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 \rho(x)\, dx} \quad (2)$$

where μ is the average obtained from this distribution function and is calculated as:

$$\mu = \int_{-\infty}^{\infty} x \cdot \rho(x)\, dx \quad (3)$$

Regarding the fits shown in Figure 4a, the parameter values obtained for the two distributions are:

	α	β
Mesophilic	2.306	0.596
Thermophilic	2.624	0.887

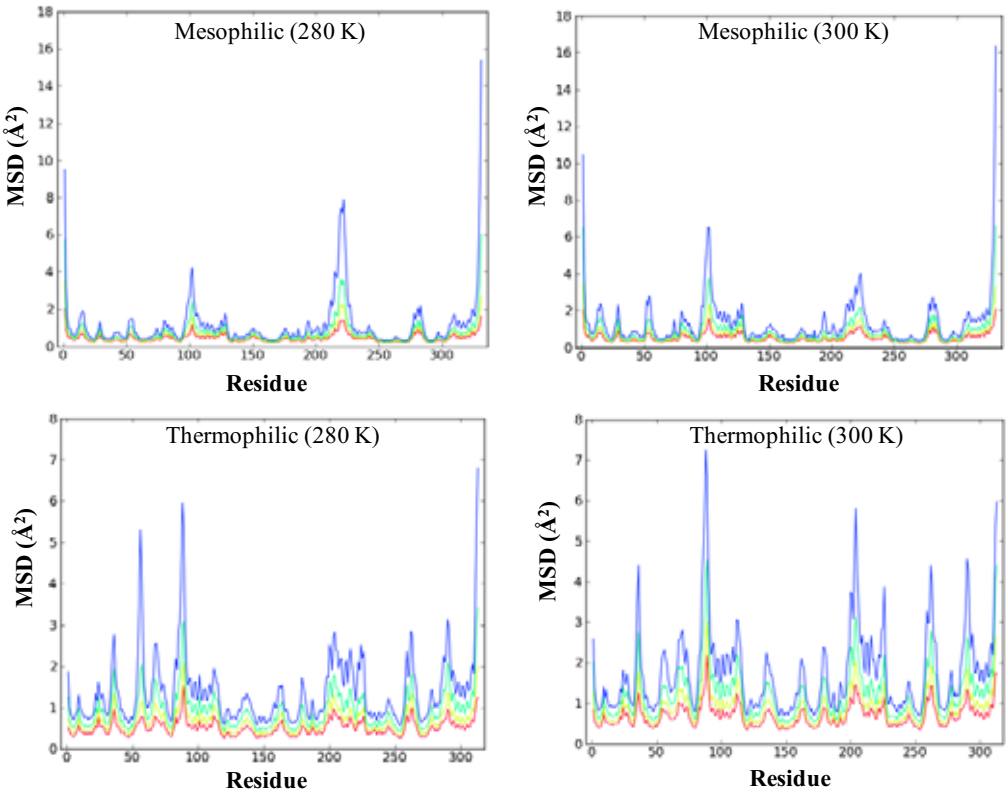


Figure S1. MSDs of the two proteins at different temperatures (280 and 300 K) and time scales: 20 ps (red), 100 ps (yellow), 1 ns (green) and 10 ns (blue).

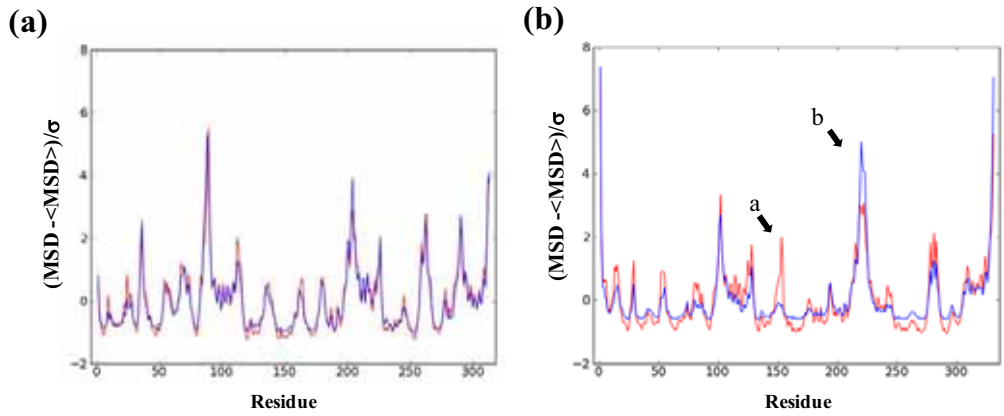


Figure S2. Normalized representation of residues flexibilities at 320 K of the (a) thermophilic and (b) mesophilic proteins. Two time scales are represented: 20 ps (red) and 10 ns (blue).

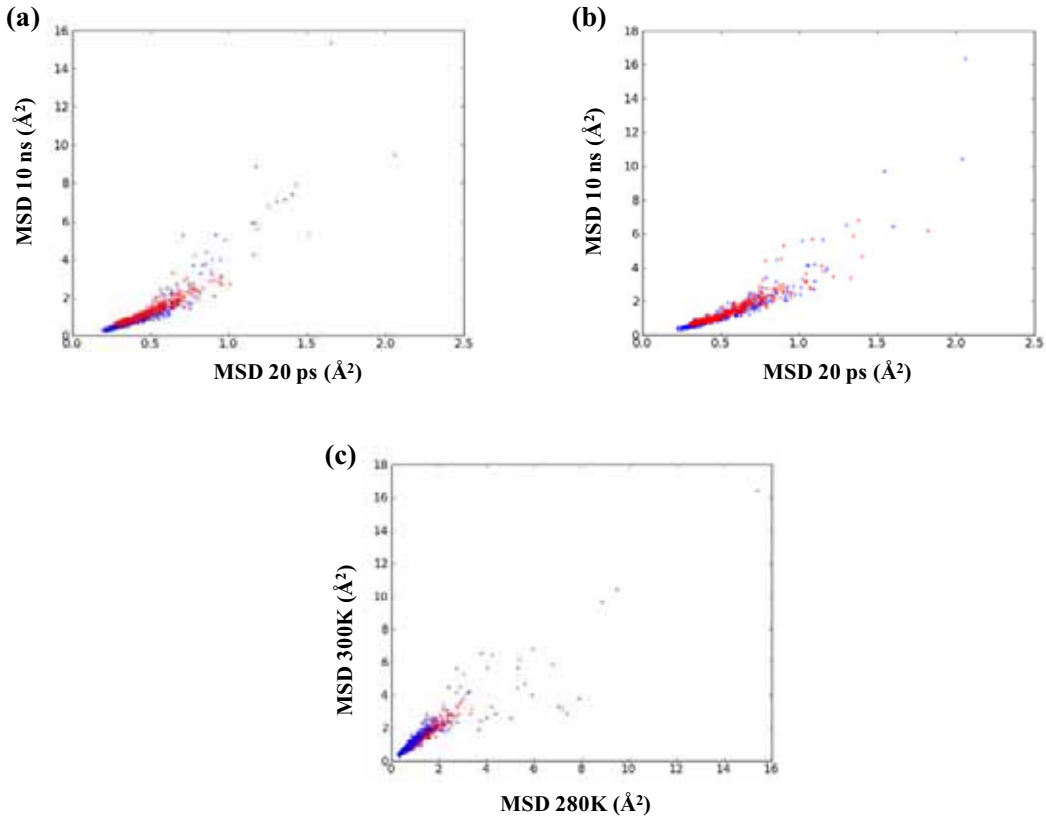


Figure S3. (a) Correlation between the MSDs at 20 ps and 10 ns of the thermophilic (red crosses) and mesophilic (blue crosses) proteins at 280 K.(b) Same as panel a at 300 K (c) Correlation between the MSDs (10 ns) at 280 and 300 K for both proteins.

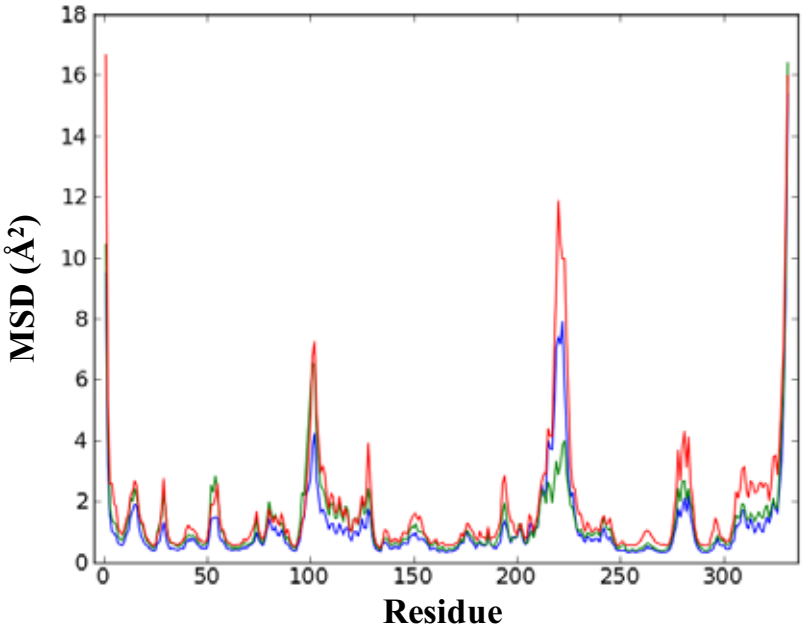


Figure S4. MSD (10ns) of the mesophilic protein at 280 (blue), 300 (green) and 320 K (red).

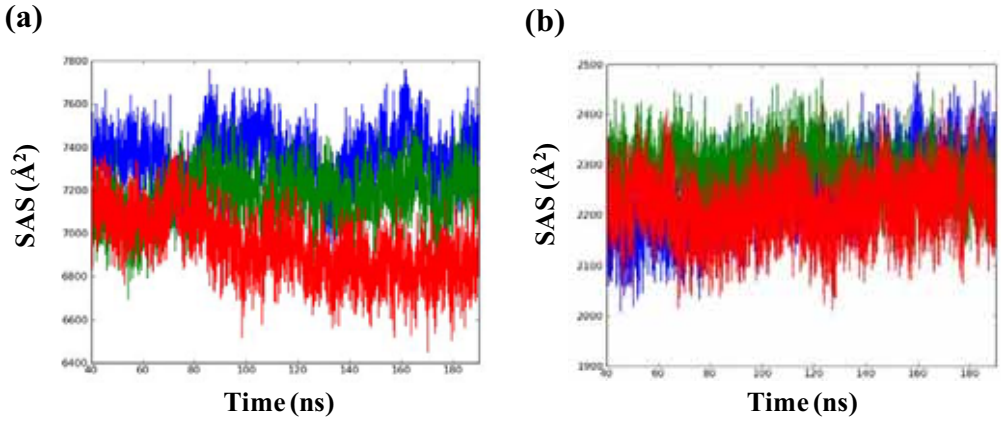


Figure S5. Solvent accessible surface (SAS) of residues of the inner hole (within 20 Å from the center of mass) of the (a) thermophilic and (b) mesophilic proteins. All non-hydrogen atoms have been considered and a 1.4 Å radius probe has been used.

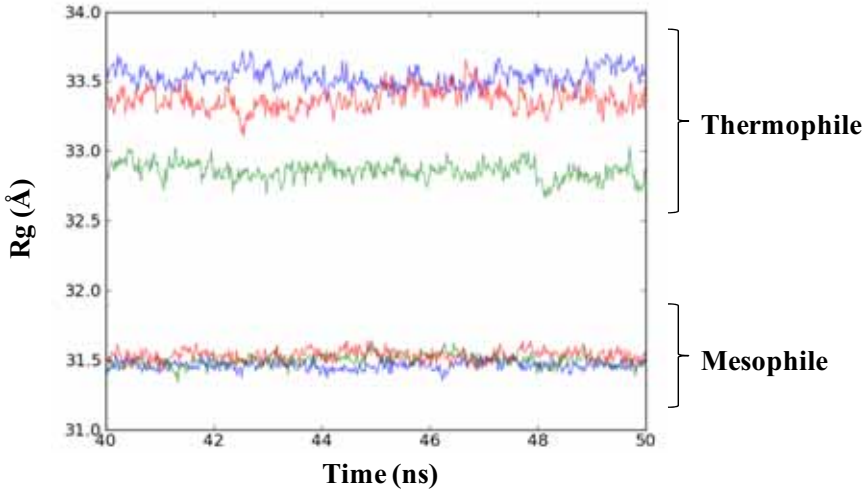


Figure S6. Time evolution of the radius of gyration in a 10 ns interval at the three temperatures studied (280 K, blue; 300 K, green; 320 K, red) for both proteins.

CHAPTER 5

CONCLUSIONS

With the aim of providing a global picture of different aspects relevant to enzymatic function (reactivity, dynamics and thermostability), in the present thesis we have studied several enzymes using a variety of computer simulation methods. Our main observations in the three aspects are the following:

1) Reactivity

a. Putative pentacoordinated phosphorus species being formed in the course of phosphoryl transfer reactions exhibit apical bonds of high polarizable character. We have observed that this makes the geometry and stability of these compounds very sensitive to three different sources of polarization. First, the electrodonor character of the apical and equatorial groups. Second, the orientation of equatorial groups. Third, external electric field effects in the apical direction. Electric fields, in particular, are observed to either enhance the kinetic stability of such intermediate or destabilize it to the extent to switch the step-wise process to a concerted one.

b. The magnitudes of the electric fields observed to induce significant changes in geometry and reactivity are likely to be present in the active site of phosphoryl transfer enzymes. This is an effect included in an implicit fashion when performing QM/MM calculations.

c. From the methodological point of view, the computational method with best compromise between accuracy and cost to describe the three sources of polarization of pentacoordinated phosphorus is the *m*PW1PW91 functional with the 6-31+G(d) basis set. To describe these effects with lower computational cost, the semi-empirical AM1/d-PhoT exhibits the best performance.

d. The studies on small phosphorane models have provided critical thinking to evaluate the viability of pentacoordinated intermediates. In particular this initially questioned the presence of a phosphorane in the X-ray structure of the *β*-phosphoglucomutase enzyme.

e. The phosphoryl transfer in *β*-phosphoglucomutase computed with high-level QM/MM methods proceeds through a pentacoordinated phosphorus that is a transition state structure. The energy barrier calculated for this process is in good agreement with kinetic experiments. Moreover, the improved refinement of the X-ray diffraction map with MgF_3^- provides additional support to the idea that the original X-ray structure was wrongly characterized as a phosphorane intermediate. Overall, this provides further support against the observation of the first phosphorane in an enzyme active site.

2) Dynamics

a. The conformational change of the *Ec*NAGK enzyme associated to substrate binding/release events is intrinsically favored by the 3D-structure. Thus the protein fold shared among Amino Acid Kinase family members entails common modes of motion that are functionally relevant.

b. The interest of the scheme devised to contrast the large-amplitude dynamics of pairs of proteins extends beyond the common dynamical mechanism of the AAK protein

family. It provides us with a rigorous framework to find out dynamic fingerprints in protein families.

c. Among the oligomers studied from the AAK family, it is commonly observed that the dynamics of the monomeric and dimeric component subunits is conserved to a high extent in the oligomeric and functional state. In addition, the oligomeric arquitecture enables new cooperative modes of motion intimately linked to the allosteric regulation of the activity as observed for hexameric NAG and UMP kinases. Inter-protein interfaces provided by oligomerization therefore are functionally relevant not only to stabilize a given fold for the monomeric subunits, but also to confer new modes of motion necessary for optimal activity.

d. In particular, the different assembly of the dimeric subunits uniquely displayed by UMPK induces rigid-body motions of the monomeric subunits that are not allowed by the NAGK arquitecture and are necessary for allosteric regulation.

3) Thermostability

a. The interpretation from neutron scattering experiments that the intramolecular motions of a thermophilic enzyme are less sensitive to temperature changes must be reconsidered. Translational and rotational diffusive motions account for ~50% of the total mobility observed as shown by molecular and Brownian dynamics simulations.

b. The different thermal behavior of the global mobility exhibited by the two enzymes arises from the different nature of inter-protein interactions present in the crowded solution. The larger number of charged amino acid residues at the surface of the thermophilic enzyme induces more intense electrostatic interactions among protein molecules that are enhanced with temperature due to a decrease in the desolvation penalty of water. These interactions are attractive and partially counterbalance the natural increase in diffusion with temperature.

c. With regard to internal dynamics, MD simulations reveal that the thermophilic enzyme exhibit larger structural fluctuations in the range of temperatures studied, which is in qualitative agreement with the experiment. At longer time scales than in the experiment, the MSD of the thermophilic protein becomes less temperature-

dependent than that of the mesophilic homologue. This supports the fundamental idea originally proposed in the experiment. Such lower dependence arises from improved packing of the protein structure and the larger proportion of charged residues which form intramolecular salt bridges that tighten with temperature.

d. The distribution of residues' flexibilities is broader in the mesophilic protein. This higher dynamical heterogeneity is attributed to more local motions that arise from a less compact structure, which presents enlarged coil regions. The thermophilic protein, on the other hand, exhibits more cooperative motions that emerge from broader energy landscapes.

A more general conclusion from this investigation is the increasing importance of computational methods in correctly interpreting experimental data, in spite of the limitations of current computational approaches. We have shown two examples of different experiments with far-reaching conclusions that after computational examination need to be reconsidered.

CHAPTER 6

SUMARIO

La presente tesis se centra en la modelización molecular de enzimas haciendo especial énfasis en tres aspectos necesarios para la función enzimática: reactividad, dinámica y termostabilidad. Se ha pretendido dar una visión general del mecanismo de funcionamiento de los enzimas dirigiendo el estudio no a un enzima en concreto sino a una variedad de enzimas con características distintas y que han abierto cuestiones diversas. El estudio de los tres aspectos considerados en esta tesis se han divido en tres partes diferentes. En primer lugar, los estudios de reactividad se han centrado en reacciones de transferencia de fosfato, las cuales están implicadas en un amplio rango de procesos biológicos y que presentan la complejidad de proceder según distintos tipos de mecanismos. En este sentido las herramientas de la química computacional pueden ayudar a discernir el tipo de mecanismo. Uno de los ejemplos que se mostrará es el del enzima β-fosfoglucomutasa. En segundo lugar, el estudio de la dinámica se ha focalizado principalmente en los movimientos lentos de gran amplitud que están asociados a procesos de unión de sustrato y regulación de la actividad por alosterismo. Esta parte se ha centrado en la familia de las quinasas de aminoácido que presentan importantes cambios conformacionales asociados a su función biológica y que han sido muy bien caracterizadas por cristalografía. En tercer lugar, se ha analizado la relación entre termostabilidad y dinámica en el contexto del experimento de Zaccai y colaboradores basado en dispersión de neutrones.

Catálisis enzimática

Los enzimas son macromoléculas biólogicas que son capaces de acelerar la velocidad de reacción en más de 15 órdenes de magnitud, lo cual les permite llevar a cabo reacciones químicas a escalas de tiempo adecuadas para los procesos biológicos. La clave de la extraordinaria eficiencia catalítica de los enzimas reside en la gran preorganización del centro activo, el cual presenta amino ácidos con distinta polaridad en una conformación óptima para unir el sustrato y estabilizar el estado de transición de la reacción. En general, la comprensión detallada del mecanismo catalítico de enzimas es de gran utilidad tanto para el diseño de inhibidores de enzimas asociados a algún tipo de disfunción, así como para rediseñar enzimas adaptados a catalizar nuevas reacciones químicas.

Estudios recientes de RMN y espectroscopía unimolecular han revelado la importancia de la flexibilidad (dinámica) de los enzimas en la catálisis. Los cambios conformacionales implicados en la catálisis pueden tener lugar en un amplio rango de escalas de tiempo. Las fluctuaciones más rápidas (escala de pico-nanosegundos) corresponden a movimientos locales que implican cadenas laterales y subdominios, mientras que la dinámica más lenta (escala de micro-milisegundos) concierne a dominios enteros que cambian de conformación en procesos de unión de ligando. El estudio de la dinámica de enzimas a nivel computacional presenta limitaciones para abarcar las escalas más lentas. Concretamente, la dinámica molecular (DM) es el método más utilizado para explorar la dinámica conformacional de proteínas. Mediante las ecuaciones de movimiento de Newton, la DM simula la evolución temporal de las coordenadas atómicas permitiendo caracterizar diferentes estados conformacionales y sus transiciones. Para estudiar procesos en escalas de tiempo de micro-milisegundos es necesario recurrir a métodos más aproximados como los basados en una representación más simplificada de la proteína, como los modelos de red elástica.

Modelización molecular

La química teórica y computacional proporciona un amplio rango de metodologías que permite describir, a distintos niveles de aproximación, los procesos reactivos y conformacionales que tienen lugar en enzimas. Por un lado, los métodos basados en la mecánica cuántica se utilizan en la descripción de la reactividad del centro activo. Estos métodos son los más exactos, pero su elevado coste computacional limita su aplicación a los pocos átomos involucrados en la reacción. Concretamente se han utilizado los métodos basados en la teoría del funcional de la densidad (DFT), cuya relación exactitud-coste es favorable. Por otro lado, es preciso muestrear la superficie de energía potencial del sistema para caracterizar los estados conformacionales más representativos así como para describir

el estado de transición implicado en la reacción química. Para un amplio muestreo de esta superficie habitualmente se recurre a los campos de fuerzas que, mediante potenciales de interacción parametrizados, permiten describir la energía del sistema a un coste computacional muy bajo. Estos potenciales se utilizan en conjunto con un método de simulación que integra las ecuaciones de movimiento de Newton para describir la evolución de las coordenadas atómicas con el tiempo: la dinámica molecular. Esta metodología permite, en consecuencia, caracterizar la dinámica de proteínas, pero su uso se limita a describir procesos en escalas de tiempo de hasta cientos de nanosegundos. Para procesos a escalas más largas se requieren aproximaciones adicionales, como es el caso de los métodos de red elástica o de dinámica Browniana.

Catálisis de reacciones de transferencia de fosforilo

Las reacciones de transferencia de fosforilo son catalizadas por distintos tipos de enzimas en una amplia diversidad de procesos bioquímicos. Por ejemplo, las quinasas son enzimas que transfieren fosfato desde el ATP a otras proteínas o biomoléculas y que están implicadas en procesos como la división celular o la glucólisis. Aparte del gran interés que despierta estudiar este tipo de enzimas por sus potenciales aplicaciones médicas, estos enzimas merecen especial atención desde el punto de vista químico ya que son de los más eficientes que existen. El origen de esta gran eficiencia reside en que estos enzimas consiguen llevar a cabo reacciones que en disolución prácticamente no tienen lugar, debido a la extraordinaria estabilidad cinética de los fosfatos, a escalas de tiempo adecuadas para los procesos bioquímicos.

Los mecanismos de reacción de una transferencia de fosforilo pueden ser disociativos (intermedio metafosfato) o asociativos, los cuales proceden a través de una estructura de fósforo pentacoordinado que puede ser un intermedio (fosforano) o un estado de transición. El metafosfato adopta una geometría plana trigonal, mientras que el fósforo pentacoordinado (Figura 1) se trata de una bipirámide trigonal en la que el nucleófilo y el grupo saliente se sitúan en posiciones axiales y otros tres sustituyentes en el plano ecuatorial de la molécula. Estudios teóricos y experimentales han mostrado que el carácter del nucleófilo y del grupo saliente, así como la carga, tiene un efecto importante en el tipo de mecanismo. En general se observa que en la gran mayoría de enzimas la reacción procede a través de un fósforo pentacoordinado que es un estado de transición (proceso concertado) y no un intermedio estable. Sin embargo, despertó un gran interés en el año 2003 la publicación en la revista *Science* de la primera estructura cristalográfica presentando un fosforano como intermedio de reacción. Estos resultados dieron lugar a una gran polémica y hoy en día se ha reconsiderado aquella observación después de diversos estudios cinéticos y de RMN que demuestran que, en realidad, se había formado una sal de MgF_3^- que actuaba como inhibidor al ser análogo al estado de transición.

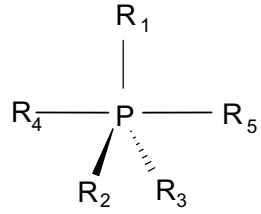


Figura 1. Estructura de un fósforo pentacoordinado. Los grupos R_1, R_2 y R_3 son ecuatoriales, mientras que los grupos R_4 y R_5 son axiales, que se corresponden al nucleófilo y al grupo saliente de la reacción de transferencia

Fósforo pentacoordinado: estructura, reactividad e implicaciones biológicas

En esta tesis se han examinado los factores que influyen en la estabilidad de compuestos pentacoordinados de fósforo. Se han estudiado distintos modelos de fósforo pentacoordinado con sustituyentes con carácter electrodador variable y se ha caracterizado el tipo de enlace de enlace presente en estos compuestos.

En primer lugar, se ha estudiado la naturaleza de los enlaces axiales en distintos modelos de fosforano con los métodos "Natural Bond Orbital" (NBO) y "Atoms in Molecules" (AIM). Los resultados muestran que ambos enlaces axiales son dativos y presentan marcada hiperconjugación, lo cual es coherente con un modelo de 3-centro 4-electrones. Los resultados del análisis sistemático, obtenidos a nivel *m*PW1PW91/6-31+G(d), muestran que la distancia de enlace axial del fósforo pentacoordinado es particularmente sensible a distintos efectos de polarización: carácter dador de los grupos axiales y ecuatoriales, la orientación de los grupos ecuatoriales y campos eléctricos externos en la dirección axial. Cuanto mayor es el carácter dador del grupo axial y más aceptor el carácter del grupo ecuatorial más cortas resultan las distancias de enlace axial. Dependiendo de los grupos presentes la distancia de enlace puede variar hasta 0.2 Å, lo cual ilustra la gran sensibilidad de este enlace a los efectos inductivos. Además, cuando el carácter dador de los dos grupos axiales es muy diferente, el enlace con el grupo dador tiende a fortalecerse a costa de debilitar el segundo enlace axial resultando en importantes diferencias en la distancia. Por otro lado, se observa que la orientación de los grupos ecuatoriales puede polarizar el átomo de fósforo afectando a la distancia de enlace axial. Teniendo en cuenta la polarizabilidad del enlace, es natural que los campos eléctricos externos tengan efectos importantes. Se observa cómo afectan a la distancia de enlace de forma que pueden estabilizar o desestabilizar el fosforano dependiendo del sentido del campo y de su magnitud. Además se aprecia como el perfil de reacción puede cambiar (Figura 2), de forma que el intermedio puede aumentar su estabilidad cinética o desestabilizarlo por completo de forma que el proceso pasa a ser concertado como consecuencia del campo. Esto puede ser de especial relevancia dada la capacidad de los centros activos de enzimas de estabilizar electrostáticamente el estado de transición.

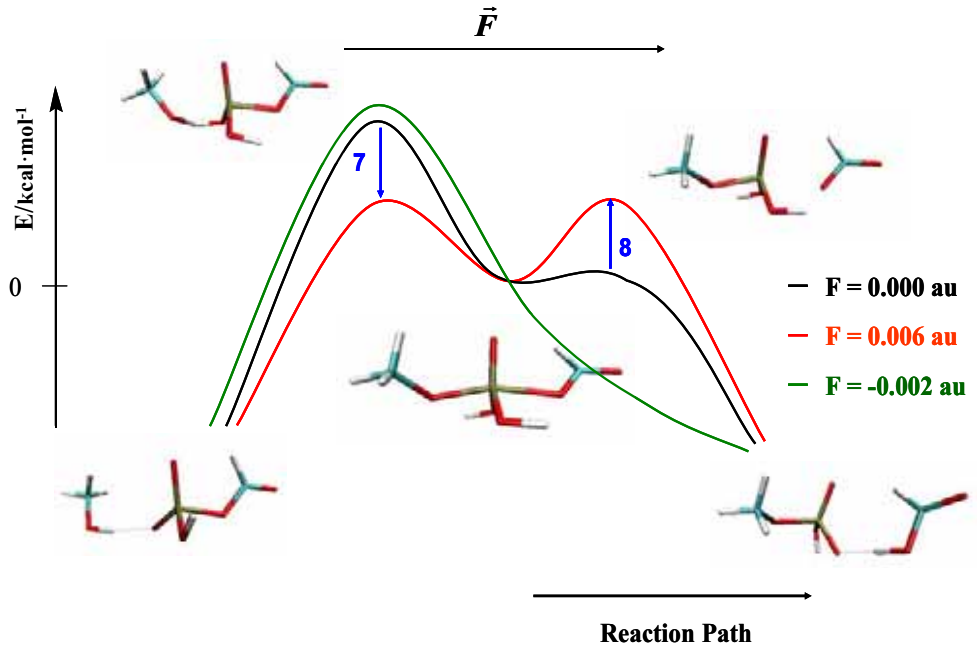


Figura 2. Esquema del perfil de energía calculado en presencia de campos eléctricos de distinta intensidad.

Una descripción detallada de este estudio se encuentra en: *Inductive and External Electric Field Effects in Pentacoordinated Phosphorus Compounds*. (2008) J. Chem. Theory Comput. 4:49-63

Fósforo pentacoordinado: evaluación de metodologías para describir los efectos de polarización

Tras la observación en el estudio previo que los campos eléctricos pueden tener un efecto importante en la estructura y reactividad del fósforo pentacoordinado, resulta necesario evaluar la calidad de la descripción que hacen diferentes métodos semi-empíricos de estos efectos. Los métodos semi-empíricos se utilizan frecuentemente en cálculos QM/MM debido a su reducido coste computacional y por lo tanto requieren describir adecuadamente la interacción de los átomos reactivos (zona QM) con el entorno electrostático de la zona MM.

Se realizó un estudio sistemático de cómo describen los semi-empíricos la geometría de dos compuestos de fósforo pentacoordinado que presentan marcados efectos de polarización debido al carácter de los sustituyentes y la orientación de los grupos ecuatoriales (Figura 3). También se evaluó como describen los correspondientes perfiles de reacción, así como el efecto de campos eléctricos externos en la barrera energética y las distancias de enlace axial. Se evaluaron semi-empíricos muy utilizados como el AM1 o el PM3 y otros que incorporan orbitales d como el PM6 o el AM1/d-PhoT. Este último, desarrollado por York y colaboradores, se ha observado que es el método que proporciona una mejor descripción de las energías, geometrías y efectos de campos eléctricos en estos compuestos. De hecho, AM1/d-PhoT se desarrolló para describir transferencias de fosforilo y consiste en una reparametrización de AM1 para átomos H, O y P que incorpora orbitales d y una corrección de la interacción core-core.

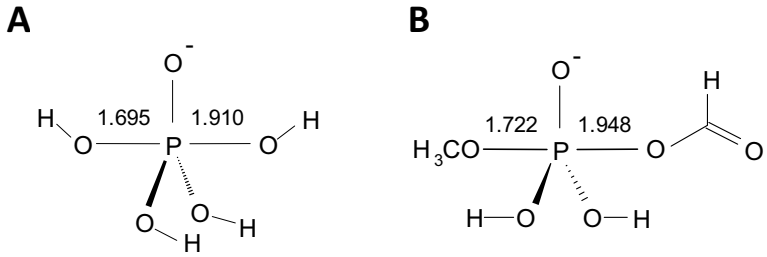


Figura 3. Estructura de los sistemas modelo estudiados que presentan efectos de polarización. (A) polarización debido a la conforamción asimétrica de los OH ecuatoriales (B) polarización inducida por la diferencia de carácter electrodador de los dos grupos axiales.

Después de plantear la posibilidad de que los campos eléctricos pueden alterar la geometría y los perfiles de reacción en centros activos de enzimas, se calculó el campo eléctrico que se genera en el centro activo de 3 enzimas de transferencia de fosforilo: β-fosfoglucomutasa, fructosa-1,6-bisfosfatasa y N-Acetil-Glutamato quinasa.. Los valores que se obtienen (0.002-0.005 au) son del mismo orden de los que se ha observado que tienen efectos

notables en los compuestos modelo. De esta forma, es plausible que los campos eléctricos en la dirección axial de la reacción enzimática tengan un efecto significativo en la reactividad. Cabe recordar que estos efectos quedan incluidos de forma implícita en cálculos QM/MM.

La descripción de los resultados de este estudio se encuentra en: *Description of pentacoordinated phosphorus under an external electric field: which basis sets and semi-empirical methods are needed*? (2008) Phys. Chem. Chem. Phys., 10, 2442-2450.

Fósforo pentacoordinado en β-fosfoglucomutasa

El conocimiento adquirido de los dos estudios previos proporciona cierta capacidad predictiva para valorar la estabilidad de una especie de fósforo pentacoordinado. De este modo, una inspección preliminar de la especie pentacoordinada de la controvertida estructura cristalográfica de β-fosfoglucomutasa señala que la marcada diferencia en el carácter dador de los grupos axiales debería dar lugar a diferencias significativas entre las dos distancias de enlace axial, en contra de lo que se observa experimentalmente.

En esta tesis se han realizado cálculos QM/MM de alto nivel para describir la reacción de transferencia de fosforilo y se confirma que el fósforo pentacoordinado es un estado de transición y no un intermedio. Además, la barrera energética calculada concuerda satisfactoriamente con el valor experimental.

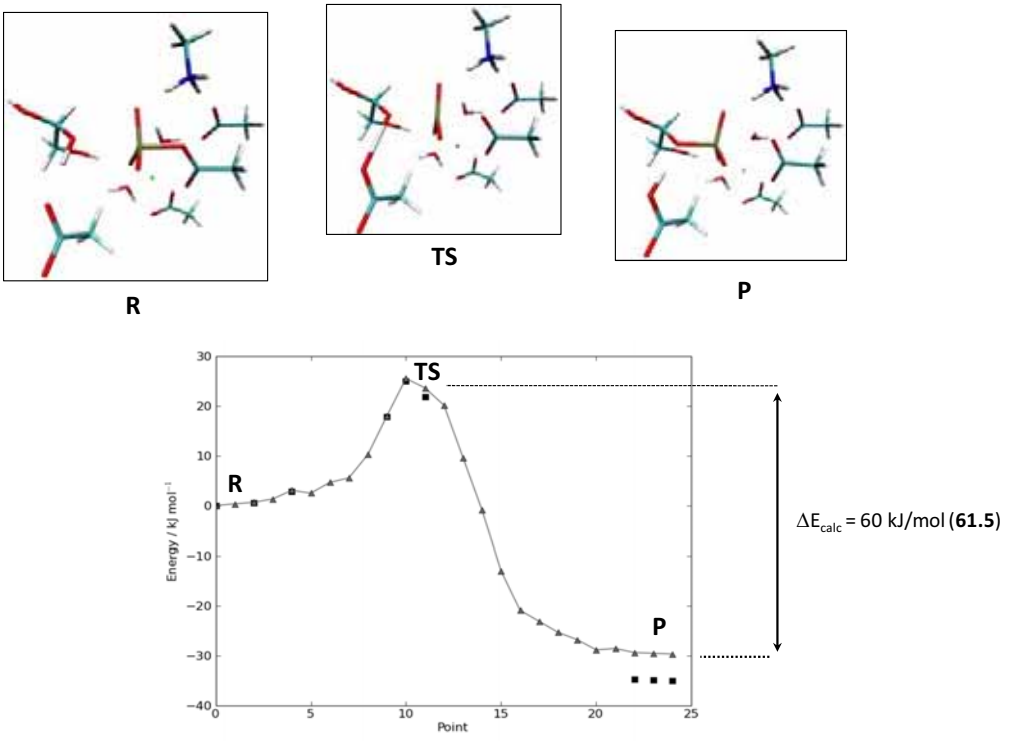


Figura 4. Perfil de reacción calculado con el método NEB obtenido a distintos niveles de teoría. Se representan las estructuras de los tres puntos estacionarios (R: reactivos, TS: estado de transición, P: productos). La barrera energética (de productos a reactivos) calculada (60 kJ/mol) se corresponde bien con la experimental (61.5 kJ/mol).

Por otro lado, el cálculo de un complejo del enzima con la sal de MgF_3^-, que se había postulado como posible inhibidor análogo al estado de transición, es efectivamente, un buen mimético del estado de transición. En resumen, se confirma la hipótesis de que el fósforo pentacoordinado no fue correctamente asignado la resolver la estructura de rayos X.

Una descripción de los resultados de este estudio se encuentra en: *Pentacoordinated phosphorus revisited by high-level QM/MM calculations* (2010) Proteins, 78, 2405-2411

Dinámica de gran amplitud en la familia de las quinasas de aminoácido

La familia de las quinasas de aminoácido la componen una serie de enzimas con un mismo plegamiento proteico y que catalizan reacciones de transferencia de fosforilo. Rubio y colaboradores han estudiado ampliamente esta familia y propusieron que la similitud estructural de estos enzimas haría que compartan el mecanismo catalítico. El enzima paradigmático de esta familia es la N-acetil-Glutamato quinasa (NAGK), que está involucrada en la ruta biosintética de la arginina en bacterias. Se piensa que la inhibición de esta enzima es una buena diana terapéutica para el desarrollo de antibióticos ya que en células mamíferas esta ruta metabólica procede a través de intermedios no acetilados y, en consecuencia, la inhibición de NAGK puede ser selectiva. Dependiendo del organismo NAGK puede adoptar una forma dimérica o hexamérica. En el caso de una estructura hexamérica se ha observado que la actividad enzimática puede regularse de forma alostérica por inhibición de arginina.

Un aspecto interesante de esta familia es que el conjunto de estructuras cristalográficas que se han resuelto en distintos estados de unión muestra que estos enzimas son altamente flexibles (Figura 5). En la forma dimérica de NAGK se ha observado un movimiento de apertura y cierre del centro activo necesario para unión de sustratos y catalizar la reacción. NAGK en su forma hexamérica presenta grandes cambios conformacionales debido a la interacción con el regulador alostérico. El enzima carbamato quinasa (CK) presenta un subdominio móvil que abre y cierra el centro activo. UMP quinasa (UMPK) también muestra importantes cambios conformacionales asociados a la regulación alostérica, pero de carácter distinto al de NAGK. De UMPK es importante destacar que aunque el plegamiento de la subunidad monomérica es muy similar al del resto de miembros de la familia, el ensamblaje de los monómeros en dímeros es diferente, pero se desconoce la funcionalidad de esta diferencia estructural. Tal y como puede apreciarse, esta familia presenta enzimas con distintos grados de oligomerización y, por lo tanto, proporciona un marco adecuado para estudiar como la estructura oligomérica añade complejidad a la dinámica intrínseca de las subunidades permitiendo la función biológica.

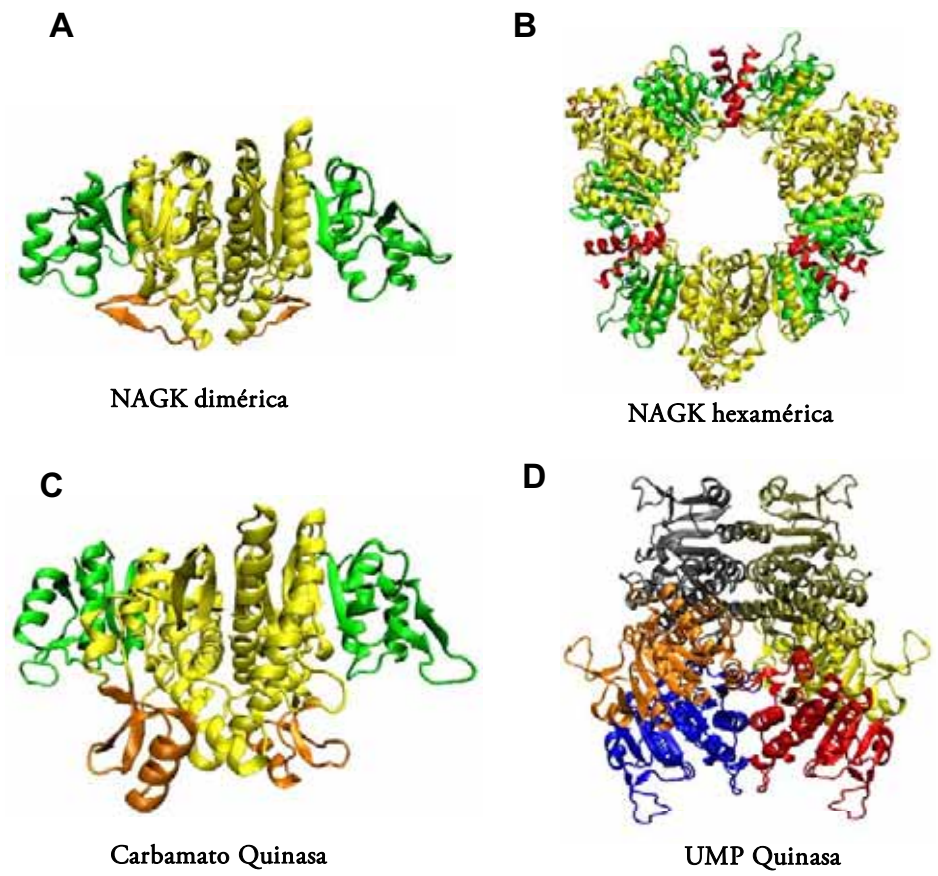


Figura 5. Quinasas de aminoácido. (A) NAGK dimérica de *E.Coli*. Los dominios flexibles corresponden a las zonas en verde y naranja, (B) NAGK hexamérica de *T. Maritima*. La estructura se corresponde a un trímero de NAGK diméricas, que interaccionan a través de una hélice N-terminal en rojo (C) Carbamato quinasa de *P. Furiosus.* Presenta mismo criterio de coloración que en los panels A y B, (D) UMP quinasa de *E. Coli*. Cada una de las subunidades monoméricas presenta una coloración diferente.

En esta tesis se han estudiado los cambios conformacionales de gran amplitud de los miembros de esta familia con un doble objetivo: (1) evaluar el grado de conservación de la dinámica lenta de estos enzimas y (2) analizar los efectos del ensamblaje oligomérico en la dinámica asociada a la unión de ligando. Para estudiar esta dinámica de gran amplitud se han utilizado los modelos de red elástica (*Elastic Network Models, ENMs*) desarrollados por Bahar y colaboradores, que hacen un análisis de modos normales sobre una descripción muy simplificada de la proteína. Los modos de más baja frecuencia que se derivan de este análisis son robustos respecto al plegamiento de la proteína y permiten caracterizar las direcciones de movimiento más accesibles.

NAGK como paradigma de movimientos de gran amplitud de la familia de las quinasas de aminoácido

Teniendo en cuenta que la forma dimérica de NAGK (*Ec*NAGK) es el paradigma estructural de la familia AAK, se ha estudiado la dinámica de gran amplitud de este enzima y comparado con otros miembros de la familia: carbamato quinasa, UMP quinasa y NAGK hexamérica. En primer lugar se analizaron los modos normales de más baja frecuencia de *Ec*NAGK. Estos modos de movimiento más accesibles se encuentra que describen en más de un 70% el cambio conformacional observado por cristalografía entre las formas abierta y cerrada del enzima. Esto demuestra que el cambio conformacional de gran amplitud necesario para la catálisis es intrínsecamente accesible por el propio plegamiento de la proteína y que el efecto del sustrato se limita a inducir ajustes en el centro activo para optimizar la unión.

En segundo lugar, para evaluar el grado de conservación de la dinámica se ha realizado una comparación de los modos normales entre los distintos enzimas mediante un método que permite estudiar la dinámica de una parte del sistema teniendo en cuenta las restricciones de movimiento que introduce el entorno. La aplicación de este esquema de trabajo no solo se limita a esta familia sino que es extensible a otras familias para encontrar similitudes en la dinámica. El resultado de este análisis muestra que los miembros de la familia AAK presentan unos patrones dinámicos como resultado de la similitud en la estructura (Figura 6). El grado de conservación de los modos normales más accesibles es elevado confirmando la hipótesis propuesta por Rubio y colaboradores acerca del mecanismo compartido.

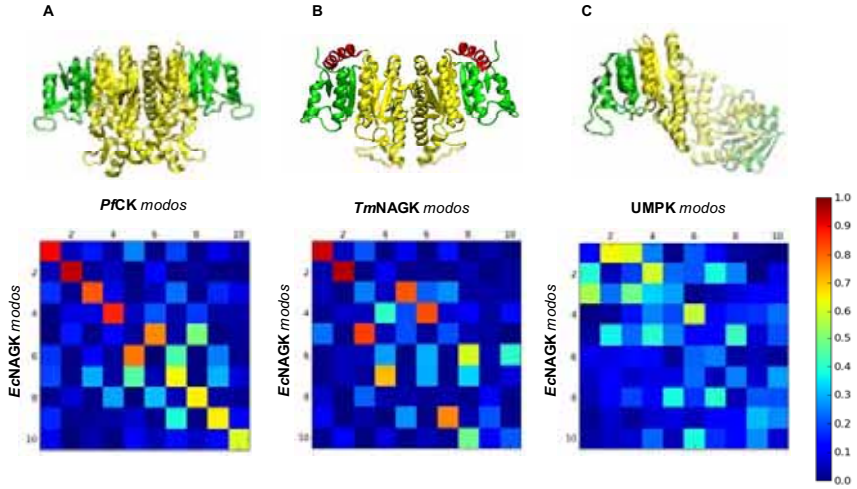


Figura 6. Comparación de los modos normales de *Ec*NAGK con otros miembros de la familia. (A) Carbamato quinasa, (B) NAGK hexamérica (subunidad dimérica) y (C) UMPK (subunidad dimérica). Se representan los productos escalares de pares de modos en una matriz.

La presentación detallada de los resultados de este estudio se encuentra en: *On the conservation of the slow conformational dynamics within the Amino Acid Kinase family: NAGK the paradigm* (2010) PLoS Comput. Biol., 6:e1000738.

Efectos de oligomerización en la dinámica de gran amplitud

Se ha utilizado el esquema de trabajo desarrollado en el estudio previo para analizar los efectos de la oligomerización en la dinámica de las subunidades. Esto implica tomar una subunidad como subsistema y el resto del oligómero como entorno para el cálculo de modos normales. La observación general que se extrae de este estudio es que la oligomerización proporciona nuevos modos de movimiento cooperativos que explotan la dinámica intrínseca de las subunidades.

En primer lugar, se han comparado para cada enzima (NAGK dimérica, CK y la NAGK hexamérica) los modos normales de baja frecuencia de las subunidades aisladas con los que se obtienen teniendo en cuenta el entorno del oligómero. Se encuentra que los modos de gran amplitud intrínsecamente accesibles a las subunidades se conservan en gran medida en el estado oligómerico. En la figura 7 se muestra una comparativa de estos modos para los enzimas *Ec*NAGK y *Pf*CK.

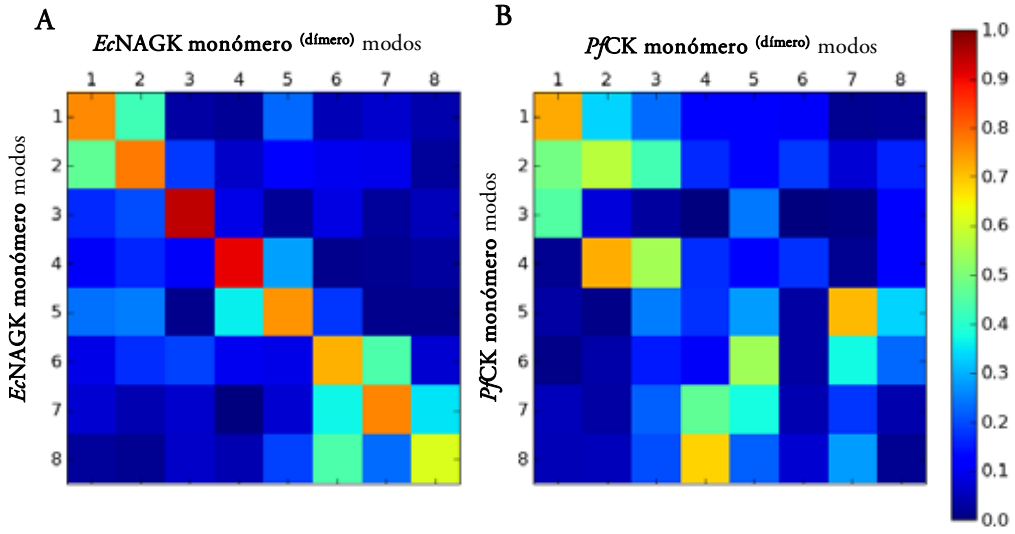


Figura 7. Comparación de los modos normales del monómero aislado con los modos obtenidos incluyendo el efecto de entorno debido a la otra subunidad monomérica del enzima. (A) *Ec*NAGK (B) *Pf*CK

En segundo lugar, se observa que hay modos de movimiento que emergen del diseño estructural de la interfaz entre subunidades. Por ejemplo, el cambio conformacional asociado a la regulación alostérica de NAGK hexamérica, que implica movimientos de cuerpo rígido por parte de las subunidades diméricas, está determinado por la estructura de la interfaz entre los dímeros que componen el hexámero. Un cálculo de modos normales

de los dos tipos de subunidades diméricas (AB y AF) que componen la proteína muestra que la interfaz AF es la que capacita a la estructura para los movimientos de cuerpo rígido de las subunidades. Dada la relevancia de esta interfaz, se ha evaluado la comunicación entre las subunidades a través de esta interfaz mostrando. Este análisis identifica a varios residuos claves en la cooperatividad de la estructura.

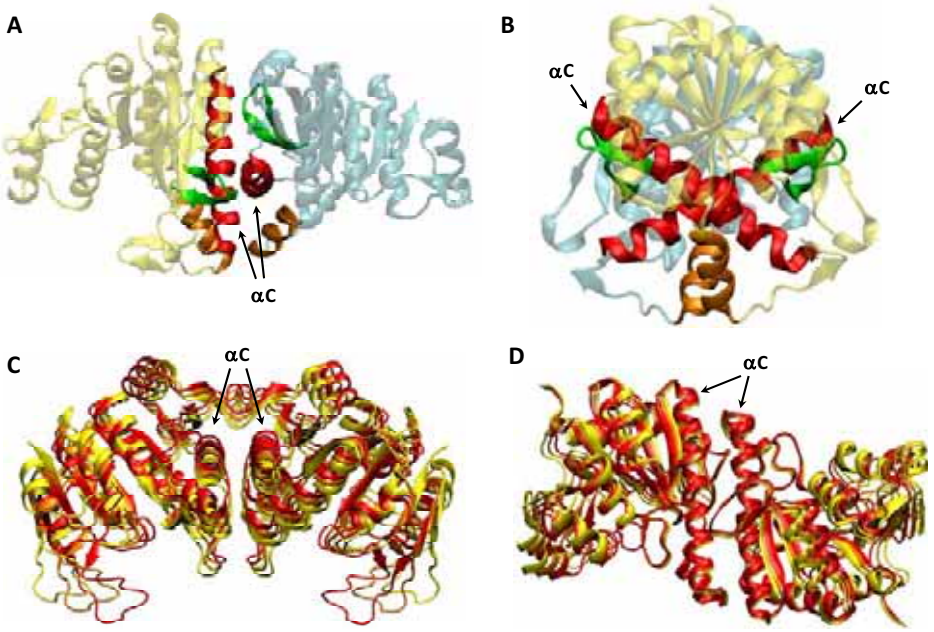


Figura 8. Comparación entre los dimeros de *Ec*NAGK y UMPK. (A) y (B) Dos vistas perpendiculares de *Ec*NAGK. Las hélices αC (en rojo) son los principales elementos de la interfaz de las dos subunidades y adoptan una disposición cruzada. (C) y (D) Dos vistas perpendiculares de la subunidad dimérica de UMPK. Las hélices αC también son los principales elementos de la interfaz de las dos subunidades y adoptan una disposición paralela. Diferentes deformaciones a lo largo del modo de más baja frecuencia se representan en rojo, naranja y amarillo.

Otro ejemplo interesante que se muestra es el de UMPK, que presenta un ensamblaje de las subunidades diméricas que es diferente al del resto de los miembros de la familia (Figura 8). Estudiando los modos de baja frecuencia de la subunidad dimérica se encuentra que esta diferencia en el diseño de la interfaz capacita a la estructura para realizar movimientos de cuerpo rígido de las subunidades monoméricas que no están permitidos con el otro tipo de diseño de la interfaz que presenta la familia. Este tipo de movimiento se ha observado a través de estructuras cristalográficas en presencia y en ausencia de GTP (regulador alostérico). Esto sugiere que el diseño de la interfaz está muy relacionado con el mecanismo de regulación alostérica.

Una presentación detallada de los resultados obtenidos en este estudio se encuentra en: *Changes in dynamics upon oligomerization regulate substrate binding and allostery in Amino Acid Kinase family members* (2011) PLoS Comput. Biol., 7: e1002201.

Termostabilidad de enzimas

Hasta este punto se han considerado propiedades reactivas y dinámicas de enzimas. No obstante, un aspecto importante de los enzimas es que han evolucionado para realizar su función biológica bajo unas condiciones ambientales específicas. De gran interés son los enzimas que trabajan a altas temperaturas (termófilas) por sus posibles aplicaciones biotecnológicas. Para comprender mejor los mecanismos de estabilidad de las proteínas termófilas, en esta tesis se ha estudiado cómo la adaptación térmica puede determinar las propiedades dinámicas de un enzima.

Para comprender los mecanismos de termoestabilidad de proteínas termófilas suelen hacerse estudios comparativos con homólogos mesófilos, cuya actividad óptima tiene lugar a temperaturas más bajas. En general, se observa que las termófilas presentan una mayor proporción de amino ácidos cargados (Asp, Glu, Arg, Lys). Se piensa que esto puede estar relacionado con una mayor capacidad de hacer fuertes interacciones de puente salino entre amino ácidos de carga opuesta que proporciona robustez a la estructura proteica. Lo interesante de estas interacciones es que se intensifican al aumentar la temperatura. Para que se forme un puente salino la carga debe desolvatarse, lo cual tiene un coste energético. Al aumentar la temperatura, la constante dieléctrica del agua disminuye, ya que el incremento de movilidad reduce su capacidad de apantallar las cargas, y en consecuencia el coste de desolvatación baja favoreciendo la formación del puente salino. Otro aspecto diferencial entre ambos tipos de proteínas es que las termófilas suelen presentar estructuras más compactas y bucles (“loops”) más cortos en la superficie.

Respecto a las diferencias dinámicas de ambas proteínas hay más controversia, ya que distintas técnicas experimentales proporcionan información a distintas escalas de tiempo. Tradicionalmente se consideraba que las proteínas termófilos son más rígidas y que por eso requieren temperaturas altas para ser activas y llegar a la desnaturalización. No obstante, hay experimentos que proponen una alternativa a esta visión. Zaccai y colaboradores, basados en medidas de dispersión de neutrones, establecieron un nuevo paradigma en la relación entre la termoestabilidad y la flexibilidad de proteínas. A partir del experimento, los autores sugirieron que la adaptación de las proteínas termófilas a las altas temperaturas se encuentra en la menor sensibilidad de la flexibilidad interna (a escalas de tiempo de picosegundos) frente a cambios de temperatura. Sorprendentemente, también encontraron que, a baja temperatura, el enzima termófilo es más flexible y menos activo. Con el objetivo de entender el mecanismo de este diferente comportamiento térmico de la flexibilidad, la última parte de esta tesis se ha focalizado en racionalizar el fundamento teórico de los resultados de este experimento y comprender mejor las diferencias en las propiedades dinámicas de los dos enzimas homológos.

Simulación de las propiedades dinámicas analizadas por dispersión de neutrones

Los experimentos de dispersión de neutrones proporcionan información dinámica a escalas de cortas de tiempo (desde picosegundos hasta nanosegundos). En un experimento de este tipo se mide el intercambio de momento y energía de los neutrones, como resultado de la colisión con la proteína, y se obtiene el factor dinámico de estructura ($S(Q,\omega)$). A partir de esta magnitud se puede extraer información dinámica de la proteína: el desplazamiento cuadrático promedio $\langle u^2 \rangle$, que cuantifica la amplitud de los movimientos accesibles a una escala de tiempo determinada por el instrumento.

En el experimento de Zaccai y colaboradores midieron el MSD de dos enzimas homólogos adaptados a diferentes temperaturas: malato deshidrogenasa (100ºC) y lactato deshidrogenasa (30ºC). El rango de temperaturas estudiado fue 280-320 K. Como los valores de $\langle u^2 \rangle$ medidos por dispersión de neutrones pueden incluir contribuciones tanto de la dinámica intramolecular como de la difusión, se han realizado simulaciones de dinámica molecular (MD) para analizar la flexibilidad interna y, por otro lado, de dinámica Browniana (BD) parar el movimiento difusivo en disolución. Cabe señalar que la BD se ha llevado a cabo para una caja de 1000 moléculas con el objetivo de incorporar los efectos de volumen excluído presentes en el experimento debido a las altas concentraciones que se utilizan.

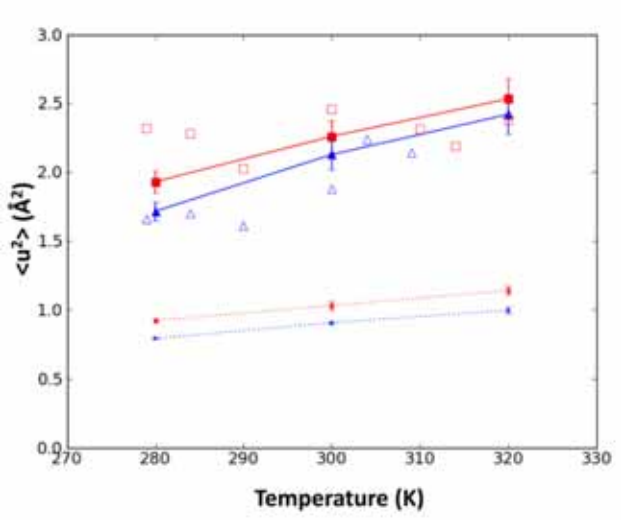


Figura 8. Comparación entre valores de $\langle u^2 \rangle$ experimentales [8] y calculados para el enzima termófilo (cuadrados rojos) y el mesófilo (triángulos azules) a las tres temperaturas estudiadas (280, 300 y 320 K). Los valores experimentales se representan con los símbolos vacíos, los valores simulados de $\langle u^2 \rangle$ para la dinámica intramolecular con líneas discontinuas y los $\langle u^2 \rangle$ que incluyen la difusión con líneas continuas.

Los resultados de las simulaciones concuerdan con las observaciones experimentales, pero sólo si se tiene en cuenta la contribución de la difusión (Figura 8). Concretamente la difusión traslacional y rotacional contribuye en un ~50% al $\langle u^2 \rangle$ medido experimentalmente. Además la diferente dependencia del $\langle u^2 \rangle$ con la temperatura proviene más por diferencias en el comportamiento térmico de la difusión (no de la flexibilidad interna) de las dos proteínas. De esta manera la interpretación original del experimento, basada exclusivamente en la flexibilidad interna, necesita ser reconsiderada. El acuerdo cualitativo que presentan las simulaciones de dinámica interna con el experimento es el hecho de que el enzima termófilo presenta valores de $\langle u^2 \rangle$ mayores que los del mesófilo, pero un 50% por debajo del valor experimental.

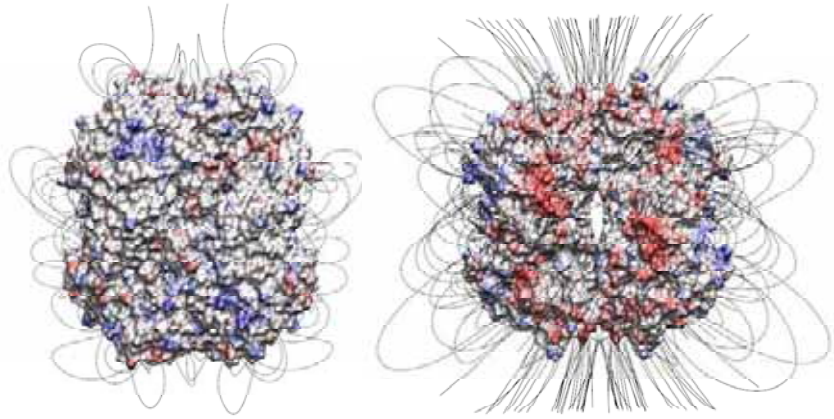


Figura 9. Potencial electrostático del enzima mesófilo y el termófilo. Representación de las estructuras molecular coloreadas según el potencial. Las regiones en rojo y azul corresponden a zonas de densidad de carga positiva y negativa respectivamente

Este estudio tiene implicaciones tanto en la interpretación de experimentos de dispersión de neutrones que se realicen en solución. Además, estos resultados muestran evidencia de que bajo condiciones de hacinamiento, como *in vivo*, enzimas termófilos y mesófilos tienen propiedades de difusión con un comportamiento térmico diferente debido a la diferente composición de su superficie. El potencial electrostático en la superficie de la proteína termófila es más intenso (Figura 9), debido a la mayor cantidad de residuos cargados, lo cual conlleva interacciones electrostáticas más fuertes entre proteínas que influyen en el comportamiento de difusión. Estas interacciones electrostáticas tienden a fortalecerse con aumentos de temperatura debido a que la constante dieléctrica del agua disminuye y en consecuencia el coste de desolvatar un residuo cargado para formar un puente salino disminuye favoreciéndose la formación de esta interacción. De esta manera, el incremento natural de la difusión con la temperatura se compensa parcialmente por un

aumento de las interacciones atractivas entre proteínas que en última instancia reduce su difusión.

Este estudio abre nuevas oportunidades para realizar nuevos estudios comparativos de la difusión de proteínas termófilas y mesófilas para determinar si se trata de una tendencia general y, si es así, que consecuencias biológicas esto podría tener.

Una descripción detallada de los resultados de este estudio se encuentra en: *Crowding induces differences in the diffusion of thermophilic and mesophilic proteins: a new look at neutron scattering results* (2011) Biophys. J. 101: 2782-2789.

Comparativa de la dinámica intramolecular del par termófilo-mesófilo

Tras haber separado las contribuciones de dinámica interna y difusión en los resultados de dispersión de neutrones, el siguiente paso fue realizar un análisis comparativo más detallado respecto a las diferencias en flexibilidad interna de los dos enzimas. En primer lugar, estudiamos movimientos en un rango de escalas de tiempo más amplio (20 ps – 10 ns) que en el experimento (100 ps). La dependencia con la temperatura de esta flexibilidad, expresada como MSD, se observa que es más fuerte en el enzima mesófilo a medida que se aumenta la escala de tiempo (Figura 10). De esta manera, para las escalas cortas todavía no se aprecian diferencias significativas en esta dependencia, pero sí a escalas más largas (>300 ps). Cabe señalar que aunque en el estudio anterior reinterpretamos el experimento por la contribución difusiva, este estudio apoya la idea originalmente propuesta por Zaccai y colaboradores a escalas de tiempo más largas.

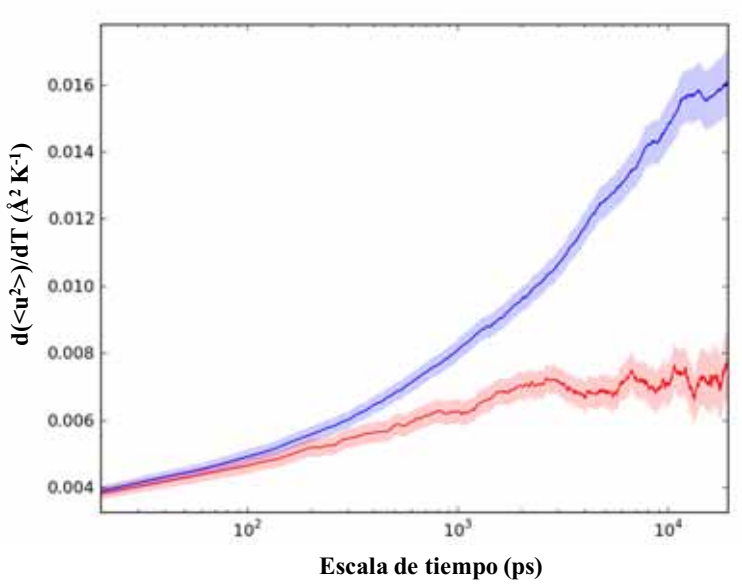


Figura 10. Dependencia de la flexibilidad interna con la temperatura a distintas escalas de tiempo para el enzima termófilo (rojo) y el mesófilo (azul)

La diferente dependencia con la temperatura la atribuimos a dos factores importantes. Primero, diferencias en la estructura secundaria. Se observa que las zonas del enzima mesófilo con una flexibilidad más dependiente de la temperatura se corresponden con bucles ("loops") y otras zonas desestructuradas como los extremos de cadena. Generalmente los enzimas mesófilos presentan bucles más largos que sus homólogos termófilos resultando en estructuras menos compactas. Segundo, las interacciones de puente salino que se forman entre amino ácidos de carga opuesta tienden a fortalecerse con

la temperatura. De esta forma, el mayor contenido de residuos cargados en el enzima termófilo proporciona más posibilidades de formar este tipo de interacciones. Se observa como los puentes salinos, además, tienen más tendencia a fortalecerse con la temperatura en el enzima termófilo. Esto provoca que haya regiones que pueden reducir su movilidad a temperaturas más altas.

Otra diferencia importante es que el enzima mesófilo presenta una mayor heterogeneidad dinámica. Esto implica que presenta una mayor diferencia de movilidad entre las zonas más flexibles y las más rígidas, respecto al homólogo termófilo (una distribución más ancha de movilidades atómicas). En consecuencia, los movimientos del mesófilo son de carácter más local y los del termófilo más cooperativos, lo cual podría estar ligado a una mayor capacidad de disipar la energía térmica.

Una presentación detallada de los resultados de este estudio se encuentran en: *Dynamic fingerprints of protein thermostability revealed by long molecular dynamics.* Enviado a *J. Chem. Theory Comput.*