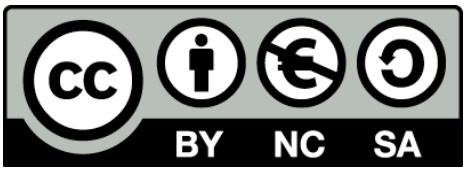
Protein Flexibility: From local to global motions. A computational study

Melchor Sánchez Martínez



Programa de doctorat en química teòrica i computacional

Protein Flexibility: From local to global motions. A computational study

Melchor Sánchez Martínez

Ramón Crehuet Simón[1]
Director de tesis

Jaime Rubio Martínez[2,3]
Tutor de tesis

1. Departament de Química Biològica i Modelització Molecular, Institut de Química Avançada de Catalunya (IQAC-CSIC)
2. Institut de química teòrica i computacional (IQTC-UB)
3. Departament de química física de la Universitat de Barcelona

Agradecimientos

En primer lugar, hay que reconocer que es difícil agradecer a todas las personas que de una u otra manera colaboran e influyen en el desarrollo de una tesis doctoral en unas pocas líneas. Así que si alguien siente que falta, espero sepa disculparme y entienda que de todas las personas con las que uno comparte vivencias en estos años, algunas están más directamente relacionadas que otras. De antemano pido disculpas.

Paradójicamente, este apartado, aunque el último que se escribe, es por lo general el primero en ser leído. No es un requerimiento académico de formato o contenido como si lo son otros (que como se suele decir por aquí … “deu n´hi do”) pero dada su “popularidad” no lo obviaré.

Si es cuestión de agradecer, la primera persona que me viene a la cabeza es Ramón, mi director de tesis, referente del trabajo bien hecho y el rigor científico. Sin él y su inagotable curiosidad científica habría sido imposible llevar a cabo esta tesis. Quiero agradecerle la oportunidad que me dio confiando en mi cuando venía de un campo relacionado aunque más o menos alejado de la química computacional, así como la paciencia y dedicación que ha tenido durante todos estos años conmigo. Gracias por todos los valores y conocimientos, humanos y científicos, que me has transmitido.

A Jaime Rubio, mi tutor de tesis, por la buena disposición que siempre ha tenido para ayudarme con los aspectos burocráticos de la tesis.

Quiero agradecer a Josep y Santiago de quienes, aún sin haber trabajado directamente con ellos, he aprendido mucho. Su pasión por la ciencia y su capacidad de trabajo ha sido un ejemplo y a la vez una motivación durante estos años.

To Martin J. Field for his warm welcome in Grenoble as well as all his scientific guidance. I also want to thank Martin our interesting discussions about football and science in general. Undoubtly my scientific career has advanced due to him. Many thanks. To my collegues at the IBS. Anirban thanks for interesting scientific discussions and all that I learned from you, at both scientific and human level. Thanks also for showed me Grenoble. All the best in your new married life in Canada. Nicholus thanks to make my last time in Grenoble better and for sure for climb with me to the Bastille. All the best.

To Sandipan for his nice welcome in Jülich and his teaching regarding Monte Carlo. My thesis has advanced a lot due to this. I also want to thank Sandipan for his good predisposition to help me with any thing scientific or not and, also, for all our discussion about scientific topics and science in general. Many thanks. Also to Olav and Jan for be a good example of scientific rigor and help me always that I needed it. And finally, also, to all the people of the department for the interesting discussions during the lunch time.

Quiero agradecer también a todos los compañeros del CSIC por los buenos momentos pasados y el gran ambiente que se ha respirado en el despacho, así como en los desayunos y comidas varias que hemos compartido. Una buena atmósfera laboral es clave para el desarrollo profesional en cualquier ámbito y más si cabe en la ciencia.

A Javi, el primero en marchar, por su agradable compañía y buenos consejos. A Miquel por la alegría que aportaba y todo lo que aprendí de él. A nivel estrictamente científico quizás menos ya que nunca trabajamos juntos, pero a nivel de toda la atmósfera de ”componente” variado que rodea la ciencia, mucho. Gracias a él se puede decir que soy un experto. Me alegro mucho de que por fin hayas conseguido esa posición por la que tanto has peleado y trabajado. Te mereces la Ikerbasque. A

Quique por su alegría, capacidad de trabajo, rigor científico y humanidad. De él he aprendido muchísimo a nivel científico y humano. Espero que acabes de exprimir la Marie Curie en Seattle y la vuelta a Barcelona sea lo mas provechosa posible. A Jelisa por toda la alegría que aportó durante el año que estuvo con nosotros, así como su inestimable ayuda para sobrellevar las largas horas de trabajo los dos solos en el despacho y todas las comidas compartidas tanto en el CSIC como fuera de él. Por último a Susana, la última en llegar pero no menos importante. Gracias por hacer que este último tramo de la tesis, probablemente el más duro y estresante, haya sido más llevadero. Gracias por toda tu alegría y compartir siempre tus inquietudes. Como dice Josep, la conciencia social del grupo. Un ejemplo de lucha por unos ideales. Si hubiera más Susanas quizás algunas cosas cambiarían a mejor. Con tu gran capacidad de trabajo, estudio y aprendizaje harás una gran tesis.

En este recordatorio de gente no pueden faltar Alberto, Guillermo y Pesquera, que llevan ahí desde incluso antes de comenzar el doctorado: la época del máster de Biofísica. Han sido las personas que han estado siempre desde que llegué a Barcelona en 2009 y eso es de agradecer. Gracias por aguantarme siempre y estar dispuestos a ayudarme cuando lo he necesitado. En este grupo quiero incluir también a Max que aunque ahora ya de vuelta en Alemania, fue y sigue siendo un buen amigo con el que he compartido y espero seguir compartiendo en un futuro grandes momentos. Por supuesto también a Anna, Cristina y Katia por aguantar a estos y todos los buenos ratos que hemos pasado.

No hay que olvidar tampoco a Manolo y Raúl, la morralla de zona franca, dos incorporaciones de última hora pero con todas las horas en “El Prat” y últimamente en casa de Raúl, han hecho más llevadera esta tesis y la vida en general.

Tampoco me puedo olvidar de la gente del máster QTC. En especial Arnau, Gian y Oriol, que son los que más me han tenido que soportar, por las comidas en la UB y todos los ratos pasados. Sin olvidarme por supuesto de Jordi, Laia, los dos Marc, Mulet y el chiquitín, y Pablo. Nos vemos poco, casi siempre en congresos, pero siempre es un placer poder compartir un rato con vosotros. Sois unos cracks.

También quiero agradecer a todos los compañeros del máster de biofísica por hacer que mi primer año en Barcelona fuera un gran año aunque luego nos hayamos distanciando. Igualmente toda la gente de Vic merece una mención especial ya que quizás ahora la relación por unos motivos u otros no sea la misma y estemos algo más alejados, pero todo el tiempo pasado juntos ha servido para que hoy esté donde estoy y sea como soy. Alex, Ana, Clara, Gael, Guille, Laia, Laura, Marina, Iñigo y Fer, muchas gracias por haber estado ahí y ser como sois. Y hablando de Vic txeic! tú aunque solo sea por pesat y preguntar de vez en cuando como van las proteínas, también mereces que te de las gracias. Ya quedaremos para unos garbanzos o algo Flix!

Y en cuanto a amistades se refiere nunca se han de olvidar los orígenes. Quizás vuestra contribución no haya sido directa pero siempre habéis estado ahí desde Ribaforada, Tafalla, Zaragoza, Londres, Estrasburgo o Arabia Saudi de una forma u otra: Aitor, Fran, Ibai, Iván, Paul, muchas gracias.

A toda la gente de Sant Pere y Cubelles por haber estado ahí siempre y haber compartido tantas comidas, cenas y demás eventos. Adri, Carla, Cuchi, Fer, Nelson, Pato y Sara. Gracias por ser como son.

Y como no agradecer a mi familia (a los que están y a los que se fueron) por su apoyo incondicional aunque no tengan ni la más mínima idea de lo que hago. A mi familia de aquí y a la de Argentina. Por que aunque a veces no os acordéis de mi nombre... no puedo consideraros de otra manera. Gracias por haberme acogido como lo habéis hecho, por estar ahí y preocuparos por mi

tesis aunque como mi familia de acá, en eso coinciden, no sepan casi ni de que va. Muchas gracias. En especial a vos Tricky que no solo me bancás en Argentina sino también lo hiciste acá en Sant Pere. Julietita, Roberto y Aurora muchísimas gracias por ser como son conmigo. Y por supuesto a vos Norma que siempre estuviste ahí. Gracias de corazón a TODA mi familia.

Especialmente quiero agradecer a mis padres, un ejemplo a seguir y de quienes me siento orgulloso. Os agradezco vuestro cariño y apoyo sin reservas. También vuestra insistencia en que no dejara de estudiar (porque aunque ahora tengo una tesis escrita estuve apunto de dejarlo en bachillerato...). Sin vosotros esto no hubiera sido posible de ninguna manera. Gracias por haberme dado siempre todo y haber hecho lo imposible siempre que os lo he pedido. Aunque a veces también os extralimitéis y hagáis más de la cuenta...de corazón muchísimas gracias por todo.

También a Roberto, mi hermano. La verdad que me fui a la Universidad cuando aún eras un crío (que por cierto, enano sigues siendo) pero aún así, a tu manera, siempre te has preocupado por mi (aunque te cueste reconocerlo...) y ya sabes que el sentimiento es mutuo casco feo. Gracias por estar ahí y alegrarme los días cuando vuelvo por casa. Y quien sabe, ahora que estás metido en esto de la química, quizás algún día seas tú el que escriba unos agradecimientos... Eres medio desastre pero aún así, no cambies y sigue intentando hacer siempre lo que te guste.

Mención aparte mereces tú Marce. Tendría que escribir una tesis entera para agradecerte todo lo que me aportas. Sé que no ha sido fácil aguantar todo un doctorado sin entender muchas veces lo que hacía, ni los horarios, ni todo el tiempo invertido... solo puedo darte las gracias. Eres uno de los motivos, sino la razón principal, por lo que cada día me levanto con ganas de hacer cosas y trato de esforzarme y mejorar día a día. Gracias por iluminar mi camino. Espero estar siempre a tu lado y compartir muchos más momentos (menos estresantes que una tesis) contigo.

En general a tod@s y cada un@ de vosotros, lo mejor en vuestras vidas a todos los niveles.

Y para terminar, quiero responder una pregunta recurrente dentro del grupo... Sí Josep, se puede decir que durante estos años he sido feliz. Gracias a todos por hacerlo posible.

Contents

Chapter 1

Introduction

1.1 Protein Dynamics

Proteins are flexible entities, and thus move. Its function is closely related to flexibility. To carry out any function is necessary a conformational change. As protein motions imply an exchange of conformations, protein dynamics is also known as Protein Conformational Dynamics. The fluctuations between the different proteic configurations can be classified according to the length-scale, the time-scale and the amplitude and directionality of them [1]. In agreement with the length-scale the movement could be a local movement, involving only the rearrangement of a few amino-acid side chains or even backbone, or it may be a large, global movement, modulating the allostery or the conformational transitions, and even involve folding of the entire protein [2,3]. In line with the time-scale these motions are divided into slow and fast dynamics, and regarding their amplitude and directionality, could be distinguish between large and small amplitude protein motions. Generally local motions are also fast and small amplitude movements whereas global motions are associated with slow and large amplitude movements [1]. This classification can lead to misunderstandings because the frequency of the local motions may be low (rare events) such as chemical reactions. Thus can be argued that they are slow motions because need some time to occur, but they are nonetheless fast.

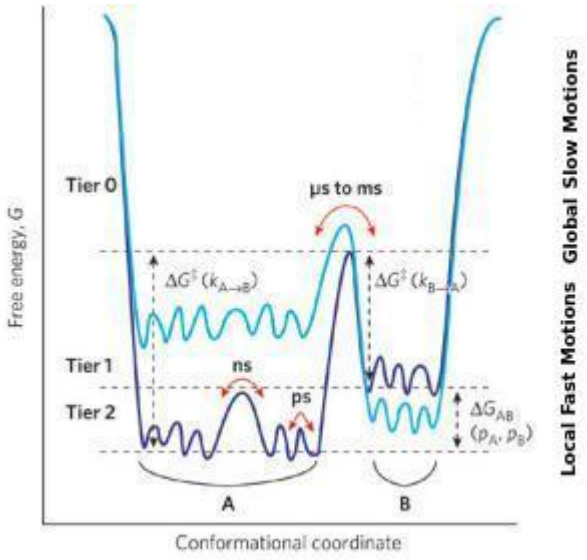


Figure 1.1: One dimensional cross section of the multidimensional energy landscape of a protein showing the hierarchy of protein dynamics and the energy barriers. Light and dark blue lines represents two different hypothetical energy landscapes of a system. Local slow motions (Tier-0) and dark blue Global Fast motions (Tier-1 and 2). Lower tiers describe faster fluctuations between a large number of closely related substates within each tier-0 state. Adapted from Henzler-Widman & Kern [1].

All these motions encompassed into protein dynamics are governed by the features of the underlying energy landscape. To fully describe a protein, a multidimensional and rugged energy landscape defining the relative probabilities of the conformational states (thermodynamics) and the energy barriers between them (kinetics) is required. To understand proteins in action, the fourth dimension, time, must be added [1,4].

1.1.1 Global Motions

The protein dynamics at this level define fluctuations between kinetically distincts states separated by energy barriers of several k_BT (k_B being the Boltzmann constant and T the temperature). Their time-scale corresponds to microseconds (μ) and slower at physiological conditions. This is the reason why they are called slow motions. Typically these are large amplitude collective movements between a relatively small number of states, involving for instance domain motions. Within each state transitions between closely related conformations constitute the local fast motions (1.1.2) [1,3,5].

Dynamics on this time-scale is very relevant and receives a lot of attention because many biological processes involving conformational transitions such as

substrate binding, allosteric events, enzymatic dynamics and even disorder to order transitions take place at this time-scale.

1.1.2 Local Motions

The protein dynamics at this level define fluctuations within the picosecond to nanosecond and even femtosecond time-scale, defined as fast motions. In contrast to slow movements they represent a large ensemble of structurally similar states, fluctuating as small amplitude motions separated by an energy barrier of less than $1k_BT$ at physiological temperature. However the chemical reaction are an exception. They present barriers of several k_BT although they are fast and local, and thus the frequency is low. This is the reason why the aforementioned controversy regarding the local motions exist. They can be fast (in relation with the time that need to take place) and also slow (related to the frequency with they occurred).

We can distinguish between different processes depending on the time-scale, such as loop motions at the nanosecond time-scale, or local atomic fluctuations on the picosecond (ps) time-scale. Chemical reactions (bond cleavage) as well as bond vibrations take place at the femtosecond (fs) time-scale[1,3,5].

1.1.3 Experimental and Computational techniques to study Protein Dynamics

1.1.3.1 Experimental techniques

The flexibility of proteins has been widely studied both experimentally and computationally. There exist a wide range of experimental techniques suitable to explore different time scales and resolutions (Fig. 1.2) (For a book review see Livesey[6]).

X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, cryo-electron microscopy and Small X-ray Angle Scattering (SAXS) are able to produce atomic resolution or near atomic-resolution snapshots of global motions. Furthermore X-ray diffraction, NMR relaxation dispersion or low-resolution spectroscopic methods, can provide a picture of the local motions. Thus, the combination of different techniques allow the multi-scale exploration of protein dynamics.

NMR spectroscopy is a powerful technique that allows proteic structural and kinetic determination. In NMR the relaxation of the nuclei after excitations with the magnetic field allows to span the atomic resolution detection of conformational transitions from picoseconds to seconds[1]. Within the microsecond to millisecond

and second time-scale this technique is able to capture the conformational transitions of the biochemical process that take place at these time ranges by measuring the backbone chemical shift assignments and fully determine the distribution of conformations using the residual dipolar couplings(RDCs) [7,8]. Based on the information given by N-H couplings, the most informative ones, the secondary structure population could be determined [9–11]. Furthermore have been shown that RDCs are able to capture the structural fluctuation at the nanosecond to microsecond time-scale [12].

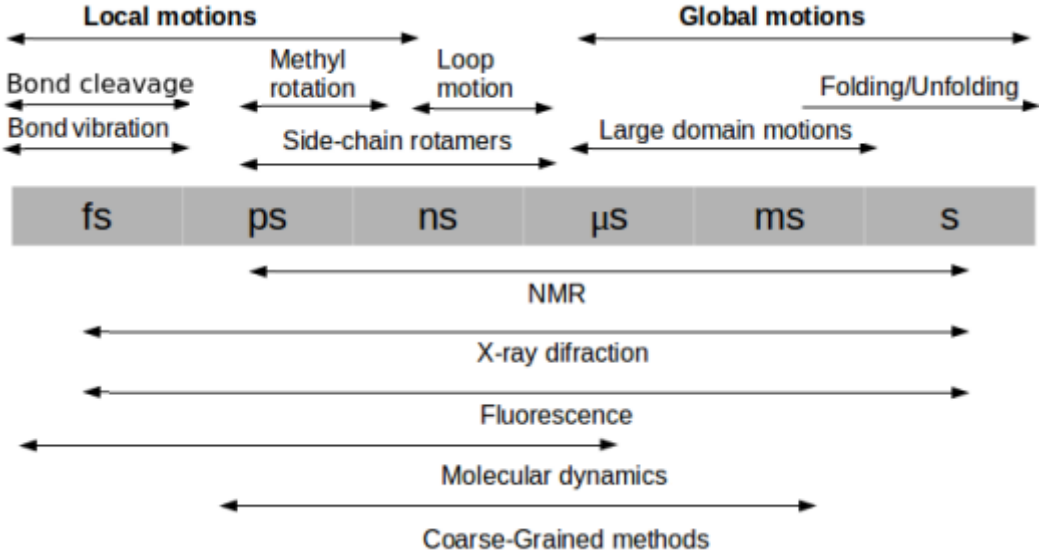


Figure 1.2: Time scales of dynamical events in proteins and techniques sensitive to different time scales, ranging from femtoseconds (fs) to seconds (s). Adapted from Henzler-Widman & Kern [1].

The family of X-ray methods that encloses some techniques as X-ray crystallography or SAXS, constitutes a helpful set of techniques to study protein dynamics. For high resolution X-ray crystallography a homogeneous crystal is required, being the substates trapped into it by biochemical techniques as site-specific mutagenesis or substrate analogs. Alternatively the reaction could be synchronized across the entire crystal usign 'cryo-trapping' techniques [13]. This strategies are nicely exemplified by the the cytochrome P450 enzymatic cycle intermediates characterization [14]. This homogeneous crystal requirement could be alleviated employing cryo-electron microscopy or SAXS techniques where the structural ensemble determination could be made but loosing resolution. Furthermore these methods can not characterize the kinetic information, as for example NMR spectroscopy can do [1].

However there are some X-ray crystallography variants aimed to account kinetics effects such as the so-called temperature-dependent X-ray macromolecular

crystallography[15]. Usually routine data are collected close to 100K in order to mitigate radiation damage (see section 1.3.1), however the usage of another temperature range, for instance from 15K to room temperature, can provide both dynamical and structural insights. At room-temperature the macromolecules are active and move but at cryo-temperatures macromolecular motions are very slow or even null and biological activity is impaired. Varying the temperature of the crystal between 100K (or lower) and room temperature one can turn the proteic activity on and off and subsequently trap functional intermediate states that then could be characterized structurally by crystals[15].

X-ray diffraction data not only contain information about the average 3D structure of a protein, but also about the conformational distribution around this state. The B-factors (also known as the temperature factor and Debye-Waller factor) are able to measure this mean-square atomic displacement. However a recent paper invite caution in their interpretation, because they showed that B-factors of some well-resolved atoms underestimate their actual values even sixfold[16].

1.1.3.2 Computational techniques

Experimental techniques provide a lot of useful data that computational techniques needed. Very often, if not always, the initial structure or information employed comes from experiments, but computation has advantages as it can describe dynamics completely and can characterize the full conformational energy landscape of proteins (although conformational substrates and the rates of interconversion can be located experimentally, the transition pathway at atomic resolution is out of reach[17]). The computational simulations can be used as a 'virtual microscope' to study processes or molecules in a cell that are not directly accessible in experiments[18]. Time to time the mixture of both 'worlds' is increasing, and the incorporation of experimental data into computational models is most usual.

Protein dynamics can be explored by all-atom (AA) simulations such as Molecular Dynamics (MD) describing 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, identifying the most relevant conformations characterizing conformational transitions as well as monitoring microscopic properties over time, allowing the prediction of equilibrium macroscopic properties of the system.

In general the description of dynamical events in the microseconds time scale or beyond is out of reach by conventional MD with current computational power. However there are efforts to span the achievable time-scale and in this direction, the impressive progress that have been done by Shaw and co-workers to

cover extremely longer time scales by using a special-purpose machine for MD, has to be highlighted. They recently reported the first 1-millisecond simulation for the bovine pancreatic trypsin inhibitor[19]. But this achievement although impressive is just one unusual case. Employing traditional all-atom simulation the microsecond-to-millisecond time scale is inaccessible, despite the increasing computing power. To overcome this problem some possible solutions are being addressed.

One of them is the ongoing development of more approximate methods to cover longer time scales as coarse-graining (CG) techniques[20–24]. This methodology reduces the number of degrees of freedom of the system accelerating the simulations and thus expanding the achievable time-scales, but at the cost of loosing certain information about the system. To overcome this limitation, hybrid AA/CG models are being developed[25–27].

Another alternative to accelerate the dynamic process and span the time-scale is apply an external force[28–30], like targeted[31,32] (TMD), steered[33,34] (SMD) or accelerated MD (AMD) methods[35]. When instead of an external force an empirical potential is added to the force field, the method is called Restrained MD[36] (RMD). The application of prior knowledge about the system, regarding the reaction coordinate or the end points, as happens with enhanced sampling methods as Metadynamics[37], Umbrella Sampling[38] or Transition Path sampling[39] is also a good and useful option.

Other interesting variant of MD is Replica Exchange (RE). Nowadays is common the use of generalized-ensemble techniques to speed up simulations of systems with rugged, multiple minima, free-energy landscapes[40]. Whereas classical MD is useful to study systems where barriers at room temperature are smaller or comparable to the thermal energy, RE is specially indicated for systems that have potential wells separated by relatively high barriers. RE allow systems of similar potential energies to sample conformations at different temperatures. By doing so, energy barriers on the potential energy surface might be overcome, enabling the exploration of new conformational space, improving the sampling by exchanging the temperature of non-interacting replicas of the system running at several temperatures.

A further alternative is Monte Carlo (MC). MC is a simulation method widely used to explore the protein energy landscape, quite different to MD. The particularity of MC is that instead of allowing the calculation of the different protein conformations along of the trajectory at a certain temperature, it constructs canonical ensembles generated randomly and accepted or rejected according to a certain criteria. This criteria usually is such that the probability to find the system in state i is proportional to the the Boltzmann weight $(exp(E_i/k_BT))$[18,41]. MC is

not suitable to calculate the kinetc properties, but it samples the configurational space much faster than MD.

To study protein dynamics regarding local motions (such as chemical reactions or electronic reorganizations) at single atomic detail (like reactivity or spectroscopic properties of enzymes), Quantum Mechanics (QM)/Molecular Mechanics (MM) methods are generally used[42–45]. QM methods allows the simulation of bond breaking/formation events such as proton or phosphoryl transfer reactions. On the other hand MM techniques are represented by parameterized force fields that allow the description of the energetics of the system in a fast way. The computational cost of simulating a whole protein with QM methods is unaffordable now a days. Thus QM/MM methods enable the simulation of the chemically active region (substrates and cofactors of the biochemical reaction studied) at QM high resolution level, combining it with an MM treatment for the surroundings (the full protein and solvent).

Depending on the kind of information to be extracted as well as the nature and the time-scale of the process to study, as happens with experimental techniques, some simulation techniques will be better suited than others. To fully characterize a proteic system multiscale simulations are needed. Nowadays going beyond of actual models and create hybrid AA/CG/QM methods is the goal (see[46] for a recent review of the state of the art).

More details about some of these techniques can be found in the Methods section (Chapter 3).

1.2 Enzyme Catalysis

Enzymes are the most proficient catalysts in nature. Enzymes are mainly globular proteins, i.e. proteins with a generally rounded, spherical shape. They make up the biological machinery, involved into the acceleration of each of the huge diversity of biochemical reactions, until reaching biologically relevant time-scales, which fall into the micro-second to second time range[47]. On the other hand most of non-catalyzed biochemical reactions take place in time scales ranging from minutes to millions of years[48,49]. In the absence of enzymes, the reaction in solution can be more than 10 orders of magnitude slower with respect to enzyme catalyzed reactions[3]. Conceptually, enzymes reduce the activation barrier between reactants and products in a (bio)chemical reaction, that is nothing but an energy barrier between reactants and products that has to be overcome by thermal activation of the reactants[47]. How enzymes are able to be so efficient reducing it, is an ongoing question.

An initially explanation is that the enzyme binds the substrate in the transition state stronger than in the ground state, but this not answer completely because opens a new question: how can the differential binding be accomplished? [50]

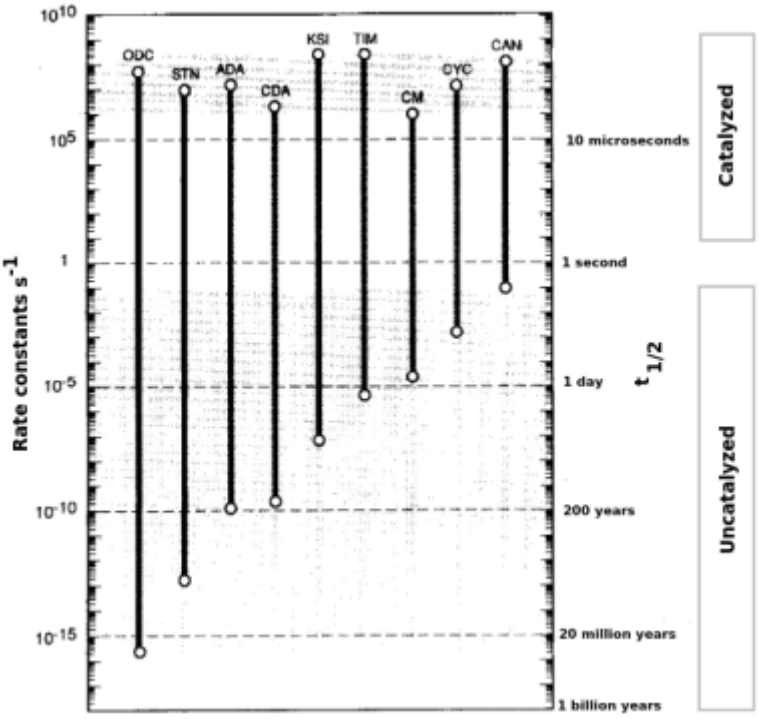


Figure 1.3: Representation of the rate-enhancement of some enzyme-catalyzed reactions (ODC, orotidine 5'-phosphate decarboxylase ; STN, staphylococal nuclease; ADA, caf intestinal adenosine deaminase; CDA, bacterial citidine deaminas; KSI, ketosteroid isomerase; CM, chorismate mutase; TIM, tryposphate isomerase; CYC, cyclophilin; CAN, carbonic anhydrase). Adapted from Radzicka & Wolfenden [48].

Unravelling the origin of such efficient catalysis has a tremendous potential, because enzymes are key targets for drug discovery [51,52], and they are increasingly used in industrial processes such as biofuels or detergent production [53–56].

Enzymes are the best example of pre-organization in nature. The orientation of the aminoacid functional groups located at the active site cavity allow to create the perfect scenario to carry out catalysis, efficiently binding the substrate and electrostatically stabilizing the reaction transition state [57]. The specific binding of the altered substrate in the transition state, is what enhances the catalytic power of the enzymes [49]. Other factors such as steric strain, desolvation or entropy have been proposed to play role in decreasing the activation energy, but its contribution is very small compare to the transition state stabilization (TSS) [3,57,58]. The specificity (the preferential and favoured binding of a certain substrate to a certain enzyme) is the hallmark of enzymes.

One of the first models addressed to understand enzymatic specificity was the well known *Lock and key* mechanism developed by Fischer[59]. This model treats the enzyme as an static entity that is complementary in shape to the substrate, but is deficient in the role of any protein motion. The amazing pictures of the precise pre-organization of the enzyme, obtained by X-ray crystallography, spread the view of enzymes as static entities. However, while the technique was evolved, the increasing number of X-ray structures together with NMR and spectroscopic studies[3] for free and bounded enzymes rapidly changed this paradigm. The experimental evidence showed that enzymes present different structures at each state of the reaction, being considered as deformable structures that require changes in conformation to bind substrates in the optimal position for efficient catalysis. Thanks to this new view, almost 70 years later of the Fischer model, Koshland proposed his *Induced fit* model[60] that considers some degree of plasticity in the enzyme and states that the enzyme conformation changes upon ligand binding. The enzymatic conformational rearrangements are 'induced' by the ligand. On the other hand the *conformational selection* model[61] state that the ligand could choose between a subset of pre-existed enzyme conformers. The latter model takes one step further toward the dynamic view emphasizing that enzymes are intrinsically flexible.

Nowadays, the dynamic nature of enzymes is commonly accepted and thus it is the subject of study by many experimental and computational groups. Several studies, mostly since the beginning of the XXI century, have highlighted the role of protein dynamics in the enzymatic function trying to characterize the vast range of dynamic events involved[58,62–70]. However the question of how exactly the protein motions help enzyme catalysis remains open.

1.2.1 Catalytic role of protein motions. Controversy between dynamics and catalysis

Despite the several studies that have related different kind motions to catalysis (see table 1.1), as has been stated above, their exact role on the catalytic cycle is still a matter of much debate[3,71]. The biochemical and biophysical community are inside an intense debate in that the major issue is what is understand by 'dynamical effects' in enzyme catalysis, that at the end is 'nothing but' disentangle whether (global) dynamics at the millisecond time-scale do catalyze the chemical step[69,72] or do not[73,74].

Motion	**Time-scale**	**Reference**
Conformational Change	ms-s	Adenylate Kinase[73,75]
Allosteric Transition	ms-s	Aspartate transcarbamylase[76], Aminoacid Kinase Family (AAK)[77]
Slow conformational sampling	ms-s	Flavin adenine dinucleotide[78], N-acetyl-l-Glutamate-Kinase[79]
Fast conformational sampling	ps-ms	Many enzymes[65,80–83]

Table 1.1: Representative experimental and computational examples of protein motion related to enzyme catalysis in the literature. Adapted from Nagel and Klinman[68]

Experimental and computational studies have proved that proteins have accessed to an ensemble of conformations encoded into their 3D structures, and thus that dynamics occur during a catalytic cycle is accepted by all scientists[17,84–86]. However some argue that the term 'dynamical effects' should only be used to asses deviations from Transition State Theory, which is an equilibrium theory, understanding them as a transfer of energy from a conformational coordinate to the chemical reaction coordinate in an inertial way[58]. Some experimental[87] and computational[58,73,88] studies suggest that indeed, these dynamical effects are small or negligible in enzymes. Other studies, that understand 'dyanmical effects' as any time-dependent change in atomic coordinates[66], suggest that fast dynamics are coupled to the enzymatic cycle[89] and there are computational studies which indicate that promoting vibrations are coupled to the catalytic reaction coordinate[90–92]. When slower conformational motions are present during the catalytic cycle, they can become the rate-limiting step. For some enzymes, NMR results seem to indicate that this is the case[75]. These motions are associated to ligand binding processes, although they seem to take place also for the free enzyme, pointing out to an intrinsic functional dynamics[93].

The problem seems to be somehow purely semantic and a clear definition of what 'dynamical effects' are, could be valuable and an step forward. A consensus approach from both experimental and computational points of view to define the role of protein motions in determining the outstanding efficiency of enzymes is required.

1.2.2 Aminoacid kinase family of enzyme (AAK). N-Acetyl-l-Glutanate-Kinase (NAGK). Phosphoryl transfer reactions in enzymes

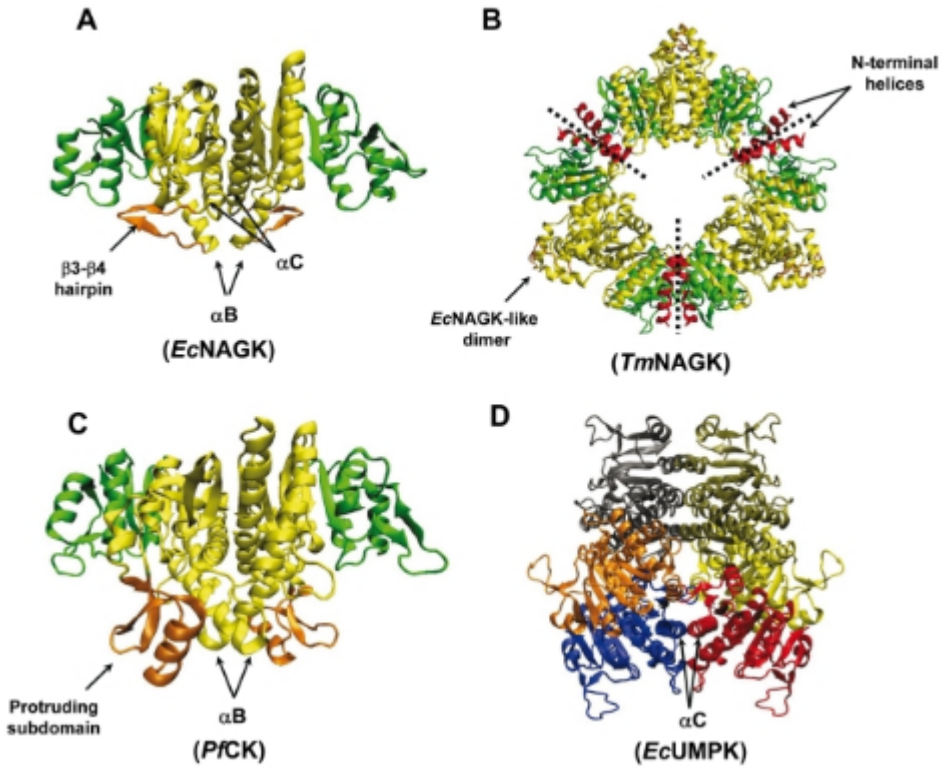


Figure 1.4: AAK family enzymes.(A) NAGK from Escherichia coli (EcNAGK), (B) NAGK from Thermotoga maritime (TmNAGK), (C) CK from Pyrococcus furiosus (PfCK), (D) UMPK from Escherichia coli (EcUMPK). Panels A, B and C show the ATP binding domains in green and N domains in yellow. The NAG-binding sites in EcNAGK (β3-β4 hairpin) and the CK-binding site in PfCK (protruding subdomain (PS) composed of the strand β5, helix D and hairpin β6-β7) are colored orange. The N-terminal helices of TmNAGK (red) interlink three EcNAGK-like dimers (delimited by dotted lines). This hexameric enzyme is indeed regarded as a trimer of EcNAGK-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. Figure reproduced, with permission, from Marcos et al. [77].

One of the most typical reactions that take place in enzymes are the phosphoryl transfer reactions, for that a very extensive literature exists mainly based on mechanistic investigations [94,95].

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. In mammals it proceeds thorough non-acetylated interme-

diates (Fig. 1.5). NAG phosphorylation by NAGK is the key controlling step of the byosinthethic route in many organisms, since NAGK is feedback inhibited by arginine, the end product. From a medical point of view, the fact that in mammalian cells the arginine biosynthesis proceeds through non-acetylated intermediates makes this route interesting. This interest is due to NAGK activity may be selectively inhibited and is a target for potential antibacterial drugs given the regulatory role of this enzyme in bacteria.

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, attracting more attention due to they are involved in processes like cancer.

Our computational studies showed that the large-amplitude motions of EcNAGK are intrinsic to the enzyme, and shared among other family members, thereby pointing to a common mechanism of action[79]. However not all NAGKs are arginine inhibited, for instance NAGK from Escherichia Coli (EcNAGK) is an example of an arginine-insensitive NAGK. EcNAGK is 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 biochemical and crystallographic studies[96–101].

The crystallographic studies[96,97,100] have given insights about the EcNAGK mechanisms of binding and catalysis. It is a homodimer of 258 residues in each monomeric subunit, being folded into an $\alpha\beta\alpha$ sandwich, without presenting cooperativity as have been shown by Kinetic studies[101]. Each subunit consists of a N domain that hosts the NAG binding site (NAG lid) and a C domain that binds ATP (see Figure 1.6). The phosphoryl transfer reaction takes place at the interface between these two domains.

X-ray structures of EcNAGK complexed with either ADP or with the inert ATP analogue AMPPNP (PDB codes 1GS5, 1OH9, 1OHA, 1OHB, and 2X2W) have active sites that are too narrow to let the substrates bind directly, whereas structures with an unoccupied ATP site (PDB code 2WXB) have a more open active site that does allow the substrates to enter. This suggests that the C-domain and NAG lid undergo a conformational closure that is likely to be triggered by nucleotide binding (see Figure 1.6), conforming a 'double drawbridge' gate[102], which is rare in active site entrances. Rubio and co-workers[97] hypothesized that in the closed form of EcNAGK the narrowness of the active site exerts a 'conformational compression' on the substrates (O-O distance in Figure 1.5) that favours catalysis. *In this thesis we have study the reactivity of the different crystal structures of EcNAGK, estimating the significance of 'conformational compression' and determining its contribution to the overall turnover of the enzyme.*

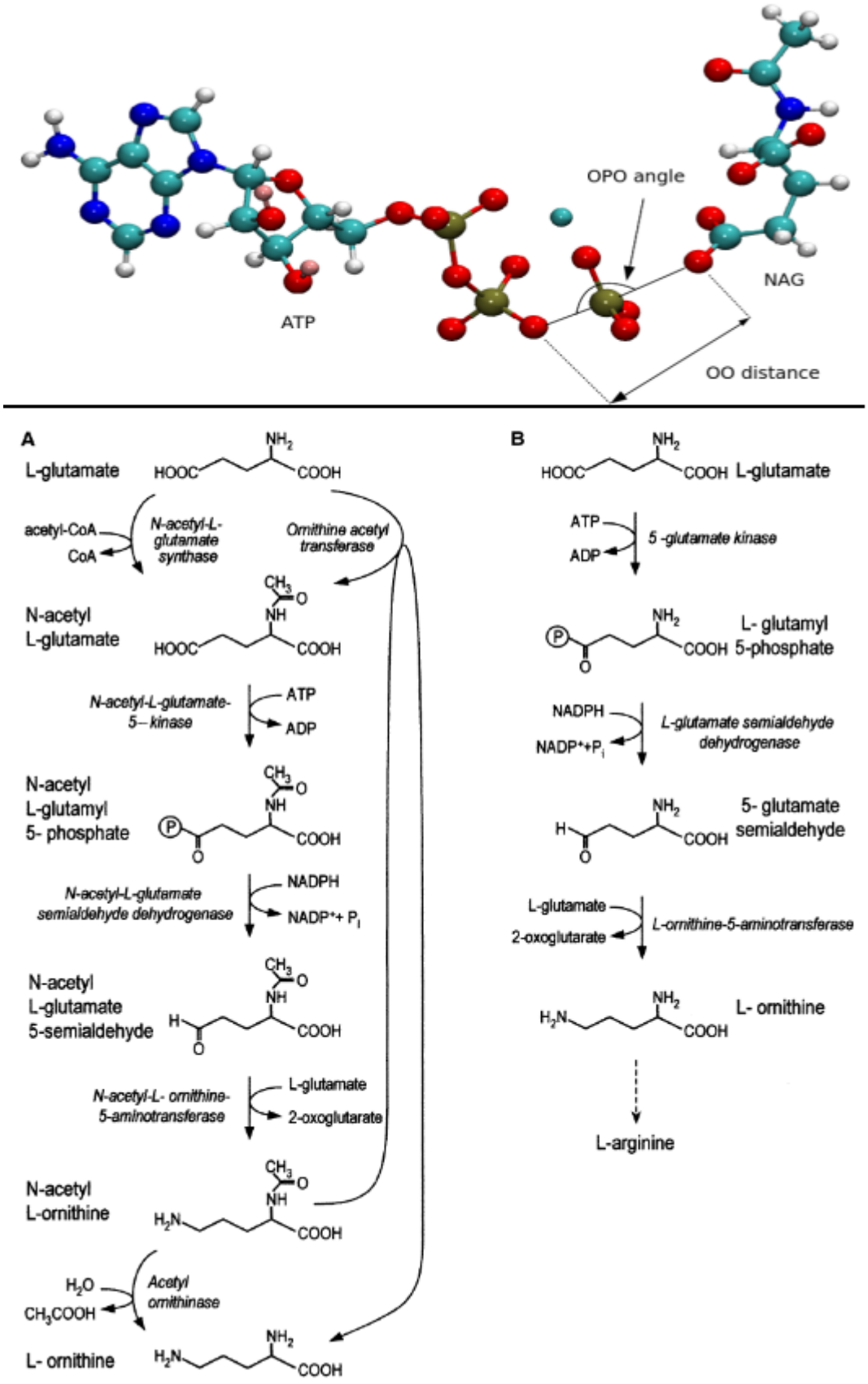


Figure 1.5: Upper panel. Schematic representation of the phosphorylation catalyzed by NAGK. Lower panel. A) Biosynthetic route of arginine in bacteria. B) Biosynthetic route of arginine in mammals. Dashed arrows denote more than one chemical step. Adapted from Ramon-Maiques *et al.*[96].

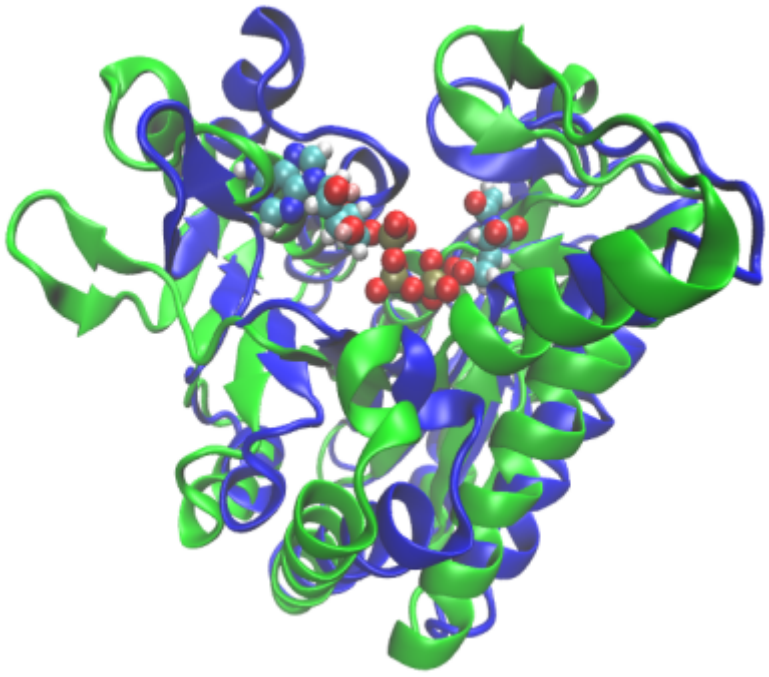


Figure 1.6: Open (green) and closed (blue) conformations of NAGK monomers. The substrates in van der Waals spheres correspond to the closed conformation.

1.2.3 Application of enhanced sampling methods to enzyme catalysis

To understand enzyme catalysis, and mechanism, it is necessary to elucidate the way in which each enzyme exerts electrostatic and other forces to binding substrate and stabilizing transition state, as we have already explained. The enzymes employ all possible strategies to achieve the ultimate objective of reducing the activation free energy. Each enzyme follow its own reaction path, that connects the several chemical species that evolve during the reaction process, within the transition between the reactants and products states. These processes are called *rare events* because they represent rare but important transition events between long lived states. These transitions can be represented over a Potential Energy Surface (PES) or a Free Energy Surfaces (FES), leading to Minimum Energy Paths (MEPs) and Minimum Free Energy Paths (MFEPs) respectively.

The potential energy surface (PES) is a theoretical concept (in chemical physics and related areas) used to describe the energy of a given system, respect to the positions of all the atoms, in other words, a relation between the energy and the geometry of the system. The PES can be characterized by their minima, which correspond to locally stable configurations, and by transition regions connecting

the minima. The Free energy surface (FES) is an averaged projection over the reaction coordinate of the representative PES minima. In other words, each minimum of the PES corresponds to a local free energy minimum, but projection onto a lower dimensional space can produce smoother surfaces (with a much simpler appearance), an important issue regarding proteins (see sections 3.4.1 to 3.4.3 for further details). For relatively rigid systems (such as many organic molecules), and for flexible systems with a small number of significant degrees of freedom, it is possible to determine the PES and FES by sampling the minimum and saddles points. However for proteins (polypeptide chains with many degrees of freedom) determine these surfaces is more complicated, almost impossible. Thus is essential to make several simplifications in the description of proteic PES and FES.

Indeed, the development of approaches for simulating rare events in complex molecular systems, such as enzymes, is a central concern in chemical physics. One of the most common problems is the finding of minimum free energy paths (MFEPs), which describes the chemical mechanism being the resulting energy barrier a good framework to estimate the rate (k) of the process[103–105] and thus to connect theory with experiments. It can be calculated using the Eyring equation under the Transition State Theory (TST)[106]:

$$k = \frac{k_B T}{h} e^{-\frac{\Delta G^{\ddagger}}{RT}} \tag{1.1}$$

where $\Delta G^{\ddagger}$ is the Gibbs free activation energy, k_B is Boltzmann's constant, and h is Planck's constant.

There are two philosophies to approach the problem. The first one is based on determining the free energy surfaces against a predefined set of collective variables (CV) as Metadynamics[37,107], adaptative biased force (ABF)[108] sampling or Umbrella sampling[38]. These techniques require a precise choice of a few CVs. If their number increases these methods rapidly become impractical due to the computational expense and difficulty of exploring multidimensional energy surfaces. Unfortunately, enzymatic reactions are complex, usually defined over roughness energy surfaces and often need many CVs to be entirely described.

The second approach is based on determining the reaction pathways without making any a priori assumption over the CVs, however the initial and final structures are needed. Chain-of-states methods as Zero temperature string method[109,110] and Nudged Elastic Band (NEB)[111,112] does not suffer from the limitation of CV imposed, but in their basic versions, these methods produce only minimum (potential) energy paths (MEPs) as they omit sampling and entropic contributions[113].

To overcome these limitations there have been developed hybrid methods that incorporate the best of the two approaches. Maragliano *et al.*[113] developed the

String method with collective variables to produce MFEPs. In a related development Roux and coworkers proposed a novel method that employed swarms of trajectories (SoT) to evolve the string and to estimate its average displacement in CV space[114]. Recently a comparision between both methods have been made[115] showing that SoT presents the suitable conditions to be applied to the study of biomolecular reactions as enzyme catalysis.

Alternative approaches to calculate free energies based on CVs have also been proposed[116], but their application to large systems such as enzymes are rare. An exception is the work of Zinovjev *et al.*[117] who adapted the method of Branduardi and coworkers[116] to study enzyme catalysis and applied it to the mechanism of isochorismate pyruvate lyase (IPL) (see section 3.4.1 to 3.4.3 for further details about all these concepts and methods). *In this thesis we have implemented the SoT method into the pDynamo library[118] defining the suitable settings to be applied to enzyme catalysis.*

1.3 Protein damage

In the previous section it has been showed that enzymes catalyze a reaction in that some ligands are involved and the global dynamics could help somehow the local dynamics of the active center to carry out the catalytic function (section 1.2). However, sometimes a chemical reaction can occur far from the enzymatic active site due to the action of external factors.

Proteins are continuously damaged by intrinsic and extrinsic factors *in vivo*, influencing several intracellular pathways and resulting in different disorders and diseases[119]. This damage is mostly produced *in vivo* but *in vitro* some experimental procedures damage the biological sample, in some cases, provoking effects that are not shown in nature.

Submitting a protein to external factors, in vitro or even computationally, like high temperature or high electrical pulses to study unfolding processes on proteins[120,121] or provoking a cellular stress to visualize the effect of aging[122] is a common procedure. The problem arises when some experimental procedure destabilize a biological sample unexpectedly in a non desired way, so a deep knowledge of the effects of protein damage is needed.

1.3.1 High radiation damage. Decarboxylation reactions

High radiation damage is one of the most common techniques producing collateral effects *in vitro*, and macromolecular crystallography the scientific field in which

it is mostly observed[123]. Sometimes the radiation damage is used to monitor functional aspects of the structural dynamics of enzymes[124], but commonly they are unexpected or at least non-desirable effects. X-ray radiation damage limits enormously the amount and quality of structural information extracted from protein and virus crystals[125]. Intense X-ray beams from synchrotron sources (see Bildernack *et al.*[126] and Paganin's book[127] for futher discussion) produces punctual chemical and structural damage in proteins during crystallographic data collection affecting the protein conformational motions along of a wide range of time-scales[125]. This damage even occurs with cryo-cooled crystals at 100K, the temperature at which the vast majority of crystallographic data is collected[128–130]. The most common signatures of protein damage in crystals are cleavage of disulfide bonds and decarboxylation of acidic residues, mainly glutamic and aspartic amino acids[125,130].

This specific damage is not made directly through the absorption of an X-ray photon by one of the atoms in the radiation-sensitive group (primary damage). Rather it is a damage inflicted by radicals created after primary photoabsorption elsewhere in the protein or the surrounding solvent (secondary damage)[123], through a charge transfer process (see section 1.3.2). In other words, the X-ray ionize the sample removing the core electrons and thus generating 'holes' in a primary step and in a secondary step these 'holes' are localized on the more stable places, the aminoacids in a biological context. A lot of 'shoot-and-trap' experiments, have been performed to show specific X-ray damage in protein structures as for instance Weik and coworkers[124]. Chemically identical groups in the same protein display differential radiation sensitivities, showing that differences in the chemical and structural environment must be at the origin of the differential sensitivities, although they have remained largely elusive[131]. Another aspect that remain unresolved and could give some insights is the understanding of structural features that might rationalize the broad distribution of, for instance, decarboxylation probabilities of chemically identical groups in a protein.

Decarboxylation reactions are of big importance in biology and a common enzymatic processes. The mechanism for these enzymatic reactions has been widely studied experimentally and computationally[132–136]. There are a high number of enzyme classes (>90), in which decarboxylases are currently organized exhibiting a variety of different catalytic mechanisms with a shared pattern: the cleavage of C-C bonds and the subsequent release of CO_2[137].

Regarding radiation damage processes, this pattern is obviously maintained but with the particularity that could take place in a region far from the active site of the enzyme and between acidic residues. Decarboxylation of acidic residues in proteins[129] might also be explained in terms of an electron migration mechanism, which is initiated by the capture of a secondary hole on the side chain,

resulting in the generation of CO_2 and a carbon-centered radical[138,139]. Several studies have been performed over these reactions trying to justify the migration processes. Some of them tried to relate the radiation-sensitivity with the solvent exposure[128,140], the distance of the damaged residue to the protein surface[141] or the influence of the pKa of a carboxyl group[129,140], but without finding any clear and consensually accepted correlation. Why this process take place remains elusive. *In this thesis we have studied the radiation damage induced decarboxylation in LDH.*

1.3.2 Charge Transfer

Charge Transfer (CT) is a basic chemical process that could be defined as the 'Spontaneous charge redistribution between a reactant state and an acceptor state'[142]. This process can take place in two different regimes, diabatic (non-adiabatic) and adiabatic. To properly understand and explain CT processes, the Marcus theory of electron transfer[143] that is the seed from CT methods have 'grown', is basic.

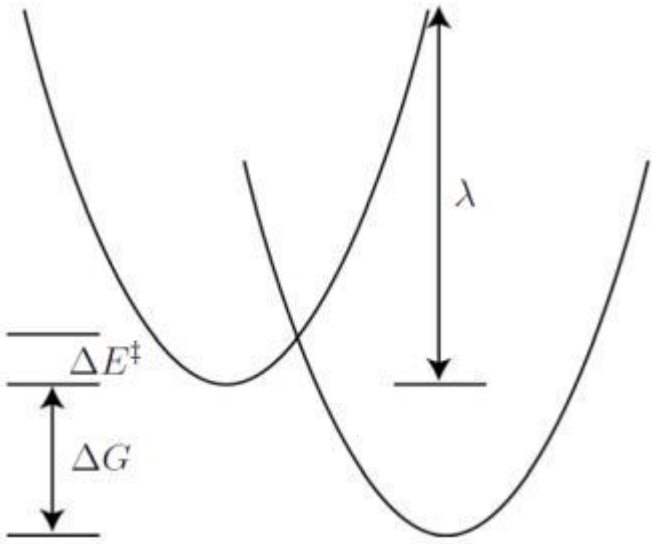


Figure 1.7: ΔG is the free energy change between the reactants on the left and the products on the right and $\Delta E^{\ddagger}$ is the activation energy. λ is the reorganization energy. This is the energy it would take to force the reactants (on the left) to have the same nuclear configuration as the products (on the right) without letting the electron transfer. Reproduced from Marcus 1994[143].

Marcus theory of electron transfer

The Marcus theory describes the electron transfer according to the following equation[143].

$$\Delta E^{\ddagger} = \frac{(\Delta G + \lambda)^2}{4\lambda} \tag{1.2}$$

The electron transfer rate is defined as:

$$k_{ET} = Ae^{\frac{-(\Delta G+\lambda)^2}{4\lambda RT}} \tag{1.3}$$

being 'A' dependent of the electronic coupling $|\mathrm{V}_{DA}|^2$.

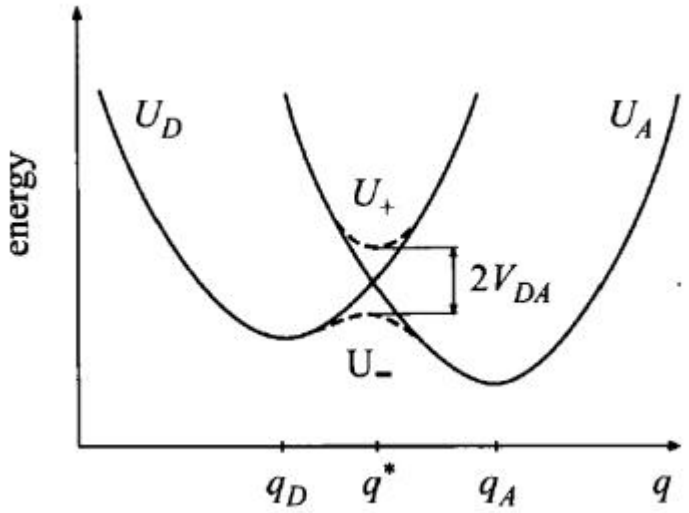


Figure 1.8: Donor and acceptor PES versus a single reaction coordinate. Non-adiabatic (full line), Adiabatic curves (dashed line). There is a splitting between the adiabatic curves which has a magnitude of $2\mathrm{V}_{DA}$ at the crossing point q^*. Reproduced from Kühn & May 2006[142].

The IUPAC gold book says that the adiabatic electron transfer process, is a process in which the reacting system remains on a single electronic surface in passing from reactants to products. In these kind of processes the electronic transmission factor is close to unity.

A non-adiabatic electronic state, is one that does not change its physical character as one moves along the reaction coordinate, whereas the adiabatic Born-Openheimer electronic states are a mixture of non-adiabatic states and changes its physical character at their crossing region. The differences between this two regimes, apart of the different way in that the the Charge transfer is calculated, Fig. 1.8, arise from the fact of the non-adiabatic CT is define as a charge transition process for which the vibrational motion is much faster than the motion of the transferred electron[142], while the adiabatic process proceeds in the opposite way.

The ground adiabatic state is thought of as arising from the avoided crossing between the two states U_+ and U_- as we can see at Fig. 1.8. The molecules that are on U_+ and U_- presents a nuclear coupling, i.e., a spin-orbit interaction. The adiabatic state thus changes its character whereas diabatic, or non-adiabatic, state does not suffer this change. The diabatic states, usually, are calculated from the adiabatic states[144]. This procedure is further explained at Chapter 3, (see section 3.6). Non-adiabatic states play a critical role in ET theory, which is based on the existence of diabatic states on the reactants and products, where the electron is localized on the donor and acceptor, respectively. Reactions are characterized by ΔG, and the reorganization energy, λ, of the diabatic free-energy surfaces[143,145] (see section 1.3.2).

The charge transfer process can be an electron transfer (ET) or a hole transfer (HT), depending on the type of charge which is transferred. There are different possible mechanisms for electron or hole transfer, that can be summarize as: direct exchange or bridge-assisted mechanism, Fig. 1.10, and within the bridge-assisted mechanism, superexchange or hopping mechanisms Fig. 1.11.

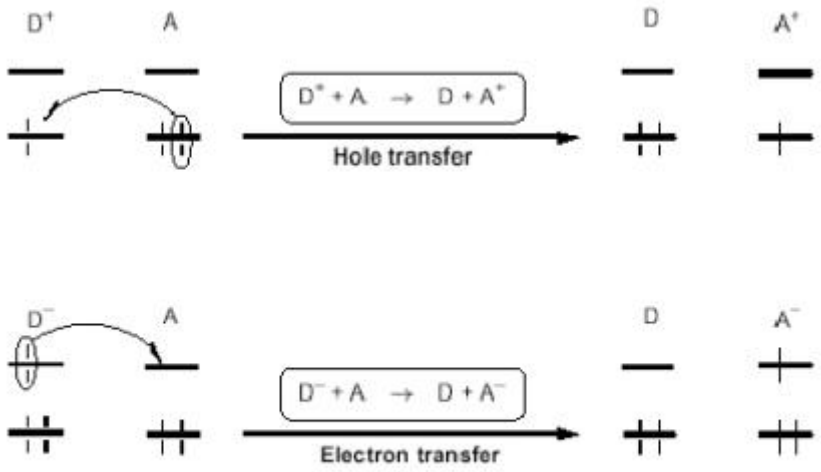


Figure 1.9: Schematic relationship among Hole and Electron transfer[143].

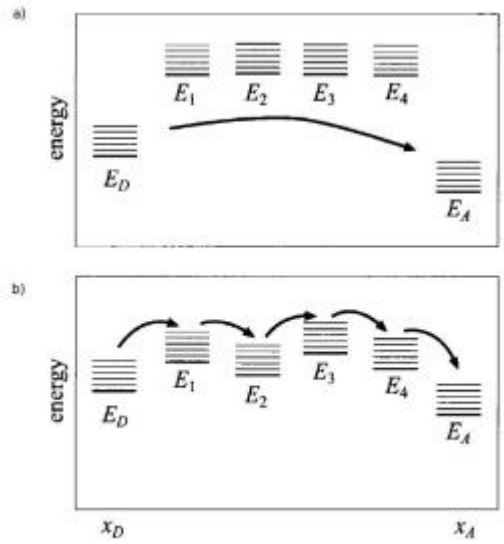


Figure 1.10: a) Direct mechanism b) Bridge-assisted mechanism. E_D represents the donor energy states and E_A the acceptor energy states. $E_{1..4}$ represent the energy states of the binding species. Reproduced from Kühn & May[142].

Direct exchange

The charge is transferred from the donor to the acceptor directly, without interact with other species present in the reaction field. However these species are necessary and (see Fig. 1.10) its presence influence the process creating the correct environment where the charge transfer take place.

Bridge-assisted mechanism

The donor and the acceptor are connected by bridging species. The charge is transferred between the donor and the acceptor but helped by the intermediate species. Depending on the way in that the help comes, is a superexchange or a hopping CT process.

- **Superexchange**

  Is a one step process where the bridge species support the delocalization of the donor state wave function. The orbitals of the intermediate species influence the reaction indirectly creating the suitable environment in which the charge transfer is produced.

- **Hopping**

  Is a multi-step process where the charge is transferred between the species that are involved, including the donor, the acceptor and the bridge. The electronic wave function is subsequently localized on the various sites during the transfer.

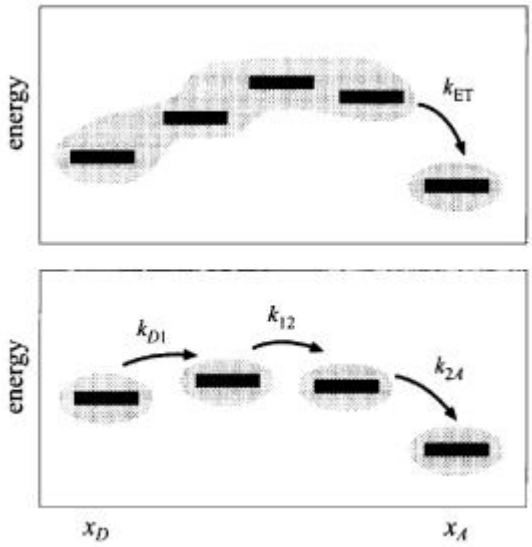


Figure 1.11: Bridge-mediated ET between a donor and an acceptor level. The upper part gives a scheme of the superexchange ET where the initial state wave function extends over the whole bridge. For the sequential ET (lower part) the electronic wave function is localized on the various sites during the transfer [142]. Reproduced from Kühn & May [142].

The charge transfer rate (in non-adiabatic systems) becomes proportional to $|V_{DA}|^2$, but it also depends on the probability at which the crossing region on the donor PES U_D (see Fig. 1.8) is reached. Accordingly, the ET rate is defined as:

$$k_{ET} \propto |V_{DA}|^2 e^{-\frac{\Delta E^{\ddagger}}{RT}} \tag{1.4}$$

$E^{\ddagger}$ denotes the activation energy needed to enter the crossing region starting at the minimum position of the donor PES, hence we have $E^{\ddagger} = U_D(q^*) - U_D(q^D)$.

1.4 Protein dynamics and conformational ensembles

Conformational ensembles, also known as structural ensembles, are the accessible set of a structures at a certain temperature describing the proteins structure. They are powerful tools to represent the range of conformations that can be sampled by proteins, thus allowing for an explicit representation of the dynamics of the protein. They are indicators of the structural heterogeneity of proteins, that can be generated purely theoretically or, as is most often the case, by fitting ensembles of conformations to experimental data [146,147]. Conformational ensembles have been employed to study different aspects related to fundamental properties of

proteins, such as molecular recognition or protein folding. They can not provide the interconversion rate of exchange between conformers or the time-scale of the dynamics, but inform about its amplitude given insights of the behaviour of the protein[148]. There are a wide range of techniques able to generate them.

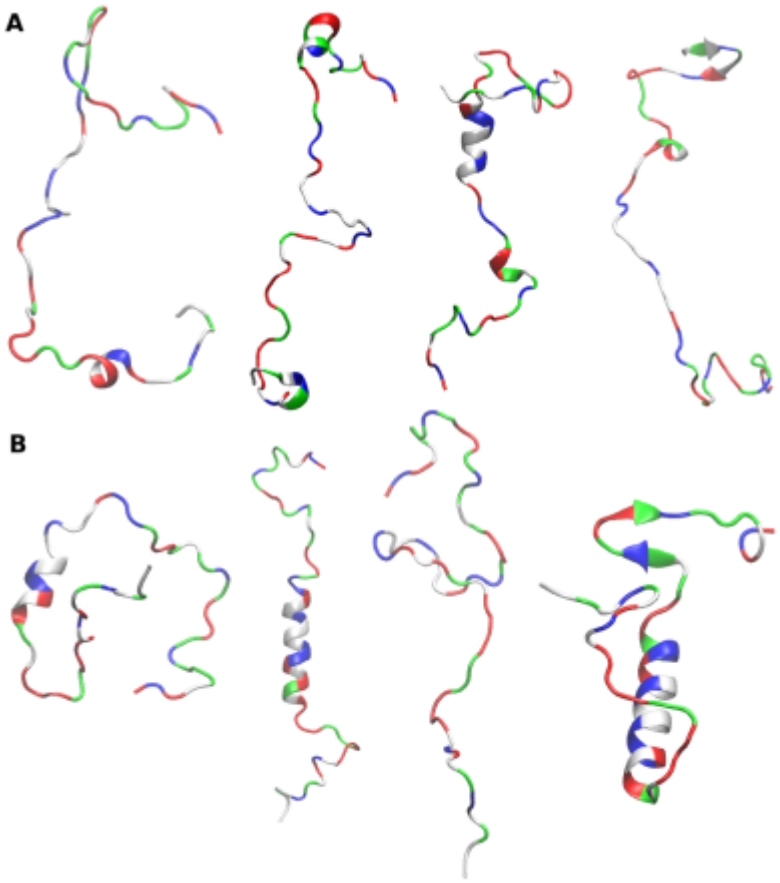


Figure 1.12: Conformational ensembles of the Sendai Virus Nucleocapsid protein generated by using A) PROFASI[149] REMC simulations and B) the statistical coils based ensemble generator FlexibleMeccano[150]. Different colors represent different residue types.

An useful method to generate conformational ensembles is classical MD. MD is a powerful technique to study protein motions generating dynamically relevant conformational ensembles as was shown by Showalter and Brüschweiler[151]. Furthermore with the new hardware advances like Graphics Processing Units (GPUs) or the specifically designed ANTON supercomputer[152] the time range that can cover is being spanned. Additionally, as MD is based on modelling interatomic interactions through empirical potentials (force fields), the continuously improvement and creation of force fields, as Amber-ff99SB-ILDN[153] parameterized against experimental results allows to a better, more realistic, description of

the protein behaviour. However there are still processes that can not be routinely simulated using MD, as when the protein is too big, the conformational space too vast or the dynamics too slow. In these cases enhanced sampling methods as accelerated MD[154] or REMD[155] could be used to overpass this limitation.

When simulations are used to generate ensembles by themselves[146,156], the unique experimental data used is the initial structure. Nevertheless there are scenarios where the conformational space is too vast to be sampled by MD or enhanced sampling methods because, for instance, when the behaviour of the system (a too much, multidimensional and roughness, complex landscape) can not be fully describe by force fields, and thus the experimental data play a bigger role. Some experimental techniques could provide kinetic and structural information (see section 1.1.3.1) as SAXS that provides information about the hydrodynamic properties of the proteins and specially NMR techniques that provide atomic resolution. Within NMR techniques it is possible to measure a wide number of parameters (RDCs, Chemical Shifts (CS), Nuclear Overhaussen Effects (NOE), Scalar Couplings (SC), chemical shifts anisotropy in aligned samples, cross-correlated relaxation rates, Paramagnetic Relaxation Enhancement (PRE) or order parameters (S^2)) that can be in principle related to quantities computed from structures, trajectories and ensembles[147].

These values are used, for instance, to verify simulated ensembles through a back-calculation. This process implies the calculation of quantities that can be compared against experimental data as the aforementioned chemical shifts[157] or RDCs[158]. It is important to take into account the range of validity of the equations employed and its parameterization as well as the accuracy of both the experimental measurements and the backcalculation process[147]. If the back-calculated values disagrees with the experimental ones, this difference could be employed as an empirical potential correction to the potential energy of the protein provided by the force field (protein-specific force-field correction) and thus run restrained simulations[146,159]. There are other variants of the method as ensemble averaged[160,161] or time averaged restrained simulations[162,163].

Furthermore the experimental data could serve as a structural filter. In this case, the data is used to select conformations from a pre-defined pool of conformations generated *a priori*[146,147]. This pool contains all physically possible conformations that the protein can sample with a certain probability in a predifined timescale. If the ensembles does not present the correct statistical weights and as experimental methods contains information about the distribution of the conformations, they could be use to statistically re-weight them[146,147].

IDPs (see section 1.4.1) are inside the processes that generally traditional MD can not cover (although there are studies realized employing MD[164,165]). They

are biologically relevant proteins that can not be represented with conventional structural determination methods such as X-ray crystallography or cryo-electron microscopy. Furthermore current force-fields can not sample their vast conformational space and describe the weak-interactions that dominate its behaviour[166]. IDPs are a perfect example of an scenario in that simulations and experiments have to converge to generate representative conformational ensembles, and a lot of efforts are put on it[167].

1.4.1 Intrinsically Disordered Proteins(IDPs)

The effects of dynamics is specially important for IDPs, which are an emerging family of proteins whose most characteristic feature is that they don't present a folded structure[168–170]. This lack of stable structure can be present over the entire protein length or only in some regions (which are called Intrinsically Disordered Regions IDRs instead of IDPs)[171,172].

Structural disorder is abundant in all species, although its level is higher in eukaryotes than in prokaryotes. By conservative estimates, have been shown that about 40% of eukaryotic proteins contain long disordered regions (of at least 30 residues)[10,168,171,173]. These IDPs or IDRs play key roles in a wide range of cellular processes including signalling, cell cycle control, molecular recognition, transcription, translation and replication. Besides, they are involved in numerous human pathologies such as neurodegenerative diseases, cancer, diabetes and amyloidoses[174–179]. However, its study started at the end of the 20th century, and has been only in the last decade[170] when their existence has been widely accepted, becoming nowadays a hot scientific topic at both experimental and computational level[174–176,180–184].

Due to their high implication in diseases, they are perfect candidates to drug design, but unluckily, this is not happening. There are very few drug targets based on IDPs in contrast to the bioinformatics studies that, for example, showed that Post Transcriptional Modifications (PTMs) prefer disordered regions or that the 79% of cancer associated proteins contain disordered regions of more than 30 residues[184]. This happens because how IDPs perform their diverse functions is not well understood[168,185,186]. Understand their functional and conformational properties, thus, it is of great interest for a wide range of biological processes. In fact, important advances have been made towards its understanding, specially using spectroscopic techniques, i.e., NMR[167,175,187], single-molecule fluroescence[188–191] and with atomistic and coarse-grained simulations[192–196].

Along the last years it has been concluded that the protein dynamics is specially important for IDPs, because owing to their structural plasticity they present

a highly dynamic conformational exchange. The structure and dynamics of IDPs (that present multiple binding sites) are closely related to their interactions with (multiple) binding partners. This phenomena is important for the functional promiscuity and regulation of these proteins[197]. For instance, many IDPs are significantly unstructured under physiological conditions, in the unbound state, and have been shown to undergo coupled folding and binding reactions only upon binding another protein[175,187,198]. These coupled between folding and binding is found commonly in biology[174,176,199–203] although not all IDPs present it[172,204,205]. Recently, the interest is put on the observation that these proteins themselves can be targeted by small molecules[176,206–211].

To study IDPs the usage of conformational ensembles is a widely accepted option. There are several methods developed to select these ensembles, some of them specially designed to study IDPs. There are procedures based on MC as the ENSEMBLE[212] method developed to study IDPs and others on genetic algorithms such as the ASTEROIDS[213] method focus also on IDPs, concretely in generate conformations from NMR data, and the EOM[214] created to generate ensembles from SAXS data. However the most important and limiting factor is the generation of the set of conformations. The pool has to be representative of the size of the conformational landscape, or else the selected ensemble couldn't represent the structural properties of the protein even being in agreement with the experimental data. The pool can be composed by statistical coils or from ensembles determined by simulations[147].

The seminal work of Dobson and coworkers[215] followed by others[150,216,217] showed that is possible to produce representative ensembles of the whole conformational landscape, generating ensembles where the distribution of the backbone torsion angles come from the structures deposited in the PDB. These methods based on statistical coils are aimed to reproduce the structural properties of the polypetide chains where there are dominated by local structural references, for instance when there are not long-range interactions stabilizing the tertiary structure as in globular proteins.

These conformational ensembles are very simple but match the experimental measurements realized over IDPs reasonably well[215–217]. However from Förster Resonance Energy Transfer (FRET) and EPR experiments have been shown that IDPs can form transient long-range interaction important for their physiological roles[174,218,219]. So there are increasing efforts to improve the description of IDPs (concretely the pool generation methods based on statistical coils).

Additionally, there are another methods not based on using statistical coils, but on molecular simulations, aimed to generate sets of conformations. An exemplifying work was the realized by Head-Gordon and coworkers[165], in that they

employed MD simulated ensembles, refined with the ENSEMBLE method, and validated against NMR data. This work was based on the Forman-Key's study [212] that showed the presence of secondary structure in the denaturated state, and thus describing the effect of the long-range interactions in Aβ peptides. It evidenced that the differences between the Aβ40 and the Aβ42 peptides behaviour is due to the differences in the sequences, the two extra residues, enabling the formation of long range contacts with hydrophobic residues along the sequence in the Aβ42 peptide. This is an example, but there are many works that have used this kind of (simulation based) pool generation methods [211,220–222]. *In this thesis we have developed the MaxEnt algorithm that is able to compare two sets of RDCs generated experimentally or by simulations and reweight statistically one set over the other.*

1.4.2 Cooperativity of secondary structure elements in protein ensembles

Some regions of IDPs can adopt secondary structures, at least for a transient time [175], as have been probed experimentally (specially by NMR) [10,223–225]. These structured regions, termed MoRFs (also known as molecular recognition elements, MoREs), are key to recognition processes mediated by coupled folding-binding events. The interpretation of this experimental data is usually done by stating that a certain segment of the protein chain adopts a certain secondary structure in a percentage of the total ensemble. However this way of interpret experimental data, imply a question: How can the ensembles be represented to better unveil their structure?

The MORFs are usually described as the ratio of residues that adopt a certain secondary structure. When we generate ensembles of IDPs it is difficult to visualize their composition or to detect MoRFs. Sometimes the conformational propensities for single residues hide the nature of cooperative structures. Thus it is important to differentiate when residues in a fragment independently adopt a conformation in a secondary structure region (MORFs), from when that fragment contains a true secondary structure, with all the residues adopting that conformation at the same time. In other words if n residues are in a certain secondary structure region the 15% of the time, that does not mean an secondary structure of n residues is present 15% of the time. Whether this happens or not will lead to different experimental results, such as different RDCs, and is related when the aforementioned open question: How can the ensembles be represented to better unveil their structure? *In this thesis we have developed SS-map, that represents the cooperativity or the correlations in secondary structure formation for IDPs*

References

[1] K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature.*, 450(7172):964–972, 2007.

[2] K. Teilum, J. G. Olsen, and B. B. Kragelund. Protein stability, flexibility and function. *Biochimica et Biophys. Acta*, 1814(8):969–976, 2011.

[3] U. Doshi and D. Hamelberg. The dilema of conformational dynamics in enzyme ctalysis: Perpectives from theory and experiment. In *Protein Conformational Dynamics*, pages 221–243. Springer, 2014.

[4] H. Frauenfelder, S. Sligar, and P. Wolynes. The energy landscapes and motions of proteins. *Science.*, 254(5038):1598–1603, 1991.

[5] S. Lukman, C. S. Verma, and G. Fuentes. Exploting protein intrinsic flexiility in drug design. In *Protein Conformational Dynamics*, pages 245–269. Springer, 2014.

[6] D. R. Livesey, editor. *Essentials of Computational Chesmistry:Theories and models*. Methods in Molecular Biology. Humana Press, 2014.

[7] A. Mittermaier and L. E. Kay. New tools provide new insights in NMR studies of protein dynamics. *Science.*, 312(5771):224–228, 2006.

[8] L. E. Kay. NMR studies of protein structure and dynamics. *J. Magn. Reson.*, 173(2):193 – 207, 2005.

[9] Y. Qu, J.-T. Guo, V. Olman, and Y. Xu. Protein structure prediction using sparse dipolar coupling data. *Nucleic Acids Res.*, 32(2):551–561, 2004.

[10] M. Jensen, P. Markwick, S. Meier, C. Griesinger, S. Zweckstetter, P. Bernado, and M. Blackledge. Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Sructure*, 17:1169–1185, 2009.

[11] J. A. Marsh and J. D. Forman-Kay. Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. *J. Mol. Biol.*, 391(2):359 – 374, 2009.

[12] L. Salmon, G. Bouvignies, P. Markwick, and M. Blackledge. Nuclear magnetic resonance provides a quantitative description of protein conformational flexibility on physiologically important time scales. *Biochemistry.*, 50(14):2735–2747, 2011.

[13] D. Bourgeois and A. Royant. Advances in kinetic protein crystallography. *Curr. Opin. Struct. Biol.*, 15(5):538 – 547, 2005.

[14] I. Schlichting, J. Berendzen, K. Chu, A. M. Stock, S. A. Maves, D. E. Benson, R. M. Sweet, D. Ringe, G. A. Petsko, and S. G. Sligar. The catalytic pathway of cytochrome p450cam at atomic resolution. *Science.*, 287(5458):1615–1622, 2000.

[15] M. Weik and J.-P. Colletier. Temperature-dependent macromolecular X-ray crystallography. *Acta Crystallogr. Sect. D*, 66(4):437–446, 2010.

[16] A. Kuzmanic, N. S. Pannu, and B. Zagrovic. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nature. Commun.*, 5, 2014.

[17] K. A. Henzler-Wildman, V. Thai, M. Lei, M. Ott, M. Wolf-Watz, T. Fenn, E. Pozharski, M. A. Wilson, G. A. Petsko, M. Karplus, C. G. Hubner, and D. Kern. Intrinsic motions along an enzymatic reaction trajectory. *Nature.*, 450(7171):838–U13, 2007.

[18] U. H. Hansmann. Sampling protein energy landscapes - the quest for efficient algorithms. In *Multiscale Approaches to Protein Modelling*, pages 209–230. Springer, 2011.

[19] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, and B. Towles. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, SC '09, pages 39:1–39:11, New York, NY, USA, 2009. ACM.

[20] V. Tozzini. Coarse-grained Models for Proteins. *Curr. Opin. Struct. Biol.*, 15(2):144–150, 2005.

[21] A. Liwo, Y. He, and H. A. Scheraga. Coarse-grained force field: general folding theory. *Phys. Chem. Chem. Phys.*, pages 16890–16901, 2011.

[22] S. J. Marrink and D. P. Tieleman. Perspective on the Martini Model. *Chem. Soc. Rev.*, 42:6801–6822, 2013.

[23] F. Sterpone, S. Melchionna, P. Tuffery, S. Pasquali, N. Mousseau, T. Cragnolini, Y. Chebaro, J.-F. St-Pierre, M. Kalimeri, A. Barducci, Y. Laurin, A. Tek, M. Baaden, P. H. Nguyen, and P. Derreumaux. The OPEP protein model: from single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems. *Chem. Soc. Rev.*, 43:4871–4893, 2014.

[24] H. I. Ingolfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole, and S. J. Marrink. The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Science.*, 4(3):225–248, 2014.

[25] M. Neri, C. Anselmi, M. Cascella, A. Maritan, and P. Carloni. Coarse-Grained Model of Proteins Incorporating Atomistic Detail of the Active Site. *Phys. Rev. Lett.*, 95(21):1–4, 2005.

[26] A. J. Rzepiela, M. Louhivuori, C. Peter, and S. J. Marrink. Hybrid simulations: combining atomistic and coarse-grained force fields using virtual sites. *Phys. Chem. Chem. Phys.*, 13(22):10437–48, 2011.

[27] M. R. Machado, P. D. Dans, and S. Pantano. A hybrid all-atom/coarse grain model for multiscale simulations of DNA. *Phys. Chem. Chem. Phys.*, pages 18134–18144, 2011.

[28] B. J. Berne and J. E. Straub. Novel methods of sampling phase space in the simulation of biological systems. *Curr. Opin. Struct. Biol.*, 7(2):181 – 189, 1997.

[29] R. Elber. Long-timescale simulation methods. *Curr. Opin. Struct. Biol.*, 15(2):151–6, 2005.

[30] S. A. Adcock and J. A. McCammon. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem. Rev.*, 21062(5):1589–1615, 2006.

[31] P. Ferrara, J. Apostolakis, and A. Caflisch. Computer simulations of protein folding by targeted molecular dynamics. *Proteins: Struct. Funct. Bioinforma.*, 39(3):252–260, 2000.

[32] P. Krger, S. Verheyden, P. J. Declerck, and Y. Engelborghs. Extending the capabilities of targeted molecular dynamics: Simulation of a large conformational transition in plasminogen activator inhibitor 1. *Protein Science.*, 10(4):798–808, 2001.

[33] B. Isralewitz, J. Baudry, J. Gullingsrud, D. Kosztin, and K. Schulten. Steered molecular dynamics investigations of protein function. *J. Mol. Graph. Model.*, 19(1):13 – 25, 2001.

[34] B. Isralewitz, M. Gao, and K. Schulten. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.*, 11(2):224 – 230, 2001.

[35] D. Hamelberg, J. Mongan, and J. A. McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.*, 120(24):11919–11929, 2004.

[36] R. Kaptein, E. Zuiderweg, R. Scheek, R. Boelens, and W. van Gunsteren. A protein structure from nuclear magnetic resonance data: lac repressor headpiece. *J. Mol. Biol.*, 182(1):179 – 182, 1985.

[37] A. Laio, A. Rodriguez-Fortea, F. L. Gervasio, M. Ceccarelli, and M. Parrinello. Assessing the accuracy of metadynamics. *J. Phys. Chem. B*, 109(14):6714–21, 2005.

[38] J. Kästner. Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Science.*, 1(6):932–942, 2011.

[39] C. Dellago, P. G. Bolhuis, and P. L. Geissler. Transition Path Sampling. *Adv. Chem. Phys.*, 123:1–86, 2002.

[40] A. Mitsutake, Y. Sugita, and Y. Okamoto. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Pept. Science.*, 60(2):96–123, 2001.

[41] D. Frenkel. Speed-up of monte carlo simulations by sampling of rejected states. *Proc. Natl. Acad. Sci.*, 101(51):17571–17575, 2004.

[42] P. Sherwood. Hybrid Quantum Mechanics / Molecular Mechanics approaches. *Mod. Methods Algorithms Quantum Chem.*, 1:257–277, 2000.

[43] H. M. Senn and W. Thiel. QM/MM studies of enzymes. *Curr. opinion chemical biology*, 11(2):182–187, 2007.

[44] H. M. Senn and W. Thiel. QM/MM Methods for Biological Systems. In *Atomistic Approaches in Modern Biology*, volume 268, pages 173–290. Springer, 2007.

[45] H. M. Senn and W. Thiel. QM / MM Methods for Biomolecular Systems. *Angewandte Chemie*, 48(7):1198–1229, 2009.

[46] M. Orozco. A theoretical view of protein dynamics. *Chem. Soc. Rev.*, 43:5051–5066, 2014.

[47] R. Wolfenden and M. J. Snider. The depth of chemical time and the power of enzymes as catalysts. *Accounts Chem. Res.*, 34(12):938–945, 2001.

[48] A. Radzicka and R. Wolfenden. A Proficient Enzyme. *Science.*, 267:90–93, 1995.

[49] C. Lad, N. H. Williams, and R. Wolfenden. The rate of hydrolysis of phosphomonoester dianions and the exceptional catalytic proficiencies of protein and inositol phosphatases. *Proc. Natl. Acad. Sci.*, 100(10):5607–5610, 2003.

[50] A. Warshel. Electrostatic Origin of the Catalytic Power of Enzymes and the Role of Preorganized Active Sites. *J. Biol. Chem.*, 273(42):27035–27038, 1998.

[51] D. Y. Ganapati, D. S. Ashwini, and B. D. Shrikant. Enzyme catalysis in fine chemical and pharmaceutical industries. In *Enzyme Mixtures and Complex Biosynthesis*, pages 79–108. Landes Bioscience, 2007.

[52] M. Rask-Andersen, S. Masuram, and H. B. Schiöth. The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annu. Rev. Pharmacol. Toxicol.*, 54(1):9–26, 2014.

[53] A. Schmid, J. Dordick, B. Hauer, A. Kiener, M. Wubbolts, and W. B. Industrial biocatalysis today and tomorrow. *Nature.*, 409:258–268, 2001.

[54] O. Kirk, T. V. Borchert, and C. C. Fuglsang. Industrial enzyme applications. *Curr. Opin. Biotechnol.*, 13(4):345 – 351, 2002.

[55] A. Schmid, F. Hollmann, J. B. Park, and B. Bühler. The use of enzymes in the chemical industry in europe. *Curr. Opin. Biotechnol.*, 13(4):359 – 366, 2002.

[56] A. T. Martinez, F. J. Ruiz-Dueñas, A. Gutierrez, J. C. del Rio, M. Alcalde, C. Liers, R. Ullrich, M. Hofrichter, K. Scheibner, L. Kalum, J. Vind, and H. Lund. Search, engineering, and applications of new oxidative biocatalysts. *Biofuels, Bioprod. Biorefining*, 2014.

[57] A. Warshel. Energetics of enzyme catalysis. *Proc. Natl. Acad. Sci.*, 75(11):5250–5254, 1978.

[58] S. C. L. Kamerlin and A. Warshel. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins: Struct. Funct. Bioinforma.*, 78(6):1339–1375, 2010.

[59] R. U. Lemieux and U. Spohr. How Emil Fischer was led to the lock and key concept for enzyme specificity. In *Advances in Carbohydrate Chemistry and Biochemistry*, volume 50, pages 1 – 20. Academic Press, 1994.

[60] D. E. Koshland, G. Némethy, and D. Filmer. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits. *Biochemistry.*, 5(1):365–385, 1966.

[61] J. Monod, J. Wyman, and J.-P. Changeux. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.*, 12(1):88–118, 1965.

[62] S. J. Benkovic and S. Hammes-Schiffer. A perspective on enzyme catalysis. *Science.*, 301(5637):1196–1202, 2003.

[63] P. K. Agarwal. Role of Protein Dynamics in Reaction Rate Enhancement by Enzymes. *J. Am. Chem. Soc.*, 127(43):15248–15256, 2005.

[64] S. Hammes-Schiffer and S. J. Benkovic. Relating protein motion to catalysis. *Annu. Rev. Biochemistry.*, 75:519–541, 2006.

[65] L. Masgrau, A. Roujeinikova, L. O. Johannissen, P. Hothi, J. Basran, K. E. Ranaghan, A. J. Mulholland, M. J. Sutcliffe, N. S. Scrutton, and D. Leys. Atomic description of an enzyme reaction dominated by proton tunneling. *Science.*, 312(5771):237–241, 2006.

[66] K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature.*, 450(7171):913–U27, 2007.

[67] S. D. Schwartz and V. L. Schramm. Enzymatic transition states and dynamic motion in barrier crossing. *Nature. Chem. Biol.*, 5(8):551–8, 2009.

[68] N. D. Zachary and J. P. Klinman. A 21^{st} century revisionist's view at a turning point in enzymology. *Nature. Chem. Biol.*, 5:543–550, 2009.

[69] V. C. Nashine, S. Hammes-Schiffer, and S. J. Benkovic. Coupled motions in enzyme catalysis. *Curr. Opin. Chem. Biol.*, 14(5):644 – 651, 2010.

[70] G. G. Hammes, S. J. Benkovic, and S. Hammes-Schiffer. Flexibility, diversity, and cooperativity: pillars of enzyme catalysis. *Biochemistry.*, 50(48):10422–30, December 2011.

[71] J. D. McGeagh, K. E. Ranaghan, and A. J. Mulholland. Protein dynamics and enzyme catalysis Insights from simulations. *Biochimica et Biophys. Acta*, 1814(8):1077–1092, 2011.

[72] M. Karplus. Role of conformation transitions in adenylate kinase. *Proc. Natl. Acad. Sci.*, 107(17):E71; author reply E72, 2010.

[73] A. V. Pisliakov, J. Cao, S. C. L. Kamerlin, and A. Warshel. Enzyme millisecond conformational dynamics do not catalyze the chemical step. *Proc. Natl. Acad. Sci.*, 106(41):17359–17364, 2009.

[74] S. C. L. Kamerlin and A. Warshel. Reply to Karplus: Conformational dynamics have no role in the chemical step. *Proc. Natl. Acad. Sci.*, 107(17):E72–E72, April 2010.

[75] M. Wolf-Watz, V. Thai, K. Henzler-Wildman, G. Hadjipavlou, E. Z. Eisenmesser, and D. Kern. Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nature. Struct. & Mol. Biol.*, 11(10):945–949, 2004.

[76] J. C. Gerhart and H. K. Schachman. Allosteric interactions in aspartate transcarbamylase. II. evidence for different conformational states of the protein in the presence and absence of specific ligands. *Biochemistry.*, 7(2):538–552, 1968.

[77] E. Marcos, R. Crehuet, and I. Bahar. Changes in dynamics upon oligomerization regulate substrate binding and allostery in amino acid kinase family members. *PLoS Comput. Biol*, 7(9):e1002201, 2011.

[78] H. P. Lu, L. Xun, and X. S. Xie. Single-molecule enzymatic dynamics. *Science.*, 282(5395):1877–1882, 1998.

[79] E. Marcos, R. Crehuet, and I. Bahar. On the conservation of the slow conformational dynamics within the amino acid kinase family: Nagk the paradigm. *PLoS Comput. Biol*, 6(4):e1000738, 2010.

[80] P. T. R. Rajagopalan, S. Lutz, and S. J. Benkovic. Coupling interactions of distal residues enhance dihydrofolate reductase catalysis: Mutational effects on hydride transfer rates. *Biochemistry.*, 41:12618–12628, 2002.

[81] J. R. E. T. Pineda, D. Antoniou, and S. D. Schwartz. Slow conformational motions that favor sub-picosecond motions important for catalysis. *J. Phys. Chem. B*, 114(48):15985–15990, 2010.

[82] S. Hay and N. S. Scrutton. Good vibrations in enzyme-catalysed reactions. *Nature. Chem.*, 4(3):161–8, January 2012.

[83] R. García-Meseguer, S. Martí, J. J. Ruiz-Pernía, V. Moliner, and I. Tuñón. Studying the role of protein dynamics in an SN2 enzyme reaction using free-energy surfaces and solvent coordinates. *Nature. Chem.*, 5(7):566–571, 2013.

[84] E. Z. Eisenmesser, D. a. Bosco, M. Akke, and D. Kern. Enzyme dynamics during catalysis. *Science.*, 295(5559):1520–1523, 2002.

[85] 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. Intrinsic dynamics of an enzyme underlies catalysis. *Nature.*, 438(7064):117–121, 2005.

[86] J. A. Hanson, K. Duderstadt, L. P. Watkins, S. Bhattacharyya, J. Brokaw, J.-W. Chu, and H. Yang. Illuminating the mechanistic roles of enzyme conformational dynamics. *Proc. Natl. Acad. Sci.*, 104(46):18055–18060, 2007.

[87] U. Doshi, L. C. McGowan, S. T. Ladani, and D. Hamelberg. Resolving the complex role of enzyme conformational dynamics in catalytic function. *Proc. Natl. Acad. Sci.*, 2012.

[88] A. Warshel, P. K. Sharma, M. Kato, Y. Xiang, H. Liu, and M. H. M. Olsson. Electrostatic Basis for Enzyme Catalysis. *Chem. Rev.*, 106(8):3210–3235, 2006.

[89] G. Bhabha, J. Lee, D. C. Ekiert, J. Gam, I. A. Wilson, H. J. Dyson, S. J. Benkovic, and P. E. Wright. A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science.*, 332(6026):234–8, 2011.

[90] P. K. Agarwal, S. R. Billeter, and S. Hammes-Schiffer. Nuclear quantum effects and enzyme dynamics in dihydrofolate reductase catalysis. *J. Phys. Chem. B*, 106:3283–3293, 2002.

[91] S. Saen-Oon, M. Ghanem, V. L. Schramm, and S. D. Schwartz. Remote Mutations and Active Site Dynamics Correlate with Catalytic Properties of Purine Nucleoside Phosphorylase. *Biophys. J.*, 94(10):4078–4088, 2008.

[92] D. Antoniou, J. Basner, S. Núñez, and S. D. Schwartz. Computational and Theoretical Methods to Explore the Relation between Enzyme Dynamics and Catalysis. *Chem. Rev.*, 106(8):3170–3187, 2006.

[93] D. Tobi and I. Bahar. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci.*, 102(52):18908–18913, 2005.

[94] A. C. Hengge. Mechanistic studies on enzyme-catalyzed phosphoryl transfer. In *Advances in Physical Organic Chemistry*, volume 40. Academic Press, 2005.

[95] W. W. Cleland and A. C. Hengge. Enzymatic mechanisms of phosphate and sulfate transfer. *Chem. Rev.*, 106(8):3252–3278, 2006.

[96] S. Ramón-Maiques, A. Marina, F. Gil-Ortiz, I. Fita, and V. Rubio. Structure of acetylglutamate kinase, a key enzyme for arginine biosynthesis and a prototype for the amino acid kinase enzyme family, during catalysis. *Structure.*, 10(3):329–42, 2002.

[97] F. Gil-Ortiz, S. Ramón-Maiques, I. Fita, and V. Rubio. 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 ALF_4-Transition State Mimic. *J. Mol. Biol.*, 331(1):231–244, 2003.

[98] S. Ramón-Maiques, M. L. Fernández-Murga, F. Gil-Ortiz, A. Vagin, I. Fita, and V. Rubio. 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(3):695–713, 2006.

[99] M. L. Fernandez-Murga and V. Rubio. 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(8):3018–3025, 2008.

[100] F. Gil-Ortiz, S. Ramón-Maiques, M. L. Fernandez-Murga, I. Fita, and V. Rubio. Two Crystal Structures of Escherichia coli N-Acetyl-l-Glutamate Kinase Demonstrate the Cycling between Open and Closed Conformations. *J. Mol. Biol.*, 399(3):476–490, 2010.

[101] C. Marco-Marín, S. Ramón-Maiques, S. Tavárez, and V. Rubio. 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(3):459–476, 2003.

[102] A. Gora, J. Brezovsky, and J. Damborsky. Gates of Enzymes. *Chem. Rev.*, 2013.

[103] M. J. Field. Simulating enzyme reactions: challenges and perspectives. *J. Comput. Chem.*, 23(1):48–58, January 2002.

[104] M. J. Field. *A Practical Introduction to the Simulation of Molecular Systems.* Cambridge University Press, 2007.

[105] E. Weinan and E. Vanden-Eijnden. Transition-path theory and path-finding algorithms for the study of rare events. *Annu. review physical chemistry*, 61:391–420, 2010.

[106] H. Eyring. The activated complex and the absolute rate of chemical reactions. *Chem. Rev.*, 17(1):65–77, 1935.

[107] A. Laio and M. Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci.*, 99(20):12562–6, October 2002.

[108] E. Darve and A. Pohorille. Calculating free energies using average force. *J. Chem. Phys.*, 115(20):9169, 2001.

[109] W. E, W. Ren, and E. Vanden-Eijnden. String method for the study of rare events. *Phys. Rev. B*, 66(5):052301, 2002.

[110] W. E, W. Ren, and E. Vanden-Eijnden. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.*, 126(16):164103, 2007.

[111] H. Jonsson, G. Mills, and K. W. Jacobsen. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*. World Scientific, 385–404.

[112] G. Henkelman and H. Jonsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.*, 113(22):9978–9985, 2000.

[113] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.*, 125(2):24106, 2006.

[114] A. C. Pan, D. Sezer, and B. Roux. Finding transition pathways using the string method with swarms of trajectories. *J. Phys. Chem. B*, 112(11):3432–40, 2008.

[115] L. Maragliano, B. Roux, and E. Vanden-Eijnden. A comparison between mean forces and swarms-of-trajectories string methods. *J. Chem. Theory Comput.*, 10(2):524–533, 2014.

[116] D. Branduardi, F. L. Gervasio, and M. Parrinello. From A to B in free energy space. *J. Chem. Phys.*, 126(5):054103, 2007.

[117] K. Zinovjev, S. Marti, and I. Tuñon. A collective coordinate to obtain free energy profiles for complex reactions in condensed phases. *J. Chem. Theory Comput.*, 8(5):1795–1801, 2012.

[118] M. J. Field. The pDynamo program for molecular simulations using hybrid quantum chemical and molecular mechanical potentials. *J. Chem. Theory Comput.*, 4(7):1151–1161, 2008.

[119] N. Chondrogianni, I. Petropoulos, S. Grimm, K. Georgila, B. Catalgol, B. Friguet, T. Grune, and E. S. Gonos. Protein damage, repair and proteolysis. *Mol. Aspects Medicine*, 35(0):1 – 71, 2014.

[120] E. Paci and M. Karplus. Unfolding proteins by external forces and temperature: The importance of topology and energetics. *Proc. Natl. Acad. Sci.*, 97(12):6521–6526, 2000.

[121] K. J. Freedman, M. Jürgens, A. Prabhu, C. W. Ahn, P. Jemth, J. B. Edel, and M. J. Kim. Chemical, thermal, and electric field induced unfolding of single protein molecules studied using nanopores. *Anal. Chem.*, 83(13):5137–5144, 2011.

[122] C. M. Dong, X. L. Wang, G. M. Wang, W. J. Zhang, L. Zhu, S. Gao, D. J. Yang, Y. Qin, Q. J. Liang, Y.-L. Chen, H. T. Deng, K. Ning, A. B. Liang, Z. L. Gao, and J. Xu. A stress-induced cellular aging model with postnatal neural stem cells. *Cell death & disease*, 5:e1116, 2014.

[123] E. F. Garman. Radiation damage in macromolecular crystallography : what is it and why should we care? *Acta Crystallogr. Sect. D*, 66(4):339–351, 2010.

[124] J.-P. Colletier, D. Bourgeois, B. Sanson, D. Fournier, J. L. Sussman, I. Silman, and M. Weik. Shoot-and-Trap : Use of specific x-ray damage temperature-controlled cryocrystallography. *Proc. Natl. Acad. Sci.*, 105(33):11742–11747, 2008.

[125] M. Warkentin, J. B. Hopkins, R. Badeau, A. M. Mulichak, L. J. Keefe, and R. E. Thorne. Global radiation damage: temperature dependence, time dependence and how to outrun it. *J. Synchrotron Radiat.*, 20(1):7–13, 2013.

[126] D. H. Bilderback, P. Elleaume, and E. Weckert. Review of third and next generation synchrotron light sources. *J. Phys. B: At. Mol. Opt. Phys.*, 38(9):S773, 2005.

[127] D. Paganin. *Coherent X-Ray Optics*. Oxford Science Publications, 2006.

[128] W. P. Burmeister. Structural changes in a cryo-cooled protein crystal owing to radiation damage. *Acta Crystallogr. Sect. D*, 56(3):328–341, Mar 2000.

[129] R. B. G. Ravelli and S. M. McSweeney. The 'fingerprint' that X-rays can leave on structures. *Structure.*, 8(3):315 – 328, 2000.

[130] M. Weik, R. B. G. Ravelli, G. Kryger, S. McSweeney, M. L. Raves, M. Harel, P. Gros, I. Silman, J. Kroon, and J. L. Sussman. Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proc. Natl. Acad. Sci.*, 97(2):623–628, 2000.

[131] K. A. Sutton, P. J. Black, K. R. Mercer, E. F. Garman, R. L. Owen, E. H. Snell, and W. A. Bernhard. Insights into the mechanism of X-ray-induced disulfide-bond cleavage in lysozyme crystals based on EPR, optical absorption and X-ray diffraction studies. *Acta Crystallogr. Sect. D*, 69(12):2381–2394, 2013.

[132] J. P. Richard and T. L. Amyes. On the importance of being zwitterionic: enzymatic catalysis of decarboxylation and deprotonation of cationic carbon. *Bioorganic Chem.*, 32(5):354 – 366, 2004.

[133] M. Garcia-Viloca, J. Gao, M. Karplus, and D. G. Truhlar. How enzymes work: analysis by modern rate theory and computer simulations. *Science.*, 303(5655):186–195, 2004.

[134] R. Z. Liao, J. G. Yu, and F. Himo. Quantum chemical modeling of enzymatic reactions: The case of decarboxylation. *J. Chem. Theory Comput.*, 7(5):1494–1501, 2011.

[135] J. P. Richard. Enzymatic catalysis of proton transfer and decarboxylation reactions. *Pure Appl. Chem.*, 83(8):1499–1641, 2011.

[136] R. Kourist, J.-K. Guterl, K. Miyamoto, and V. Sieber. Enzymatic decarboxylation−an emerging reaction for chemicals production from renewable resources. *Chem. Cat. Chem.*, 6(3):689–701, 2014.

[137] T. Li, L. Huo, C. Pulley, and A. Liu. Decarboxylation mechanisms in biological system. *Bioorganic Chem.*, 43(0):2–14, 2012.

[138] M. D. Sevilla, J. B. D'Arcy, and K. M. Morehouse. An electron spin resonance study of γ-irradiated frozen aqueous solutions containing N-acetylamino acids. *J. Phys. Chem.*, 83(22):2893–2897, 1979.

[139] H. C. Box and E. E. Budzinski. Oxidation and reduction of amino acids by ionizing radiation. *J. Chem. Phys.*, 55(5):2446–2449, 1971.

[140] E. Fioravanti, F. M. D. Vellieux, P. Amara, D. Madern, and M. Weik. Specific radiation damage to acidic residues and its relation to their chemical and structural environment. *J. Synchrotron Radiat.*, 14(1):84–91, 2007.

[141] D. H. Juers and M. Weik. Similarities and differences in radiation damage at 100K *versus* 160K in a crystal of thermolysin. *J. Synchrotron Radiat.*, 18(3):329–337, 2011.

[142] V. May and O. Kühn. Chapter 6. In *Charge and Energy Transfer Dynamics in Molecular Systems*, pages 285–376. Wiley-VCH, 2004.

[143] R. A. Marcus. Chemical and electrochemical electron-transfer theory. *Annu. Rev. Phys. Chem.*, 15(1):155–196, 1964.

[144] A. A. Voityuk and N. Rösch. Fragment charge difference method for estimating donor−acceptor electronic coupling: Application to DNA π-stacks. *J. Chem. Phys.*, 117(12):5607, 2002.

[145] T. Van Voorhis, T. Kowalczyk, B. Kaduk, L.-P. Wang, C.-L. Cheng, and Q. Wu. The diabatic picture of electron transfer, reaction barriers, and molecular dynamics. *Annu. Rev. Phys. Chem.*, 61(1):149–170, 2010.

[146] M. Vendruscolo. Determination of conformationally heterogeneous states of proteins. *Curr. Opin. Struct. Biol.*, 17(1):15 – 20, 2007.

[147] X. Salvatella. Understanding protein dynamics using conformational ensembles. In *Protein Conformational Dynamics*, pages 67–85. Springer, 2014.

[148] R. B. Fenwick, S. Esteban-Martín, and X. Salvatella. Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Eur. biophysics journal*, 40(12):1339–55, 2011.

[149] A. Irbäck and S. Mohanty. PROFASI : A Monte Carlo Simulation Package for Protein Folding and Aggregation. *J. Comput. Chem.*, 27:1548–1555, 2006.

[150] V. Ozenne, R. Schneider, M. Yao, J.-R. Huang, L. Salmon, M. Zweckstetter, M. R. Jensen, and M. Blackledge. Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. *J. Am. Chem. Soc.*, 134(36):15138–48, 2012.

[151] S. A. Showalter and R. Brüschweiler. Quantitative molecular ensemble interpretation of NMR dipolar couplings without restraints. *J. Am. Chem. Soc.*, 129(14):4158–4159, 2007.

[152] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. In *Proceedings of the 34th Annual International Symposium on Computer Architecture*, ISCA '07, pages 1–12. ACM, 2007.

[153] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct. Funct. Bioinforma.*, 78(8):1950–1958, 2010.

[154] M. J. Harvey, G. Giupponi, and G. D. Fabritiis. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.*, 5(6):1632–1639, 2009.

[155] W. Xu, C. Zhang, L. Morozova-Roche, J. Z. H. Zhang, and Y. Mu. pH-dependent conformational ensemble and polymorphism of amyloid-β core fragment. *J. Phys. Chem. B*, 117(28):8392–9, July 2013.

[156] A. R. Fersht and V. Daggett. Protein folding and unfolding at atomic resolution. *Cell*, 108(4):573 – 582, 2002.

[157] S. Neal, A. M. Nip, H. Zhang, and D. S. Wishart. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR.*, 26(3):215–40, 2003.

[158] M. Zweckstetter. NMR: prediction of molecular alignment from structure using the PALES software. *Nature. protocols*, 3(4):679–90, January 2008.

[159] S. Esteban-Martin, R. Bryn Fenwick, and X. Salvatella. Synergistic use of NMR and MD simulations to study the structural heterogeneity of proteins. *Wiley Interdiscip. Rev. Comput. Mol. Science.*, 2(3):466–478, 2012.

[160] A. M. Bonvin and A. T. Brünger. Do NOE distances contain enough information to assess the relative populations of multi-conformer structures? *J. Biomol. NMR.*, 7(1):72–76, 1996.

[161] K. Lindorff-Larsen, R. B. Best, M. A. Depristo, C. M. Dobson, and M. Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature.*, 433(7022):128–32, 2005.

[162] A. E. Torda, R. M. Scheek, and W. F. van Gunsteren. Time-dependent distance restraints in molecular dynamics simulations. *Chem. Phys. Lett.*, 157(4):289 – 294, 1989.

[163] A. E. Torda, R. M. Scheek, and W. F. van Gunsteren. Time-averaged nuclear overhauser effect distance restraints applied to tendamistat. *J. Mol. Biol.*, 214(1):223 – 235, 1990.

[164] S.-H. Chong and S. Ham. Conformational entropy of intrinsically disordered protein. *J. Phys. Chem. B*, 117(18):5503–9, 2013.

[165] K. Ball, A. Phillips, D. Wemmer, and T. Head-Gordon. Differences in β-strand populations of monomeric aβ40 and aβ42. *Biophys. J.*, 104(12):2714 – 2724, 2013.

[166] K. Lindorff-Larsen, N. Trbovic, P. Maragakis, S. Piana, and D. E. Shaw. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J. Am. Chem. Soc.*, 134(8):3787–3791, 2012.

[167] M. R. Jensen, R. W. Ruigrok, and M. Blackledge. Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr. Opin. Struct. Biol.*, 23(3):426 – 435, 2013.

[168] V. N. Uversky. Natively unfolded proteins: a point where biology waits for physics. *Protein Science.*, 11(4):739–756, 2002.

[169] V. N. Uversky. Unusual biophysics of intrinsically disordered proteins. *Biochimica et Biophys. Acta*, 1834(5):932–51, 2013.

[170] P. Tompa. Intrinsically disordered proteins: a 10-year recap. *Trends biochemical sciences*, 37(12):509–16, 2012.

[171] A. L. Fink. Natively unfolded proteins. *Curr. opinion structural biology*, 15(1):35–41, 2005.

[172] P. Tompa and M. Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends biochemical sciences*, 33(1):2–8, 2008.

[173] P. Tompa. Intrinsically unstructured proteins. *Trends Biochem. Sci.*, 27:527–533, 2002.

[174] V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Intrinsically disordered proteins in human diseases: Introducing the d2 concept. *Annu. Rev. Biophys.*, 37(1):215–246, 2008.

[175] H. J. Dyson. Expanding the proteome: disordered and alternatively folded proteins. *Q. reviews biophysics*, 44(4):467–518, 2011.

[176] P. Tompa. Unstructural biology coming of age. *Curr. Opin. Struct. Biol.*, 21(3):419–25, 2011.

[177] O. Coskuner and O. Wise-Scira. Arginine and disordered amyloid-β peptide structures: Molecular level insights into the toxicity in alzheimer's disease. *ACS Chem. Neurosci.*, 4(12):1549–1558, 2013.

[178] H. Wang, D. I. Hammoudeh, A. V. Follis, B. E. Reese, J. S. Lazo, S. J. Metallo, and E. V. Prochownik. Improved low molecular weight Myc-Max inhibitors. *Mol. Cancer Ther.*, 6(9):2399–408, 2007.

[179] J. Gsponer, M. E. Futschik, S. A. Teichmann, and M. M. Babu. Tight regulation of unstructured proteins: From transcript synthesis to protein degradation. *Science.*, 322(5906):1365–1368, 2008.

[180] V. N. Uversky, A. S. Karnoup, R. Khurana, D. J. Segel, S. Doniach, and A. L. Fink. Association of partially-folded intermediates of staphylococcal nuclease induces structure and stability. *Protein science*, 8(1):161–73, 1999.

[181] V. N. Uversky, J. Li, and a. L. Fink. Evidence for a partially folded intermediate in alpha-synuclein fibril formation. *J. Biol. Chem.*, 276(14):10737–44, 2001.

[182] P. Tompa. Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays*, 25(9):847–55, 2003.

[183] P. Tompa, C. Szász, and L. Buday. Structural disorder throws new light on moonlighting. *Trends biochemical sciences*, 30(9):484–9, 2005.

[184] C. J. Oldfield, B. Xue, A. K. Dunker, and V. N. Uversky. Binding promiscuity of unfolded peptides. In *Protein and Peptide Folding, Misfolding, and Non-Folding*, pages 239–277. John Wiley & Sons, Inc., 2012.

[185] Y. Wang, J. C. Fisher, R. Mathew, L. Ou, S. Otieno, J. Sublet, L. Xiao, J. Chen, M. F. Roussel, and R. W. Kriwacki. Intrinsic disorder mediates the diverse regulatory functions of the Cdk inhibitor p21. *Nature. Chem. Biol.*, 7(4):214–21, 2011.

[186] V. N. Uversky. A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein science*, 22(6):693–724, 2013.

[187] P. E. Wright and H. J. Dyson. Linking folding and binding. *Curr. Opin. Struct. Biol.*, 19(1):31–8, 2009.

[188] A. C. M. Ferreon, J. C. Ferreon, P. E. Wright, and A. A. Deniz. Modulation of allostery by protein intrinsic disorder. *Nature.*, 498(7454):390–4, 2013.

[189] A. Soranno, I. Koenig, M. B. Borgia, H. Hofmann, F. Zosel, D. Nettels, and B. Schuler. Single-molecule spectroscopy reveals polymer effects of disordered proteins in crowded environments. *Proc. Natl. Acad. Sci.*, 2014.

[190] B. Schuler, A. Soranno, and D. Nettels. Application of confocal single molecule FRET to intrinsically disordered proteins. In *Intrinsically Disordered Protein Analysis*, volume 896, chapter 2, pages 21–45. Springer New York, 2012.

[191] A. C. M. Ferreon, C. R. Moran, Y. Gambin, and A. A. Deniz. Single-molecule fluorescence studies of intrinsically disordered proteins. In *Methods in enzymology*, volume 472, chapter 10, pages 179–204. Elsevier Inc., 1 edition, 2010.

[192] A. H. Mao, N. Lyle, and R. V. Pappu. Describing sequence-ensemble relationships for intrinsically disordered proteins. *Biochem. journal*, 449(2):307–18, 2013.

[193] C. K. Fisher and C. M. Stultz. Constructing ensembles for intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, 21(3):426–31, 2011.

[194] N. Lyle, R. K. Das, and R. V. Pappu. A quantitative measure for protein conformational heterogeneity. *J. Chem. Phys.*, 139(12):121907, 2013.

[195] R. K. Das, A. Mittal, and R. V. Pappu. How is functional specificity achieved through disordered regions of proteins? *BioEssays*, 35(1):17–22, 2013.

[196] S. L. Crick and R. V. Pappu. Thermodynamic and kinetic models for aggregation of intrinsically disordered proteins. In *Protein and Peptide Folding, Misfolding, and Non-Folding*, pages 413–440. John Wiley & Sons, Inc., 2012.

[197] J. H. Fong and A. R. Panchenko. Intrinsic disorder and protein multibinding in domain, terminal, and linker regions. *Mol. BioSyst.*, 6:1821–1828, 2010.

[198] J. D. Forman-Kay and T. Mittag. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure.*, 21(9):1492–9, 2013.

[199] S. Gianni, A. Morrone, R. Giri, and M. Brunori. A folding-after-binding mechanism describes the recognition between the transactivation domain of c-Myb and the KIX domain of the CREB-binding protein. *Biochem. Biophys. Res. Commun.*, 428(2):205–9, 2012.

[200] Y. Huang and Z. Liu. Kinetic advantage of intrinsically disordered proteins in coupled folding-binding process: a critical assessment of the 'fly-casting' mechanism. *J. Mol. Biol.*, 393(5):1143–59, 2009.

[201] S. L. Shammas, J. M. Rogers, S. a. Hill, and J. Clarke. Slow, reversible, coupled folding and binding of the spectrin tetramerization domain. *Biophys. J.*, 103(10):2203–14, 2012.

[202] J. Dogan, T. Schmidt, X. Mu, A. Engstrm, and P. Jemth. Fast association and slow transitions in the interaction between two intrinsically disordered protein domains. *J. Biol. Chem.*, 287(41):34316–24, 2012.

[203] E. A. Cino, R. C. Killoran, M. Karttunen, and W. Y. Choy. Binding of disordered proteins to a protein hub. *Sci. reports*, 3:2305, 2013.

[204] M. Fuxreiter. Fuzziness: linking regulation to protein dynamics. *Mol. BioSyst.*, 8:168–177, 2012.

[205] T. Mittag, L. E. Kay, and J. D. Forman-Kay. Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.*, 23(2):105–116, 2010.

[206] A. V. Follis, D. I. Hammoudeh, H. Wang, E. V. Prochownik, and S. J. Metallo. Structural rationale for the coupled binding and unfolding of the c-Myc oncoprotein by small molecules. *Chem. & Biol.*, 15(11):1149–55, 2008.

[207] M. Bisaglia, L. Tosatto, F. Munari, I. Tessari, P. P. de Laureto, S. Mammi, and L. Bubacco. Dopamine quinones interact with alpha-synuclein to form unstructured adducts. *Biochem. Biophys. Res. Commun.*, 394(2):424–8, 2010.

[208] F. E. Herrera, A. Chesi, K. E. Paleologou, A. Schmid, M. Vendruscolo, S. Gustincich, H. A. Lashuel, and P. Carloni. Inhibition of $\alpha-$Synuclein Fibrillization by Dopamine Is Mediated by Interactions with Five C-Terminal Residues and with E83 in the NAC Region. *PLoS ONE*, 3(10):e3394, 2008.

[209] J. A. Lemkul and D. R. Bevan. Morin inhibits the early stages of amyloid β-peptide aggregation by altering tertiary and quaternary interactions to produce 'off-pathway' structures. *Biochemistry.*, 51(30):5990–6009, 2012.

[210] J. A. Lemkul and D. R. Bevan. Destabilizing Alzheimer's Aβ(42) protofibrils with morin: mechanistic insights from molecular dynamics simulations. *Biochemistry.*, 49(18):3935–46, 2010.

[211] D. Dibenedetto, G. Rossetti, R. Caliandro, and P. Carloni. A Molecular Dynamics Simulation−Based Interpretation of Nuclear Magnetic Resonance Multidimensional Heteronuclear Spectra of α-Synuclein.Dopamine Adducts. *Biochemistry.*, 52(38):6672–6683, 2013.

[212] W. Y. Choy and J. D. Forman-Kay. Calculation of ensembles of structures representing the unfolded state of an sh3 domain. *J. Mol. Biol.*, 308(5):1011 – 1032, 2001.

[213] G. Nodet, L. Salmon, V. Ozenne, S. Meier, M. R. Jensen, and M. Blackledge. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *J. Am. Chem. Soc.*, 131(49):17908–17918, 2009.

[214] P. Bernado, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun. Structural characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *J. Am. Chem. Soc.*, 129(17):5656–5664, 2007.

[215] L. J. Smith, K. A. Bolin, H. Schwalbe, M. W. MacArthur, J. M. Thornton, and C. M. Dobson. Analysis of main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil conformations. *J. Mol. Biol.*, 255(3):494 – 506, 1996.

[216] A. K. Jha, A. Colubri, K. F. Freed, and T. R. Sosnick. Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl. Acad. Sci.*, 102(37):13099–13104, 2005.

[217] P. Bernadó, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl. Acad. Sci.*, 102(47):17002–17007, 2005.

[218] M. Fuxreiter, I. Simon, P. Friedrich, and P. Tompa. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.*, 338(5):1015 – 1026, 2004.

[219] K. Sugase, H. J. Dyson, and P. E. Wright. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature.*, 447(7147):1021–5, 2007.

[220] J. Michel and R. Cuchillo. The impact of small molecule binding on the energy landscape of the intrinsically disordered protein c-myc. *PLoS ONE*, 7(7):e41070, 07 2012.

[221] S. A. Jónsson, S. Mohanty, and A. Irbäck. Distinct phases of free α-synuclein–A Monte Carlo study. *Proteins:Structure, Funct. Bioinforma.*, 80(9):2169–77, 2012.

[222] X. Cong, N. Casiraghi, G. Rossetti, S. Mohanty, G. Giachin, G. Legname, and P. Carloni. Role of prion disease-linked mutations in the intrinsically disordered N-terminal domain of the prion protein. *J. Chem. Theory Comput.*, 9(11):5158–5167, 2013.

[223] M. R. Jensen, K. Houben, E. Lescop, L. Blanchard, R. W. H. Ruigrok, and M. Blackledge. Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: Application to the molecular recognition element of sendai virus nucleoprotein. *J. Am. Chem. Soc.*, 130(25):8055–8061, 2008.

[224] R. Schneider, J.-R. Huang, M. Yao, G. Communie, V. Ozenne, L. Mollica, L. Salmon, M. Ringkjobing Jensen, and M. Blackledge. Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. *Mol. BioSyst.*, 8:58–68, 2012.

[225] J. A. Marsh, C. Neale, F. E. Jack, W. Y. Choy, A. Y. Lee, K. A. Crowhurst, and J. D. Forman-Kay. Improved structural characterizations of the drkn SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure. *J. Mol. Biol.*, 367(5):1494 – 1510, 2007.

Chapter 2

Thesis Scope

The main objective of this thesis is to provide a global picture of protein motions, studying processes that take place at both local and global level. By means of a wide variety of computational methods we have examined them using some representative examples.

1) Local motions

The first part of the thesis is devoted to study the local, fast and small-amplitude, movements of the proteins. Events taking place at the active site level. The main objectives of this part are:

- Study the catalytic effects of the conformational dynamics unravelling the conformational compression effect in NAGK, providing another picture into the controversy between dynamics and catalysis.
- Implement a method (Swarms of Trajectories) to calculate free minimum energy paths in the pDynamo library for subsequent application in enzyme catalysis.
- Study of decarboxylation due to high radiation damage in proteins (concretely in LDH), finding a methodology properly describing a charge transfer produced by synchrotron techniques for subsequent application in QM/MM studies.

2)Global motions

The second part of the thesis is aimed to study global, slow and large-amplitude motions specially focused on IDPs, defined by the cooperativity effects of the secondary structure elements and the characterization of conformational ensembles. Events that take place at the structure level. The main objectives of this part are:

- Provide a new tool to refine Protein conformational Ensembles based on Residual Dipolar Couplings.
- Study the cooperativity effect of the secondary structure elements in protein ensembles providing a new view to visualize its contribution.

Chapter 3

Methodology

3.1 Quantum Mechanical Methods

Quantum Mechanical methods (QM) are the most rigorous and suitable framework to describe a molecular system at the atomic level. *Ab initio* methods aims to solve the time-independent Schrödinger equation finding the wave function which concentrates all the information of the microscopic system.

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

where $\hat{H}$ is the non-relativistic, non-magnetic electronic Hamiltonian, which consists of five operators: kinetic energy of the electrons, kinetic energy of the nuclei, nuclei-electrons coulomb attraction, electron-electron repulsion and nuclei-nuclei repulsion.

$$\hat{H} = -\frac{1}{2}\sum_{i}\nabla_i^2 - \sum_{A}\frac{1}{2M_A}\nabla_A^2 - \sum_{i}\sum_{A}\frac{Z_A}{r_{iA}} + \sum_{i}\sum_{j>1}\frac{1}{r_{ij}} + \sum_{A}\sum_{B>A}\frac{Z_A Z_B}{r_{AB}} \tag{3.2}$$

where r_{iA} is the distance between i electron and A nuclei, r_{ij} the distance between the electron i and j and r_{AB} the distances between nucleus A and B. M_A is the ratio of the mass of nucleus A to an electron and Z_A the atomic number of nucleus A. The Laplacian operators ∇_i^2 and ∇_A^2 involve differentiation with respect the coordinates i^{th} electron and the A^{th} nucleus.

The complexity of this problem lies on the correlation between the nuclei and electrons of the system. An analytical solution of Eq. 3.2 is not possible, and some approximations have been made to avoid the problem, leading to different accuracy and computational cost.

The Born-Oppenheimer approximation

The nuclei of molecular systems move slower than the electrons. If the electronic motion is instantaneous compared to the nuclei, the Born-Oppenheimer approximation separates the wave function in two parts computing electronic energies over *fixed* nuclear positions.

$$\Psi(\mathbf{r}_i;\mathbf{R}_j) = \Psi_e(\mathbf{r}_i;\mathbf{R}_j)\Psi_N(\mathbf{R}_j) \tag{3.3}$$

This approximation, means that the nuclear kinetic energy term is independent of the electrons, and thus cancelled, the correlation in the attractive electron-nuclear potential energy term is eliminated, and the intra-nuclear repulsion is reduced to a constant parameter ($\mathbf{q}_k$), dependant of the system geometry, added to the electronic energy term. Thus the *electronic* Schrödinger equation is taken to be:

$$(\hat{H}_{el} + \hat{V}_N)\Psi_{el}(\mathbf{q}_i;\mathbf{q}_k) = E_{el}\Psi_{el}(\mathbf{q}_i;\mathbf{q}_k) \tag{3.4}$$

where $\hat{H}_{el}$ include the electronic terms of Eq. 3.1, $\hat{V}_n$ is the nuclear-nuclear repulsion and $\mathbf{q}_i$ are the electronic coordinates, which are independent variables

From this approximation emerges the concept of Potential Energy Surface (PES), the surface defined by the electronic energy (E_{el}) over all possible nuclear coordinates (potential energy). This idea, is of central interest in Computational Chemistry, and will be addressed in more detail in section 3.4.

The Electronic problem

The Born-Oppenheimer approximation simplifies the Schrödinger equation but it is not enough to analytically find a wave function of the molecular system. If we neglect the electron-nuclear correlation, the correlation between electrons is still a problem for poly-electronic systems. Many computational methods addressed to find a solution to this electronic problem, are based on the Hartree-Fock method.

3.1.1 Hartree-Fock method

Approximations aimed to solve the inter-electronic interactions problem are needed. The Hartree-Fock (HF) method is not only a very useful approximation itself, but also the basis of other accurate models of molecular electronic structure. Let us assume a system of N non-interacting electrons. Within this context, 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 de product of the N monoelectronic wave functions, known as the Hartree Product.

This product, however, does not fulfil the antisymmetry principle that describes the behaviour of electrons and other fermions. The exact wave-function does not only have to satisfy the Schrödinger equation, it is also must to be antisymmetric. This requirement is enforced using the Slater determinants, where each row correspond to an electron and each column to a monoelectronic orbital with a given spin, known as spin-orbital χ_i. Using the exact Hamiltonian, the h_i operator have a set of eigenfunctions that we can take to be a set of spin orbitals χ_j.

Applying the exact Hamiltonian to the Slater determinant, with a closed-shell configuration, the Energy takes the form:

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

where H_{ii} corresponds to the kinetic energy 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 is consequence of the Pauli (antisymmetry) principle and reflects the reduced probability of finding two electrons of the same spin close to each other.

According to the variational principle, the best wave-function is the one with the lowest energy, and the simplest antisymmetric wave function that can be used for describe the ground state of a N-electron system is a single Slater determinant. In order to find the best poly-electronic wave functions described by a Slater determinant the energy as expressed in Eq. 3.5 has to be minimized. If this minimization is done respect to the molecular orbitals, subject to the constraint that the molecular orbitals are orthonormal, the Hartree-Fock (HF) equations are obtained.

$$\hat{f}\chi(x_i) = \epsilon\chi(x_i) \tag{3.6}$$

where $\hat{f}_i$ is the mono-electronic fock operator, that is defined for each electron i as:

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

where J_j and K_j are the one-electron Coulomb and Exchange operators. Within this mono-electronic Hamiltonian, the kinetic energy and nuclear attraction are strictly one-electron operator, but Coulomb and exchange operators, are effective operators in the average field of the remaining electrons. Electrons only feel an effective potential created by the rest of electrons and do not interact instantaneously, i.e. their motion is not correlated. The fock operator depends therefore on the solution of the Eq. 3.6. In practice, to solve them, is convenient to expand the orbitals by means of the Linear Combination of Atomic Orbitals (LCAO). Roothaan and Hall proposed to use an orbital basis set and therefore transform the Hartree-Fock equations into linear equations, where the variational parameters are the linear coefficients of the expansion[1,2].

The resulting equations are known as Roothaan-Hall equations. The solution of the equations is achieved iteratively starting from an initial guess of the solution until convergence. For this reason the Hartree-Fock method is also called the *Self-Consistent field* (SCF) method. Starting from an initial guess of orbitals, one can calculate the average field seen by each electron and then solve Eq. 3.6 for a new set of spin orbitals, that are used for calculate new fields. The process is repeated until self-consistency is reached.

Correlation Energy

The HF assumptions imply a huge progress to carry out molecular orbitals (MO) calculations. However, as HF neglect the electron correlation in the calculation of the inter-electronic interactions, it can have important chemical consequences when it comes to determining accurate wave functions and molecular properties derived there-from. A consequence of that, the energy difference between that obtained with the Hartree-Fock method E_0, in the limit of an infinite basis set, and the exact, non-relativistic, energy of the system ζ_0 is named *correlation* energy.

$$E_{corr} = \zeta_0 - E_0 \tag{3.8}$$

Subsequent improvements are aimed to incorporate the correlation energy, including more Slater determinants, for improve the wave function in order to obtain the exact energy, and also to reduce the computational cost, the dynamical bottleneck of the method.

Electron correlation is frequently divided into *dynamic* and *non dynamic*(static) correlation. The first one, arises from a Hartree-Fock wave function that is improved by small contributions of many other determinants representing alternative configurations, reflecting the inter-dependence of the motion of electrons. The second one, is related to wave functions in which the contributions of few determinants dominate the description of the wave function; typical of molecules with

nearly degenerate Slater determinants, in which different electronic configurations are necessary for the description of the system.

The Hartree-Fock method can be considered as a bifurcation point, from which emerge two different ways for compute the wave function.

- **Post Hartree-Fock methods**. These methods use the Hartree-Fock wave function as a starting point toward finding an improved wave function and recovering the correlation energy. The computational demand of this alternative is notoriously higher.
- **Semi-Empirical methods**. These methods simplifies the Hartree-Fock calculations by parameterizing integrals, against experimental data, with the aim to make the calculations much faster. In some cases can increase the accuracy of Hartree-Fock.

3.1.2 Post Hartree-Fock methods

There are three main commonly called post Hartree-Fock methods: Perturbation Theory, Configuration Interactions and Coupled Cluster theory. Here, we are going to explain only the one used in this thesis, the Perturbation theory.

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. This theory is based on that the '*true*' Hamiltonian operator $\hat{H}$ is expressed as the sum of the more tractable '*zeroth-order*' Hamiltonian H_0 (for which a set of molecular orbitals can be obtained) and a perturbation term, V.

$$\hat{H} = \hat{H}_0 + \lambda V \tag{3.9}$$

where λ varies from 0 to 1 and allows the expression of the wave function and the energy as a Taylor expansion of increasing order corrections.

A variant of this theory is the so-called Moller-Plesset (MP-n) theory, in that the '*zeroth-order*' Hamiltonian operator is a lineal combination of Fock operators. The eigenvalue of this Hamiltonian is the sum of the energies of the occupied Hartree-Fock orbitals. This leads to a '*error*' because each orbital energy includes the electron-electron repulsion of the occupying electrons with all of the other electrons, being this repulsion counted twice. This does not correspond with the Hartree-Fock energy, so the perturbation term V has to correct this double counting and include the '*true*' Hamiltonian's repulsion term.

$$V = \sum_{i}^{occ.} \sum_{j>i}^{occ.} \frac{1}{r_{ij}} - \sum_{i}^{occ.} \sum_{j} (J_{ij} - \frac{1}{2} K_{ij}) \tag{3.10}$$

The (MP1) first order correction to the zeroth-order energy does not advance beyond the Hartree-Fock level in determining the energy, in fact returns the Hartree-fock energy. Thus, we must consider, at least, the second order correction (MP2) to recover the correlation energy, which is computed as:

$$E_{MP2} = E_{HF} \sum_{i}^{occ.} \sum_{j>i}^{occ.} \sum_{a}^{virt.} \sum_{b>a}^{virt.} \frac{[(\phi_i \phi_j \mid \phi_a \phi_b) - (\phi_i \phi_a \mid \phi_j \phi_b)]^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b} \tag{3.11}$$

Eq. 3.11 shows that the approximation to the correlation energy is made by considering many excited configurations, wich implies 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, showing a convergence behaviour as a function of n (MP1 = HF, MP2, MP3, MP4).

Furthermore there are other approximations based on Moller-Plesset. Specially interesting is one, used in this thesis, that is based on MP2: the spin-component-scaled MP2 (SCS-MP2) method[3,4]. It was 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. *In this thesis we have employed the SCS-MP2 method to perform single point calculations over different images of reaction profiles calculated by QM/MM methods*

3.1.3 Semi-Empirical methods

The semi-empirical methods are based on the parameterization of some integrals against experimental data, with the objective of reducing the computational cost and allowing the increase of the size of the system, that is one of the Hartree-Fock bottlenecks. Bigger systems imply an increment of the number of integrals to solve as N^4. Semi-empirical methods in order to reduce the computational cost only compute a fraction of these integrals, parameterizing the rest of them. These parameters reproduce thermochemical and structural experimental data.

All semi-empirical methods ignore the core electrons because of they are 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 these methods lie in the number of neglected (not-computed) integrals and the way they are parameterize.

ZDO (Zero Differential Overlap)

The most of the semiempirical methods are based upon this approach. In this approximation, the overlap between pairs of different orbitals is set to zero for all volume elements $d\nu$.

If the ZDO approximation is applied to the two-electron repulsion integral $(\mu\nu|\lambda\sigma)$, the integral will be equal to zero. In addition all three and four-center integrals are neglected. If we apply this approach to all orbital pairs, the Roothaan-Hall equations could be obtained in a very simplified version, at least for a closed-shell regime.

CNDO (Complete Neglect of Differential Overlap) [5,6]

The CNDO was the first approach that implemented the ZDO approximation. Computationally, it represents a vast simplification of Hartree-Fock theory. It reduces the number of two-electron integrals having non-zero values from formally N^4 to simplify N^2, because of the number of integrals to compute is dramatically reduced and the remaining ones are already parameterized and do not require explicit calculation.

INDO (Intermediate Neglect of Differential Overlap) [7]

Coming from CNDO emerges the INDO model. The key change from CNDO is that the integrals between different types of orbitals are distinguished and adopt different parameterized values.

NDDO (Neglect of Diatomic Differential Overlap)

This approach complements CNDO and INDO by adding flexibility to the description of the two-center two-electron integrals. In this approximation only differential overlap between atomic orbitals in different atoms is neglected. 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.

MNDO (Modified Neglect of Differential Overlap) [8]

Based on the NDDO formalism, Dewar and Thiel reported the MNDO method. They suggested to modify the two-center two-electron integrals as interactions between multipoles replacing the continuous charge clouds, thus simplifying the calculation and reducing the computational cost.

The nuclear repulsion energy, named as core-core term, in NDO methods has to be modified, since the electron-electron terms do not compensate repulsion between nuclear charges and, at long distances, uncharged atoms or molecules

experience a net repulsion. The way this is corrected underlies the difference between the NDDO-based methods.

$$V_{NN}(A,B) = Z'_A Z'_B (s_A s_B | s_A s_B)(1 + e^{-\alpha_A R_{AB}} + e^{-\alpha_B RAB}) \tag{3.12}$$

where Z'_A denotes that the nuclear charge has been reduced by the number of core electrons and α exponents are taken as fitting parameters.

- AM1 [9]

  One critical limitation of MNDO is that it does very poorly in the prediction of hydrogen bonding geometries and energies.

  Recognizing this to be a major drawback, particularly with respect to modelling systems of biological interest, Dewar and co-workers modified the functional form of their NDDO model. To alleviate this problem, they modified the nuclear repulsion term by adding up to four gaussian functions to each atom, creating the so called Austin Model 1 (AM1). The nuclear repulsion energy between any two nuclei A and B is computed as:

$$\begin{aligned} V_{NN}^{AM1}(A,B) = V_{NN}^{MNDO}(A,B) + \frac{Z'_A Z'_B}{RAB} * \\ * \left(\sum_k a_{kA} e^{-b_{kA}(R_{AB}-c_{kA})^2} \sum_k a_{kA} e^{-b_{kB}(R_{AB}-c_{kB})^2} \right) \end{aligned} \tag{3.13}$$

  where k is between two and four depending on the atom, and the constants a_k,b_k and c_k are fitted to molecular data.

  Although AM1 is one of the most broadly used methods in a wide variety of applications, the parameterization process had not be optimal, so Stewart reported a reparameterization fo AM1 adding two gaussian functions for each atom in the so called PM3 [10] method. RM1 [11] is another more recent reparameterization that keeps the AM1 core-core expression and gives better results than AM1 and PM3.

d orbitals in MNDO models

With only s- and p- functions included, the MNDO/AM1/PM3/RM1 methods are unable to treat a large part of the periodic table, specially from the third row and lower. Furthermore from *ab-initio* simulations it is know that d-orbitals significantly increase the flexibility and improve the description of the wave function of hypervalent atoms such as phosphorus. It is patently obvious that such orbitals need to be included.

Thiel and Voityuk[12–14] described the first NDDO model with d orbitals, called MNDO/d. Following the same philosophy as MNDO, new one- and two- electron integrals involving d-orbitals were parameterized. Based on the MNDO/d formalism, extensions of AM1 have been reported by multiple groups. Voytiuk and Rösch first described an AM1/d parameter set for Mo[15] and Lopez and York[16] reported a parameter set for P. Winget and coworkers described an alternative model, named AM1*, that adds d orbitals to P, S, and Cl[17]. 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.

- AM1/d-Phot[18]

  Another AM1 based model using d-orbitals, was described by Nam and coworkers, called AM1/d-PhoT for P, H and O atoms involved in phosphoryl transfer reactions. To avoid the overstabilization of hypervalent structures given by AM1, the core-core term includes a parameter (g_{scale}) that attenuates the artificially attractive interactions involving P atoms.

  In this thesis we have employed the AM1 and AM1/d-Phot semi empirical methods to perform different enzyme related calculations.

3.1.4 Basis sets

- STOs and GTOs

  The basis set is the group of mathematical functions from which the wave function is constructed. These functions are usually based on the molecular orbitals of the hydrogen atom such as the Slater type orbitals (STOs) that are centered at the nuclei to construct the wave-function of the molecules and its radial part is described as:

$$R(r) = Nr^{n-1}e^{-\zeta r} \tag{3.14}$$

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

  However, practically, STOs leads to an inefficient evaluation of three- and four-center integrals, given the large number of integrals to compute. An alternative to STOs are the Gaussian Orbitals, GTOs:

$$R(r) = Nx^i y^j z^k e^{-\alpha r^2} \tag{3.15}$$

where α determines the width of the Gaussian, x,y,z are the cartesian coordinates and the integers j,i,k, determine the type of orbital. However, GTOs even an alternative to STOs present problems regarding their poor representation at short and large nuclear distances. This is the reason why semiempirical methods employs STOs discarding three- and four-center integrals.

STOs and GTOs are usually combined, keeping the accurate radial shape of the first ones, and the computational efficiency of the second ones. In these cases a linear combination of GTOs, called *primitives* is used to represent a given STO. When a basis function is defined as a linear combination of GTOs is called *contracted*.

- Split-Valence basis set

  Another commonly used type of basis set are the called *Split-Valence* developed by Pople and coworkers. They are employed given that valence orbitals are those involved in chemical bonding and, thus, are more sensitive to the environmental changes than core orbitals. Such basis contains one contracted basis function for describe core orbitals and double- or triple-ζ basis sets for valence orbitals. One example is the popular 6-31G basis set.

- Polarization and Diffuse functions

  Moreover to increase the flexibility and thus improve the description of the molecular orbitals, functions with higher angular momentum than that the valence orbitals, called *polarization* functions are employed. *Diffuse* functions are also used to add more flexibility, enabling the basis set to locate electron density far from the nucleus, specially necessary for negatively charged atoms. For instance, the 6-31+G(d) is a double-ζ basis set with polarization and diffuse functions.

- Ahlrichs basis set

  Ahlirchs basis sets follow the same philosophy of Pople basis sets but are optimized to higher extent, giving the same accuracy but using smaller basis sets and thus reducing the computational cost. Some examples are the Ahlrichs split valence plus polarization SVP and diffusion SVP+ or the Ahlrich's triple-ζ TZV(2d) basis set, used in this thesis.

  In general for a proper use of Post Hartree-Fock methods it is important to use larger basis set with diffuse and polarization functions of high angular momentum, to recover a high percentage of the correlation energy. Usually these basis sets imply a huge computational cost, so a common strategy is perform single point calculations using larger basis sets over molecular

geometries energy minimized with a smaller basis set. Due to the geometry parameters are less sensitive than the energy to the size of the basis set, this approach achieves a good balance between computational cost and energy and structure determination. *In this thesis we have employed different basis set schemes, Split-Valence and Ahlrichs basis sets, with polarization and diffuse functions to perform QM/MM optimizations as well as single point calculations.*

3.1.5 Density Functional Theory methods

In view of the poor predictions of chemical bonds and molecular properties afforded by HF approximation and semi-empirical methods and the high computational cost of post HF approaches, it is beneficial to seek out methods that circumvent the need to represent the many-body electronic wavefunction. Within this necessity appeared the Density Functional Theory (DFT).

Density Functional Theory based methods are an alternative to the ab-initio methods to introduce the correlation effects into the solution of the electronic Schrödinger equation. With respect to previous attempts DFT methods follow an alternative route. In DFT what fully determine the properties of a molecular system is the electronic density as demonstrated by the Hohenberg-Kohn theorems[19].

Hohenberg-Kohn theorems

The first (*existance*) theorem establishes the existance of a 1:1 relation between the electron density and the wavefunction. In this regard, the system energy depends exclusively on the density and thus the energy is a functional of the density. Anyway, this first theorem only prove the existence of the functional, but does not indicates its expression.

The second (*variational*) theorem proves that the electron density follows the variational theorem as the wavefunction, so the better the approximation to the exact electron density the lower the associated energy. The Hohenberg-Kohn theorems provide an alternative way to the Schrödinger equation, but the lack of knowledge about the exact form of the functional made the DFT theory impractical.

Kohn-Sham equations

It is not until 1965, when Kohn and Sham[20] found a practical way to find the system properties directly from the density, that the breakthrough in DFT-based methods started. The crucial 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 adopts the following form:

$$E[\rho(r)] = T_{ni}[\rho(r)] + V_{ne}[\rho(r)] + V_{ee}[\rho(r)] + \triangle T[\rho(r)] + \triangle V_{ee}[\rho(r)] \tag{3.16}$$

where T_{ni} refers to the kinetic energy of the non-interacting electrons, V_{ne} to the nuclear-electron interaction, V_{ee} to the classical electron-electron repulsion, $\triangle T$ is the correction to the kinetic energy due to the inter-electronic interaction, and $\triangle V_{ee}$ represents the quantum corrections to the electron-electron repulsion energy. The corrections to the kinetic energy and inter-electronic repulsions are gathered into the so-called Exchange correlation term $E_{xc}\rho(r)]$.

The resulting Kohn-Sham equations are very similar to the HF ones:

$$h_i^{KS}\chi_i = \epsilon_i\chi_i \tag{3.17}$$

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

$$h_i^{KS} = -\frac{1}{2}\nabla_i^2 - \sum_k^N \frac{Z_k}{|r_i - R_k|} + \int \frac{\rho(r')}{|r_i - r'|}dr' + V_{xc} \tag{3.18}$$

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

$$V_{xc} = \frac{\delta E_{xc}}{\delta \rho} \tag{3.19}$$

The main difference between this method and Hartree-Fock is that this last one is an approximate theory whereas Kohn-Sham method provides the exact solution for the exact $E_{xc}[\rho(r)]$ functional.

Exchange-Correlation functionals

However the exact form of this functional is not known and thus some approaches have been developed to calculate the exchange and correlation energy terms, gathered into the exchange-correlation functional. These approaches differ in using either only the electron density (LDA) or the electron density and it's gradients (GGA). The hybrid functionals are another approach, in which mixtures of DFT and Hartree-Fock exchange energies are used.

All the functionals are composed by mathemathical expressions and parameters that are fitted to experimental data because of DFT could be regarded as semi-empirical methods, although their number of parameters is much lower than the actually classified as semi-empiricals.

- Local density approximation functionals (LDA)

  The *LDA* term was originally used to indicate any functional where E_{xc} for some position r is determine from ρ exclusively, i.e, for the 'local' value of ρ. In these functionals, the analytical expression of the exchange functional was derived by Slater,

$$E_x[\rho r] = -\frac{9\alpha}{8}\left(\frac{3}{\pi}\right)^{1/3}\int \rho^{4/3}(r)dr \tag{3.20}$$

  and has a simple form, in contrast to the most wide used correlation energy functional which corresponds to the mathematical model of Vosko, Wild and Nusair (VWN)[21].

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

- Generalized Gradient Approximation (GGA)

  Because the electron density of a molecule is not uniform, it is reasonable to improve the LDA approximation making it depend not only on the local density, but on the extent to which the density is locally changing, the gradient of the density. The functionals that improve LDA approaches using the gradient of the density, are known as *GGA* functionals.

$$\epsilon_{xc}^{GGA}[\rho(r)] = \epsilon_{xc}^{LDA}[\rho(r)] + \nabla\epsilon_{xc}\left[\frac{|\nabla\rho(r)|}{\rho^{4/3}(r)}\right] \tag{3.21}$$

  where ϵ_{xc} is defined as the energy density, thus the exchange functional $E_{xc}[\rho(r)]$ is defined as:

$$E_{xc}[\rho(r)] = \int \rho(r)\epsilon_{xc}[\rho(r)]dr \tag{3.22}$$

  Despite the improvements respect to LDA, GGA also has drawbacks, that can be overpassed by including an additional correction to the GGA approach using 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 can be computed from the non-interacting system according to the following expression:

$$E_{xc} = (1 - a)E_{xc}^{DFT} + E_{xc}^{HF} \tag{3.23}$$

The basic idea behind the hybrid functionals is approximate the E_{xc} by mixing exchange energies calculated in an exact manner (adding part of the exact Hartree-Fock exchange energy) with those obtained from DFT (GGA) methods in order to improve the results of the pure DFT. Probably the the most widely used functional of this type is the B3LYP[22–24], although other useful functionals such as BHLYP[24,25] or mPW1PW91[26] were also developed. Indeed, as all these functionals have a huge parameter dependence and thus are very system specific, a plethora of them have been designed.

The inclusion of HF exchange in a hybrid functional, present advantages as the compensation of the underestimation by pure functionals of the importance of ionic terms in describing polar bonds[27], or the improvement in the description of the energy barrier. The GGA functionals tend to underestimate the barrier and HF, on the other hand, overestimate it. Thus the addition of HF could act as a back-titration to the barrier description accuracy. For instance, following this idea, the MPW1K[28] was optimized for properly describing the kinetics of H-atom abstractions or the mPW1N[29] that was developed for halide/haloalkane nucleophilic substitution reaction.

However, despite of the mentioned improvements, current functionals still present important shortcomings. One of the main limitations is their inaccuracy describing long-range dispersion interactions, because of in current exchange-correlation functionals the energy depends on the local density and its derivatives, which are also local, so they cannot describe accurately the electron correlation at long distances.

Several approximations for including dispersion interactions have been made. For instance, a modified version of the exchange functional by Perdew and Wang (PW) was obtained by Adamo and Barone in the mPWPW91 functional, used in this thesis to describe phosphoryl transfer reactions. It gives remarkable results both for covalent and noncovalent interactions in a quite satisfactory theoretical framework encompassing the free electron gas limit and most of the known scaling conditions[26]. Nowadays, however, the M06 family of functionals[30] are among the most accurate and widely used for describing non-covalent interactions as well as the kinetics and thermochemistry. Within the dispersion interaction context, in recent years, there have been important advances due to the development of the DFT-D methods[31,32] by including semi-classical (MP2-like) corrections of dispersion interactions to standard exchange-correlation functionals, the so-called DHDF double hybrid density functionals.

All these subsequent developments, and the ones that are coming, have contributed to increase noticeably the number of functionals currently available.

In this thesis to study enzyme catalysis reactions, we have performed several DFT QM/MM and single point calculations using the BHLYP and the mPWPW91 functionals.

3.2 Molecular Mechanics

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. Furthermore the explorations of complex systems with degrees of freedoms as proteins by QM methods is unaffordable because the computational expense makes this exploration unachievable. Thus methods that require less computational resources are needed.

Molecular Mechanics (MM) methods calculate the interaction potential of the particles using a force field, reducing the computational cost by lowering the cost of the energy calculation. A force field is a set of parametrized equations that allow the evaluation of the energy and the gradient of the system with a low computational cost. Nevertheless they have a limitation: they are not able to reproduce the formation and rupture of chemical bonds. This happens because they are parameterized with mathematical expressions that only depend on the nuclear positions and ignore the electrons. Due to this limitation, the combination of MM methods with QM methods, is a good and useful option (see section 3.3).

Molecular mechanics force fields express the energy of a molecular system as a summation of different contributions that are expressed as mathematical functions. The parameters of these functions have been optimized against experimental data and QM calculations. The most used are the AMBER[33], CHARMM[34], GROMOS[35] and OPLS[36] force fields (encompass within Class I or diagonal force fields[37]), from that exists different versions (extended atom force fields) aimed to obtain more realistic results[38,39]. The energy of the system in any MM force field is divided into bonding and non-bonding terms ($E = E_{bonding} + E_{non-bonding}$). As the common form of potential energy, molecular mechanics assumes additivity of energy potentials thus could be expressed (for Class I force fields) as:

$$V(r) = V_{str} + V_{ben} + V_{tors} + V_{imp} + V_{cross} + V_{vdw} + V_{elec} \tag{3.24}$$

being the five first terms, stretching, bending, torsion, impropers and cross terms, such as Urey-Bradley, that constitutes the bonding terms, and the last ones Van der Waals interactions and electrostatic interactions encompass within the non-bonding terms. Each force field implement them in its own way.

3.2.1 Bonding-Interactions

The energy associated with bonding terms is computed using functions that model the energy penalties due to deviations of internal coordinates from their reference values.

$$E_b = \sum_{i}^{bonds} K_b(b-b_0)^2 + \sum_{i}^{angles} K_\Theta(\Theta-\Theta_0)^2 + \sum_{i}^{torsions} K\phi(1+cos(n\phi-\delta)) + \\ + \sum_{i}^{impropers} (K_\psi(\psi-\psi_0)^2 + \sum_{i}^{Urey-Bradley} K_{UB}(UB_{1,3}-UB_{1,3,o})^2 \tag{3.25}$$

The first term corresponds to the stretching between each pair of bonded atoms described by an harmonic potential, whose force constant reflects the bond strength. The second term is the angle bending contribution, also modelled by an harmonic potential. The use of a simple harmonic potential is, in principle, enough because there are non significant deviations from the equilibrium position expected. The third term corresponds to proper torsions which model the energy changes due to bond rotations, which are responsible of the main conformational changes of the molecule, indicating the number of minimum energy conformations resulting from the bond rotation. These three first terms are included in all the force fields, however the two last terms are present only in some of them. *In this thesis we have employed the AMBER, GROMOS and OPLS force fields.*

3.2.2 Non-Bonding-Interactions

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

$$E_{nb} = \sum_{VanDerWaals}^{i<j} \left\{ \frac{A_{ij}}{R_{ij}^{12}} - \frac{C_{ij}}{R_{ij}^{6}} \right\} + \sum_{Electrostatic}^{i<j} \frac{q_i q_j}{\epsilon r_{ij}} \tag{3.26}$$

The Van der Waals interactions describe the attraction or repulsion between atoms that are non-bonded. A common way to model them is with the popular 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 simplest model for describe electrostatic interactions is the Coulomb's law, which defines the interaction energy between two point charges separated by a given distance.

Non-bonded interactions represent the most time-consuming part of the MM calculations. The evaluation of these interactions scales as N^2, being N the number of atoms. To alleviate this computational cost some approaches could be employed. For instance using spherical cutoff schemes that restrict the evaluation for all possible pairs to only some of them. Three different cutoff schemes have been developed. In the simplest scheme (truncation), only the interactions within a cutoff distance are computed. However, this introduces discontinuities in the distance-dependent non-bonding interaction energy, and the corresponding forces, leading to potential artifacts. To avoid this problem, there are other schemes aimed to gradually set to zero the distance-dependent interactions: the switch and shift functions. The shift functions alter the interaction energy function, $E(r)$, gradually from the beginning in order to reach the zero value at the cutoff distance, while the 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. These cutoff schemes has to be large enough due to the Lennard-Jones potential present a rapid decay (as $1/r^6$).

On the other hand electrostatic interactions decay much more slowly (as $1/r$), thus the effects of the long-range interactions contributions to the electrostatic energy are non-negligible. Even for non-charged particles, dipole-dipole interactions decay more slowly (as $1/r^3$). Therefore the cutoffs schemes have to be larger than for Van der Waals interactions to account them. 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 section 3.2.5). *In this thesis we have computed the full NB interactions or we have employed switch functions as well as a variant of the Ewald summation method, the Particle Mesh Ewald (PME), depending on the simulation performed.*

3.2.3 Solvent treatment: Explicit solvation

In biomolecular systems, such as proteins or small ligands, the effect of the solvent is very important, thus its modelling is a key aspect. There are two philosophies to do it: the explicit and the implicit (also known as continuum) solvation. In the first one the molecules are placed around the simulated solute (protein) molecule

only accounting for their electrostatic influence on the solute, while in the second one the various physical influences of solvent molecules on the solute (electrostatic (including induction), cavitation, exchange repulsion, and dispersion attraction) are taken into account. The solvation effects are calculated by extra terms added to the force fields. In this thesis only the explicit model have been employed.

The explicit solvation of the system requires the definition of a water model, simple MM models that describe water-mediated polar interactions. These models assume a fixed geometry for the water molecules, which are treated as rigid entities, and only consider non-bonding terms. They differ in the number of interaction sites. For instance the SPC (Single Point Charge) water model[40] presents three interaction sites as well as the TIP3P[41], in which a point charge is defined at the oxygen and two hydrogen atoms. There are also 4 interaction sites models as TIP4P[41] and even with 5 like TIP5P[42] model. However has to bear in mind that an increment of the number of interaction sites imply an increment of the computational cost. Furthermore current force fields have been parameterized in conjunction with a given water model, thus has to take it into account to choice the proper water model. The use of a water model not compatible with the force field may lead to some inconsistencies. *In this thesis we have employed the TIP3P and the SPC water models.*

3.2.4 Periodic Boundary conditions (PBC)

The simulation of a solute immersed in a solvent is usually done under Periodic Boundary Conditions (PBC). That means immerse the system into a unit cell that is infinitely replicated in the three spatial dimension avoiding, in principle, surface effects. Thus solvent molecules at the edge of the cell interact with solvent molecules as a bulk.

The use of PBC imply following the *minimium-image* convention, which means that when a particle crosses the boundary of the unit cell, an image of that particle enters to replace it, conserving the total number of particles in the cell. Within this approximation, non-bonding interactions are limited to use a cutoff of a maximum Length of L/2 where L is the length of the dimension of the box. Depending on the shape of the system different unit cell geometries can be used to construct the lattice. The cubic shape is the most broadly used, however others are more compact for a given thickness of the water layer reducing the amount of solvent molecules needed in the system, as for example rhombic dodecahedron. Thus the selection of the simulation box is important to make the calculation more computationally efficient. *In this thesis we have employed orthorhombic and rhombic dodecahedron boxes.*

3.2.5 Ewald summation method

PBC are used by a lot of simulation schemes as the *Ewald summation method*, which employ them advantageously. The Ewald sum was first devised by Ewald to study the energetics of ionic crystals. This scheme use PBC to compute long-range electrostatic interactions in a more precise form than using cutoff schemes. This technique calculates the electrostatic energy of the system with an infinite number of periodic images adopting a *reciprocal-space technique*. In this method, a particle interacts with all the other particles in the simulation box and with all of their images in an infinite array of periodic cells. The position of each image box (simply assumed to be a cube of side L containing N charges) can be related to the central box by specifying a vector, each of whose components is an integral multiple of the length of the box.

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

$$V = \frac{1}{2} \sum_{|\mathbf{n}|=0}^{\infty} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{|\mathbf{r}_{ij} + \mathbf{n}|} \tag{3.27}$$

where $\mathbf{n}$ is the vector $(\mathbf{n}_x L, \mathbf{n}_y L, \mathbf{n}_z L)$, being n_x , n_y and n_z integers and L the size of the box.

The problem comes from that the summation in Eq. 3.26 converges extremely slowly and is *conditionally convergent*. A conditionally convergent series contains a mixture of positive and negative terms, that when are accounted alone give divergent series. Thus the order in that they appear is important. An additional problem is that Coulomb interaction can vary rapidly at small distances. Ewald devises a trick to convert this sum into two series, each of which converges much more rapidly, which essentially is based on perform one summation in the real space and another in the reciprocal space.

The trick, more in detail, is based on surrounding each charge of the system with a Gaussian charge distribution of opposite sign. Thus the summation arising from point-charges and Gaussian charges is convergent and is carried out in the real space. Then, the 'neutralizing' Gaussian charge distribution is re-neutralized by a second Gaussian charge distribution over the infinite summation that is performed in the reciprocal space by Fast Fourier Transformation.

In practice, there are implementations of this method that improve the performance of the reciprocal sum, as the Particle Mesh Ewald method (PME) [43] which scales as $Nlog(N)$ and finds wide application in MD simulations. Some linear-scaling implementations have also been done for hybrid QM/MM calculations [44].

3.3 Hybrid Quantun Mechanics / Molecular Mechanics

The systems of chemical interest in computational 'biosciences'(biology, biochemistry and biophysics) and catalysis, are often condensed phase systems with many thousands of participating atoms[45]. The usage of QM methods to describe these systems most of the time implies an unaffordable computational demand as we have already commented. MM methods, on the other hand, are only an efficient alternative to QM methods if there are not bond-breaking/formation events and/or other related electronic processes. From this situation arises a necessity to develop new methods to treat these systems computationally, and the most useful and logical tool are the QM/MM methods, a quantum mechanics calculation embedding into a classical molecular mechanics model of the environment. Typically the events aimed to study tend to occur in a small part of the whole system, such as enzyme active sites, thus the small reactive region is described with QM methods and the remaining part of the system with MM force fields.

The QM/MM approaches were first introduced by the seminal work of Warshel and Levitt in 1976[46] and along the years several distinct schemes have been devised. Within the QM/MM framework the Hamiltonian is defined as:

$$H = H_{QM} + H_{MM} + H_{QM/MM} \tag{3.28}$$

where H_{QM} describes the interaction between the quantum mechanical particles, H_{MM} accounts for the interaction of all particles represented by a MM force field and $H_{QM/MM}$ evaluates the interaction between both QM and MM particles. The most important differences between the exisitng QM/MM schemes arising from the treatment of the QM/MM coupling term.

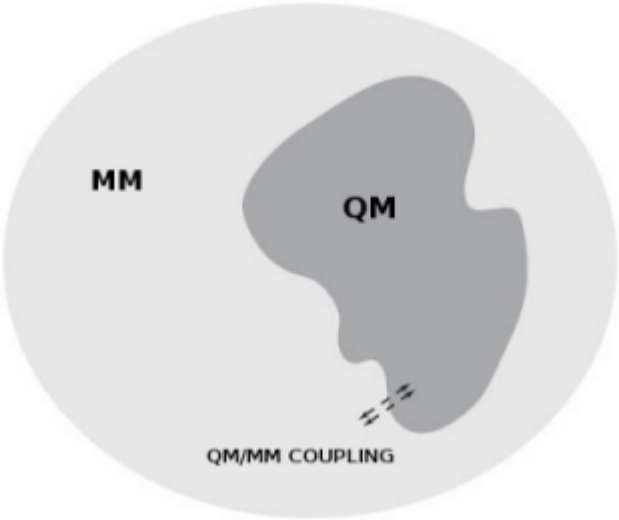


Figure 3.1: Schematic representation of the different components in a QM/MM scheme. Adapted from Field 2007 [47]

QM/MM (electrostatic) coupling schemes

Some schemes have been devised to account the electrostatic coupling between the QM charge density and the charge model used in the MM region. All of them are characterized essentially by the extent of mutual polarization and are classified accordingly as mechanical embedding, electrostatic embedding, and polarized embedding [48]. Here we are only going to explain the second one, because is the only one used in this thesis.

Electrostatic embedding

The electrostatic embedding is the most used scheme in biomolecular simulations. It eliminate the major shortcomings of mechanical embedding by performing the QM calculation in presence of the MM charge model. The MM atomic partial charges are readily available from the force field and their inclusion in the QM Hamiltonian is efficient.

$$H_{QM/MM} = \sum_{i}^{soluteelectrons} \sum_{m}^{MMatoms} \frac{q_m}{r_{im}} + \sum_{k}^{solutenuclei} \sum_{m}^{MMatoms} \left(\frac{z_k q_m}{r_{km}} + 4\epsilon_{km} \left\{ \left(\frac{\sigma_{km}}{r_{km}} \right)^{12} - \left(\frac{\sigma_{km}}{r_{km}} \right)^{6} \right\} \right) \tag{3.29}$$

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. In this scheme the Lennard-Jones contribution avoid that both regions be in excessively close contact as its effect is primarily limited to boundary atoms.

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 a proteic system.

Boundary schemes

The first practical step using QM/MM methods consists on dividing the entire system into an inner, modelled at QM level, and outer, modelled at MM level, regions. For small biomolecular systems this division is trivial because the solute is QM treated, surrounded by MM solvent molecules. For instance if the reactants of a chemical reaction (cofactors, ligands) are not covalently bound to the enzyme and no protein residue is directly involved in the chemical transformation. If it is not the case, as usually happens, the division is turn into a delicate step because it implies the cut of covalent bonds. Some approaches have been devised to treat it, that could be categorized into link atoms, boundary atoms and frozen localized orbitals (see Fig.3.2), being the first and the last one the two most used[45,48]. Here as happened with the coupling schemes we are only going to explain the one used in this thesis, the Link atoms scheme.

Link Atoms

The link-atom method is conceptually simple and the most widely used boundary scheme. In this type of scheme additional atomic centers (L), normally hydrogen atoms, are added to saturate the free valence of the QM atoms bonded to MM atoms of the inner region when the covalent bond between them is cut. The free valence at Q1 created by the QM-MM division is capped by an additional atom that is covalently bonded to Q1 (see Fig. 3.2).

The hydrogens, given that the boundary is usually defined as cutting C-C bonds, are not expected to alter significantly the original environment of the QM atom at the boundary when replace the original carbon atom. These hydrogens are only taken into account by the QM calculations being invisible to the MM calculations. The function of these atoms is cap the QM subsystem, thus are not part of the entire system.

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. There are several alternatives to alleviate it such as 1) delete the one-electron integrals associated with the link atoms, 2) delete the MM point charges in the link region from the Hamiltonian, 3) redistribute the MM atom charge between

their bound MM atoms or 4) using more physically realistic representations such as gaussian charge distributions centered on the MM boundary atoms.

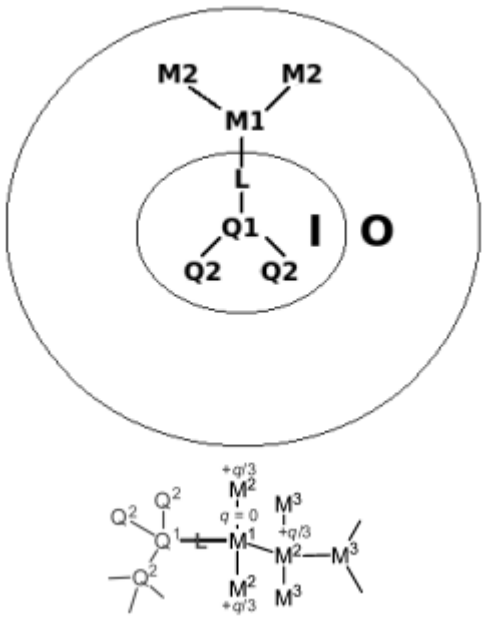


Figure 3.2: Link atoms scheme. I represent the Inner region, O the Outer and B the Boundary region. Adapted from Sherwood 2000[45], Senn and Thiel 2007[48] and Senn and Thiel 2009[49]

In this thesis we have employed pDynamo[50] to performing QM/MM calculations. pDynamo uses a link atom scheme for treat the division of the system within an electrostaic embedding.

3.4 Conformational Sampling

Everywhere in nature, dynamical processes occur constantly. We can state fairly that the execution of dynamics governed by forces is the only work that nature do. For instance biomolecules (whether in a test tube or inside the cell) move in space and change its shape and its size by binding or unbinding another molecules as well as by altering the overall concentration(s) of the system, i.e., changing the pH or the temperature. The way to understand these phenomena and to connect experiments and theory is by the use of statistical mechanics. In that sense thermodynamics will be taught as a natural outcome and statistical mechanics used to transform the detailed view of the microscopic system into thermodynamic magnitudes.

3.4.1 Potential and Free Energy Surfaces

Potential Energy Surface

So far we have described the wide variety of methods available to evaluate the potential energy for a given nuclei configuration. However, to describe a molecular system this is not enough. Although the energy calculation for one or a small number of configurations may sometimes be necessary, it can give only limited information about the properties of a system. As we have commented at the Introduction, according to Frauenfelder, the dynamical behaviour of a protein is closely related to the underlying energy landscape, thus to study a molecular (proteic) system the energy surface has to be taken into account. Under the Born-Oppenheimer approximation the nuclei move throughout a hypersurface, with 3N-6 internal degrees of freedom, whose topology determines the reactivity and other molecular properties of the system, connecting its microscopic and macroscopic (observable) properties. Thus the sampling of the full conformational space through the exploration of this hypersurface is a must.

A potential energy surface (PES) is a theoretical concept used to relate the energy and the geometry of a given system. It describes the energy respect to the position of all the atoms. Mathematically it is described as a multidimensional function of the positions of all the atoms of a given system. The PES defined by the individual terms (i.e. bond stretching, angle bending, torsions, and van der Waals) gives the contribution to the internal energy, but does not say nothing about the entropy (constituting its weakest point). However, the entropy is proportional to the number of states accessible to the system, and thus to the internal energy.

The energy quantifies the molecular interactions of a system and the entropy its structural variation. The study of the free energy, which takes entropy, and usually enthalpy, into account, is a good framework to properly describe biomolecular systems, and to compare against experimental measurements (that usually measure free energies and thermodynamical parameters), although is not always necessary. Specially having into account the high computational cost of the free energy computations compared with the ones based on potential energy. By using statistical mechanics the potential and free energies can be related.

Statistical mechanics: The partition function

In order to treat a collection of molecules (macroscopic systems) in statistical mechanics, a requirement is that certain macroscopic conditions be held constant by external influences. Depending on the constant conditions an ‘ensemble’ is defined. For instance an ensemble where the total number of particles N, the Volume, V and the Temperature, T remain invariant, is called canonical (NVT) ensemble.

In Quantum mechanical methods the wave function is a fundamental function that characterize the microscopic system. In statistical mechanics there is an equivalent function: the partition function.

In non-rigid condensed phase systems (such as proteins) we have to compute the whole partition function through the phase space integral, within the canonical ensembles, over all spatial **r** and momentum **p** coordinates.

$$Z = \frac{1}{h^{3N}} \int_r \int_p d\mathbf{r} d\mathbf{p} e^{-\frac{H_{\mathbf{r},\mathbf{p}}}{K_B T}} \equiv Z(N,V,T) = \sum_i e^{E_i(N,V,T)/K_B T} \tag{3.30}$$

Applied to Eq. 3.30 the usual Hamiltonian expression allows to separate the atomic momenta from the potential energy V,

$$H(r,p) = \sum_i^N \frac{\mathbf{p}_i^2}{2m_i} + V(\mathbf{r}) \tag{3.31}$$

therefore the integral over the phase space (p,q) becomes the configurational integral multiply by a constant:

$$Z \propto \int_r d\mathbf{r} e^{-\frac{V(q)}{k_B T}} \tag{3.32}$$

As our condensed-phase system is a biomolecule (protein) made by atoms, it is necessary to compute the internal molecular motions until the integral of configuration converged.

Free Energy Surface

The reaction free energy (Helmholtz or Gibbs functions) is the magnitude that describe the spontaneity in NVT and NPT (constant Pressure P, Temperature T and number of particles N) ensembles respectively, that is the tendency of molecular systems to associate or react. In addition under the framework of the Transition State Theory (see later in this section) it will describe the kinetics of such process.

Although different both types of free energies are closely related. The Helmoltz free energy is defined as $F(N,V,T) = -k_B T ln Z(N,V,T)$ and the Gibbs free energy as $G(T,P,N) = F + P\langle V \rangle$. Surely G is the correct measure under constant pressure conditions but in biomolecular systems the difference between G and F does not really matter. The difference between G and F can be important in dilute systems, but not for solid and liquid phases and large biomolecules in solution. Basically all of molecular-level biology and experimental biophysics

processes occurs in the condensed phase plenty of molecules. In such conditions molecules are basically attractive and fix their own volumes, thus can be consider a constant volume. Furthermore volume fluctuations in experimental biology systems are tiny on a relative scale.

To going in more detail let us assume a system described by the cartesian coordinates $\mathbf{x} \in R^n$, having into account Eq. 3.32, with a standard equilibrium distribution:

$$p(\mathbf{x}) = Z^{-1} e^{-\frac{V(\mathbf{x})}{k_B T}} \tag{3.33}$$

We assume that there are no constraints in the system, and that the part of the density arising from the momenta has been integrated out [51]. We now introduce M CVs that are functions of x and that can distinguish distinct reacting configurations of the system.

$$\tilde{\mathbf{z}}(\mathbf{x}) = \{\tilde{z_1}(\mathbf{x}), z_2(\mathbf{x}), ...z_M(\mathbf{x})\} \tag{3.34}$$

The free energy, also known as the potential of mean force, associated with $\tilde{z}(\mathbf{x})$ is a function that depends on $z = (z_1...z_M)$ and is calculated as:

$$F(\mathbf{z}) = -k_B T(Z^{-1} \int_{R^n} e^{-\beta V(\mathbf{x})} \delta(z_1 - \tilde{z}_1(\mathbf{x}))...\delta(z_M - \tilde{z}_M(\mathbf{x}))d\mathbf{x}) \tag{3.35}$$

This constitute the M-dimensional free energy surface described in the Introduction.

As the free energy is related to the potential energy, the FES is also related to the PES. To establish a clear connection between them, we must think in terms of representative samples of the local minima of the PES, from which a free energy surface is projected by averaging over the reaction coordinates [52].

Based on all the arguments given, one can reasonably conclude that exploring the FES is the best option to study biomolecular processes. However sometimes it is even better to use potential energy instead of free energy. For instance where the enthalpic and entropic contributions to the free energy barrier are not significant, and/or when an accurate description of the electronic structure is more advantageous. In these cases potential energy barriers are calculated, although a direct comparison with the experimentally-derived free energy barrier is not possible. However, a qualitative comparison can then still be done. *In this thesis we have employed both type of energies determining potential and free energy barriers and constructing the corresponding energy paths.*

Transition state theory

The Transition State Theory (TST) gives the framework of chemical reaction rate theory.

$$k = Ae^{\left(-\frac{\Delta G^{\ddagger}}{RT}\right)} \tag{3.36}$$

Within TST are encompassed different theories that have a common point: the assumption of the existence of a hypersurface (transition state) between two minima (usually reactants and products) in the phase space (energy surface). The TST assumes that 1) Reactants are in local equilibrium along the reaction coordinate, 2) the trajectories that cross this hypersurface do not recross, being thermalized in the products or reactants states and 3) the reaction coordinate degrees of freedom are separated from the rest and treated with classical mechanics. An important point of TST is that connects theory and experiments allowing the calculation of the reaction rate using the Eyring equation (Eq. 1.1) as we have commented. For a further review of the TST see for instance Garcia-Viloca and coworkers 2004[53] and Truhlar and coworkers 1996[54].

3.4.2 Stationary points and Energy minimization methods

Stationary points

Statistical mechanics (also called 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 each of these ensembles (x) is represented by the Boltzmann law (Eq. 3.33).

The low energy configurations are the most representative of a given system because the population of a given configuration decreases exponentially with the energy (such as reactants and products) although the high energy configurations play important roles such as connecting two minima (like the Transition state). Within a (bio)chemical reaction, the reactants and products (minima of the energy surface) are connected by a high energy surface (the transition state surface). The TS structure is at the same time a minimum of this surface and the highest energy structure along the reaction path.

All the minima and transition states correspond to stationary points of the energy surface, i.e., the first-derivative of the energy respect to the nuclear coordinates is zero. To distinguish between the stationary points frequency calculations over them have to be made. If the number of the resulting imaginary frequencies is zero, it is a minima, if is one, it is an ordinary TS (and a 1^{st} order saddle point)

and if present more it is an n^{th} order saddle point that depends on the number of imaginary frequencies (n).

However has to be taken into account that for systems with rugged potential energy landscapes, where entropic effects play role (as used to happen with multi-dimensional systems), the saddle point do not necessarily play a role of transition state.

Energy minimization methods

The search of stationary points is done by using numerically iterative algorithms. We can classify the optimization methods to find stationary points into three groups: no-derivatives, first-derivatives and second-derivatives respect to the way in that the the energy is derived over the atomic coordinates (we are going to explain only the first and second derivatives methods employed during this thesis). 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. The location of saddle points is also a challenging and demanding task. Oppositely to the location of minimum energy points whose energy values are reducing until find the minimum, to locate saddle points one has to find a point that is a maximum in one direction but a minimum in all the others, balancing both searches.

First-derivative methods

The two most widely used first-order minimization algorithms are the Steepest Descent and the Conjugate Gradient. These methods gradually change the coordinates of the atoms as they move towards the minimum point, since the force acting on each atom is equal to minus the gradient, $F = -\mathbf{g}_k$.

The steepest descent method takes a step along the direction of the force, which is the steepest direction at a given point of the energy surface. The direction of the gradient is determined by the largest interatomic forces, being orthogonal to the direction of the successive steps. 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 gradient method moves in a direction v_k from point x_k where v_k results from a combination of the gradient and the previous line search (directions vector) v_{k-1}:

$$\mathbf{v}_k = -\mathbf{g}_k + \gamma_k \mathbf{v}_{k-1} \tag{3.37}$$

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

Second-derivative methods

Second-order methods use not only the first derivatives, i.e. the gradients, but also the second derivatives to locate a minimum. The second derivation of the energy, i.e. the Hessian matrix, provides information about the curvature of the function. The Newton-Raphson (NR) method is the simplest second-order method. On the basis of the Taylor expansion of a function to second-order:

$$f(\mathbf{x}) \simeq f(\mathbf{x}_0) + \mathbf{g}^t(x - x_0) + \frac{1}{2}(x - x_0)^t\mathbf{H}(\mathbf{x} - \mathbf{x}_0) \tag{3.38}$$

being each step of NR expressed as:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}^{-1}\mathbf{g} \tag{3.39}$$

Moreover, the Hessian must be positive definite (all eigenvalues are positive) to ensure that the process minimizes the energy.

It requires the calculation of the inverse of the Hessian matrix, which is computationally demanding and problematic with near-zero eigenvalues. When the Hessian matrix is not positive, then the NR method moves to points (e.g. saddle points) where the energy increases. This method performs better near the minimum where the quadratic approximation is more valid, far from it becomes unstable.

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. They aim to eliminate the necessity of calculating the full matrix of second derivatives. These methods are known as Quasi-Newton methods and, among them, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) is a widely used one. Furthermore there are variants known as reduced or approximate second derivatives such as the Limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS). The difference between this one and the original method is that LBFGS avoid the storage of a dense **n** x **n** approximation to the inverse Hessian.

For a system with a large number of atoms, such as a protein, only first derivatives, like CG or a reduced set of approximate second derivatives, as the LBFGS are numerically affordable. *In this thesis we have employed mainly the Conjugate Gradient and LBFGS methods although the steepest descent has also been used.*

3.4.3 Determination of transition state structures and reaction pathways

As we have commented above, 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, and a reaction pathway is the path connecting two minima (reactants and products) passing through the transition state structure where the energy increases to a maximum and then falls.

There are several methods aimed to locate transition state structures and to elucidate reaction pathways, that are often closely related. However, we are going to restrict our discussion to the ones used in this thesis. We can classify them as 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). Within the second methods, another classification could be done according to the use or not of predefined collective variables or reaction coordinates to elucidate the reaction path.

Reaction paths

Any chemical reaction, complex formation or conformational change of a given system encompass a transition between two different basins, reactants and products, whose relative stability determines the thermodynamics of the process. This transition proceeds through the transition state structure, that is a high energy point along the reaction coordinate. The reaction coordinate(s), also called collective variable(s) (CVs), should be something that can be used to parameterize the reaction paths, usually geometrical parameters such as a dihedral angle or the difference between the bond forming and breaking. They have to describe the conformational change, the conversion between reactants and products, being the major problem that their choice is based on intuition and sometimes the intuition fails. In systems for which little is known about the reaction path, CVs are not used and a wide number of methods have been developed for finding its best description, encompassing between the so-called chain of states methods[55] (see section 3.4.3).

In their simple description, a chemical reaction takes place along the lowest potential energy, the Minimum Energy Path (MEP), which in mass-weighted coordinates is called the Intrinsic Reaction Coordinate (IRC). When the path refers to the lowest minimum free energy is the Minimum Free Energy Path (MFEP). Usually instead of second derivative calculations respect to the path, accurate approaches are made. These approaches leads to paths that formally are not true MEPs or MFEPs and thus has to be called (potential or free energy) reaction paths. We shall use the term 'reaction path' or 'pathway' to describe the path

between two minima. Has to be remarked that our use of the word 'reaction' does not necessarily imply a bond formation and/or breaking.

MFEP and MEP

A MFEP is defined on a free energy surface of Eq. 3.35 in the same way as a MEP is defined on a potential energy surface. It is the path between two minima on the surface such that the following condition holds:

$$[\mathbf{M}(\mathbf{z}).\nabla F(\mathbf{z})]^{\perp} = 0 \tag{3.40}$$

where $\nabla F(\mathbf{z})$ is the gradient of the free energy, $\mathbf{M} > \mathbf{z}$ is the metric tensor and $\perp$ indicates projection in the direction perpendicular to the path curve. Further details can be found for isntance in Maragliano *et al.*[51,56]. Remark that the metric tensor is not exclusive of the Free Energy, it also appears when calculating Minimum Potential Energy Paths in CV.

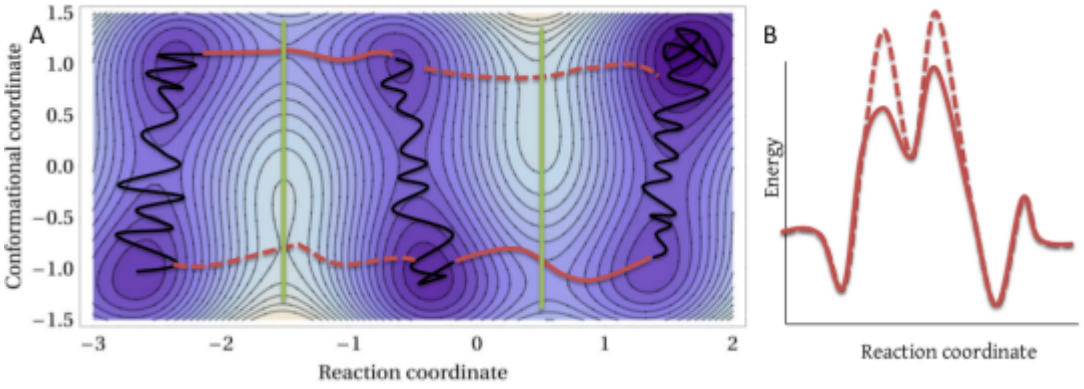


Figure 3.3: A) A model surface depicting the combination of a reaction coordinate for the chemical step and a conformational coordinate. The enzyme needs to adapt to a different conformation to optimally catalyse the reaction (solid red lines). If the sampling of the chemical transition (green lines) state is sufficient, we will capture the necessary conformational reorganization (black lines). (B) A one dimensional energy profile of the same model, where only the chemical step is considered. In such a case, the energy barrier seems to change with time, as the enzyme samples different conformations. Without considering this sampling the enzyme would have to surmount in one of the two steps, a higher energy barrier (dotted curve). Reproduced from Marcos *et al.*[57]

Local methods

Returning to the previously mentioned classification of techniques (aimed to locate TS structures and to determine reaction pathways), the first category is based on Newton-Raphson methods. 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. In general, to obtain a good starting structure to find the TS for simple reactions is enough by performing systematic constraint minimizations at different points (scanning) along the hypothetical reaction coordinate (the key is choose it properly). 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 towards the reactants and products. If the minimization process is brusque during the scanning, in some cases, the transition state can converge to a transition state connecting two other minima, not the desired ones. To check whether the TS actually connects the reactants and products, a widely used approach is going forward and backward from the saddle point until the obtained energy profile is unique[58,59]. This strategy is specially important for condensed phase systems, such as biomolecules, where multiple reaction paths may exist.

Chain of States

For the second category, several methods have been developed that make interpolations based on the two minima, such as the Zero Temperature String method[60,61] or the Nudged Elastic Band method (NEB)[62,63] developed by Jónsson and co-workers. The NEB 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:

$$\mathbf{F}_i = -\nabla V(\mathbf{R}_i)|\perp + \mathbf{F}_i^s|_{||} \tag{3.41}$$

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:

$$\mathbf{F}_i^s = k_{i+1}(\mathbf{R}_{i+1} - \mathbf{R}_i) + k_i(\mathbf{R}_{i-1} - \mathbf{R}_i) \tag{3.42}$$

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.

Regarding the reaction pathway, these kind of methods such as the string method or the NEB present drawbacks. Mainly, there are not able to produce minimum free energy paths, only minimum (potential) energy paths[51].

SoT

Other methods are based on CVs, as we have already commented, and thus are able to produces MFEPs but also present drawbacks. For instance, the CVs has to be precisely chosen because as its number increase these methods becomes computationally inefficient and unable to fully explore multidimensional energy surfaces. Unfortunately proteins usually needs a reasonable high number of CV to be completely described and also present a roughly energy landscape.

However, there are hybrid methods that incorporate the best of the two approaches. In this sense, the String method with collective variables[51] is able to produce MFEPs, and in a related development Roux and coworkers proposed a novel method that employed Swarms of Trajectories (SoT) to evolve the string and to estimate its average displacement in CV space[64]. The main advantage of the SoT is that it avoids the estimate of the potential of mean force and the metric tensor (see Eq. 3.35 and Eq. 3.40) simplifying the calculations and speeding up them.

The SoT method needs an initial reaction path to start. Let us assume that the initial path is composed by N images and it is defined by M CVs that are functions of x and that can distinguish distinct reacting configurations of the system. Eq. 3.35 can be used to determine the MFEP if a convenient representation of the path as a function of the variables, z, is available. In the SoT method this is done by parameterizing the path $z(\alpha)$ as $\alpha \in (0,1)$, where $\alpha = 0$ represents the reactants sate and $\alpha = 1$ the products sate. It is then assumed that the CVs evolve according to a non-inertial Brownian dynamics over some time step, according to:

$$z_i(\Delta t) = z_i(0) + \sum_j \Bigg(-\beta D_{ij}[\mathbf{z}(0)] \frac{\partial F}{\partial x_j}[\mathbf{z}(0)] + \partial_{zj} D_{ij}[\mathbf{z}(0)] \Bigg) \Delta t + R_i(0) \quad (3.43)$$

where D_{ij} is he diffusion tensor (that is equal to the metric tensor $\mathbf{M}(z)$ multiply by $k_B T$) and $R_i(0)$ is the gaussian thermal noise with mean of 0, and that is equivalent to the average drift[65] (or displacement) when the average of the thermal noise in Eq. 3.43 is averaged to zero. Once Eq. 3.43 has been defined, it can be employed to locate the most probable transition path (MPTP)[51,64] which is the path such that a system anywhere along it will have the highest probability

of remaining on it as it evolves. This is so because the most probable value of the Gaussian noise is zero.

Once the initial path is obtained, for each of the images of the path a trajectory is generated restrained around the collective variables (z). Thereafter unbiased trajectories are performed and the average displacement, $\overline{\Delta z^N}$, of the collective variables is calculated along them for each of the images of the path. After it the path is reparameterized according to the average displacement.

$$\begin{aligned}\overline{\Delta z_i(\Delta t)} &= \overline{z_i(\Delta t) - z_i(0)} \equiv z_i(\Delta t) = \\ &= z_i(0) + \sum_j \left(- \beta D_{ij}[\mathbf{z}(0)] \frac{\partial F}{\partial x_j}[\mathbf{z}(0)] + \partial_{zj} D_{ij}[\mathbf{z}(0)] \right) \Delta t \end{aligned} \tag{3.44}$$

All these steps are repeated until the convergence is reached.

An important insight of the SoT is the reparameterization [51,60,61,66] which consists in interpolating a curve through the path image structures and then redistributing them along the interpolated path. This is essential because it avoids the problem of the path images converge to regions of low free energy after repeated cicles applicating the Eq. 3.44. *In this thesis we have computed MEPs and MFEPs to study enzyme catalyzed reactions and protein damage events. Furthermore we have implemented the SoT method into the pDynamo library[50].*

3.4.4 Sampling Techniques

The above mentioned methods are able to energy minimize a molecular system. In fact they are conceived to yield the lowest-energy structure of a given basin. In reality, however, that is merely an approach to the state defined by this basin of the PES since temperature promotes fluctuations within the basin implying that there are a lot of similar structures that contributes to characterize this state. There exits a wide rage of techniques aimed to sampling the energy surface. Here, we are going to address the commonly employed MD simulation techniques as well as other methodologies (less) employed in this thesis such as the MC techniques.

3.4.4.1 Molecular Dynamics

One of the most used methodologies to explore the potential energy surface of complex systems with some degrees of freedom, is the molecular dynamics. This technique integrates the Newton's laws of motion, constructing trajectories that

allow to describe the temporal evolution of the positions and velocities of the particles of the system.

Additionally, MD calculations allow to calculate macroscopic properties of the system. These properties are calculated averaging the values obtained from a certain property along a trajectory long enough. The simulations has to be long enough to extract statistically significant information from time trajectories to predict relevant observable properties. Assuming the ergodic hypothesis, that postulates that 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, we could extract thermodynamic information of the system, i.e., the macroscopic properties of the system.

Integration of the equations of motion

The integration of the equations of motion can not be done analytically and for this reason, the use of algorithms based on finite differences, as the *Velocity Verlet algorithm*, is required.

These algorithms, divide the integration in some steps of time Δt and require the calculation of the forces that actuates on each particle of the system at a time t. These forces, allow the calculation of the acceleration and new velocities and positions at a time $t + \Delta t$, according to the second Newton law. Integration algorithms assume that the time-dependent positions can be expressed with a Taylor expansion:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \nu(t)\Delta(t) + \frac{1}{2}\mathbf{a}(t)\Delta t^2 \tag{3.45}$$

By adding the former equation to the corresponding expansion for the reverse time step, $r(t - \Delta t)$, the widely used *Verlet algorithm* is obtained:

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

where the acceleration is computed directly from the force at time t. Its main drawback is that the velocities are not included explicitly tending to loose numerical precision.

An important aspect of these algorithms is the size of the time step chosen, as one of the methodological limitations is the required size of the time steps. Due to this, the exploration is limited to dynamical processes that take place within the ps-ns time scale. If the time step is too large, it could produce high energy overlaps causing instabilities in the integration algorithm. On the other hand, excessively short time steps will not allow to cover biologically relevant (long enough) time scales and thus will not obtain pertinent chemical information

of the system. Therefore, a detailed balance between computational expense and stability in the numerical integration is needed. The time step usually adopted is within the fs time scale, which is around one order of magnitude shorter than the fastest molecular processes.

To overpass the time step limitation, some strategies have been proposed and a useful one is freezing the bonds. This procedure allow to increase the time step size without causing instabilities. To do this, several methods including constraints in the equations of motion have been developed, being the most widely used the SHAKE[67] and LINCS[68] algorithms.

Temperature and Pressure control

To get the macroscopic properties of the simulated systems the proper statistical mechanical (conformational) ensembles need to be calculated. By following the equations of motion the NVE ensemble can be described because of the potential and kinetic energy of the system will fluctuate and exchange, thus the total energy is conserved. 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 (the Berendsen thermostat and barostat are ones of the most used) to mimic these constant variables. By constant does not mean constrain the variable to a certain value during the whole simulation, but that along the simulation the variable only oscillates around an average value and does not drift.

MD system setup

To start a simulation experimental data, such as a crystallographic structure or an NMR ensemble, or a theoretical model is required. From them the atomic coordinates are extracted and the initial velocities are assigned. The assignment is usually based on the Maxwell-Boltzmann distribution at a given temperature.

Equilibration and production phase

Once the initial velocities are assigned the system is equilibrated until start the data collection phase. The equilibration is important to ensure that the kinetic energy (atomic velocities) is equally distributed among all degrees of freedom and oscillates around a mean value, i.e., the system is relaxed. After it, the production run starts.

The limiting, and most computationally demanding, part of each simulation step is the calculation of forces, that determines the accessible time scale. The current computational power allows MD simulations with molecular mechanics force fields to reach hundreds of nanoseconds, whereas those using a QM potential

energy function (usually a semi-empirical) have access to hundreds of picoseconds at most. Depending on the molecular properties of interest this is enough or not.

MD limitations

The main limitation of molecular dynamics is the limited amount of conformational space that can be explored, that at the end is related to the timescale that is able to cover. MD simulations describe conformational fluctuations over a broad range of time scales, i.e. from picoseconds to hundreds of nanoseconds. This allow, for instance, for accurate sampling of local motions of amino acid side chains or subdomains that take place at fast time scales. However, large-amplitude conformational changes, such as substrate-binding or allosteric events, occur at slower time scales (micro-milliseconds) that are inaccessible by standard MD techniques (as we explained at the Introduction).

A plethora of sampling methods aimed to broaden the exploration of the conformational space has been emerged. Of increasing importance and utility are Coarse-grained models, which vastly reduce the number of degrees of freedom and interaction sites by replacing sets of atoms by beads, and Replica exchange methods that running independent simulations at different temperatures are able to improve the MD description of the energy landscape. *In this thesis we have performed several MD calculations within the Gromacs[69] and the pDynamo[50] programs, using the Berendsen thermostat and barostat and in some of them the LINCS algorithm.*

3.4.4.2 Coarse Grained methods

The Coarse-Grained (CG) models vastly reduce the number of degrees of freedom and interaction sites, by replacing sets of atoms by beads, due to the potential energy surface is smoothed out leading to reduced friction, allowing the use of larger time steps. Generally, the less number of beads, the less expensive the simulation is. Furthermore, most CG models only compute short-range interactions, typically cut-off at a distance around 1 nm. All these strategies are aimed to reduce the computational expense[70].

Combining accuracy and predictive power in a few parameters is a difficult task achieved through different strategies giving rise to a variety of models and parameterization recipes. A typical classification of the coarse-grained models for proteins is based on the level of coarse-graining, i.e, the number of beads[71–73].

The advantages of using CG models are obtained at the cost of a number of emerging problems in the parameterization. The elimination of internal degrees of freedom have to be compensated, because their effect must be taken into account, in an implicit manner. Their effect have to be accounted in the effective forces

acting over the explicit degrees of freedom of the system[72]. Depending on the class of the CG model the potential energy is described accordingly, by the force field, to account the particularities of each of them[71,73].

Class	Number of beads	Type of bead
I	1 bead	C_α [74–77]
II	1 bead	C_β [78]
III	2 beads	C_α , Side chain (CM, C_β or centroid) [79–82]
IV	2 beads	Backbone CM , Side chain CM [83]
V	1-3 beads	C_α , 0-2 beads for side chain [84]
VI	1-6 beads	Backbone centroid , 0-5 beads for the side chain [85,86]
VII	1-2 beads	C_α , Backbone centroid , Side chain centroid [87–89]

Table 3.1: Classification of the CG (minimalist) models for proteins according to the number of beads. CM, center of mass. Adapted from Tozzini 2010[73]

The CG models apart of differing by the level of coarse graining are classified according to the philosophy of the force fields. There are two main categories of approaches, bottom up and top-down. In bottom-up approaches (also called structure-based coarse graining), effective CG interactions are extracted from reference atomistic simulations in a systematic way by using inverse Monte Carlo (IMC)[90], iterative Boltzmann inversion (IBI)[91], force matching (FM)[92,93], or related methods. Top-down approaches (also known as thermodynamic-based coarse graining), are based on match experimental data, especially thermodynamic properties. Typically, simple analytical interaction potentials are used and the parameters are optimized in an iterative procedure. As each approach has its own beneficial properties (bottom-up used to capture more fine details of the interaction and top-down are most transferable), many CG force fields rely on a combination of these two routes[70] (see the works of Tozzini[71–73] for a review on the different types of protein CG models, Brini *et al.* 2009[94] for a review on CG parameterization philosophies and Ingólfsson et al. 2014[70] for a recent review on the biomolecular applications of CG models).

Depending on the system to simulate one or another approach is better suited due to the different CG schemes are designed to study different properties. Some CG models are used to reproduce the solvent and other the system atoms, i.e. the protein. *In this thesis we have used the Profasi[95] and the Campari[96,97] force fields that represent the solvent with CG schemes.*

3.4.4.3 Monte Carlo techniques

Profasi and Campari are able to run MC simulations. The MC techniques are another methods widely used to explore the protein energy landscape. In MD simulations the system is simulated along the time, integrating the Newton's equation of motion, observing how the conformational ensembles goes changing whereas MC constructs canonical ensembles generated randomly and accepted or rejected according to a certain criteria. There are several MC applications but thereafter we refer MC to the methods referred to compute equilibrium properties of classical many-body systems.

MC methods are a extended class of computational algorithms oriented to obtain the probability distribution of some characteristic of the system. These methods usually rely on repeated random sampling, i.e, simulations are run many times, although this is not a must because performing one run can be enough (it depends on the simulating system). MC methods simulate accurately equilibrium thermodynamic and physical properties of a system of interest, as they are designed by construction to do so. MC simulations can be conducted in several different statistical mechanical ensembles, depending the sampled distributions on the ensemble employed. However for proteins as happens with other methods usually the NVT (canonical) ensemble is employed.

MC methods although unable to provide kinetic information are a good choice because they easily treat different thermodynamic ensembles performing a constant temperature simulation in contrast to the often required thermostat techniques in MD simulations. Furthermore they only need energies to generate the atomic trajectories without requiring expensive force calculation. Besides, these methods do not suffer of inaccuracies due to discrete-time approximations of the equations of motion.

A Monte Carlo simulation generates configurations by making random changes to the positions of the atoms over an initial system (as well as their orientations and conformations when necessary) from that the potential energy of each configuration of the system, together with the values of other properties, could be calculated by deterministic computations[98].

MC step

A MC algorithm constitutes a markov process in which a random walk is constructed in such a way that the probability of visit a particular configuration of the system is proportional to the Boltzmann factor (Eq. 3.33). Depending on the change in the probabilistic value of the energy function ($\Delta U = U_2 - U_1$) the step is accepted or rejected based, commonly, in the Metropolis criterion.

- If $\Delta U \leq 0 \rightarrow$ is accepted with a probability, $P(accept) = 1$

- If $\Delta U > 0 \rightarrow$ compute a uniform random number $Ranf$ within the interval $[0, 1)$. If the probability $P(accept) = e^{-\frac{\Delta V}{k_B T}} > Ranf$, is accepted.

This process constitute the called MC step. There are different MC methods but all of them follow the mentioned pattern.

In fact, a MC algorithm is composed by a group of MC steps that generate a Markov chain of states and has no history dependence. That means that the movement depends only on the actual and the previous step, not in the more previous (if exist). Using the definition of MC steps, i.e. the randomly generation of N MC points in the configurational space, the average value of a certain observable can be calculated as:

$$\langle A \rangle \approx \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} A\big(\overrightarrow{r_i}\big) \tag{3.47}$$

MC move

Once the MC step is accepted the system moves (this movement constitutes a MC move) adopting a new atomic configuration. As in MC methods the dynamic principles by which the atomic positions evolve incorporate random moves over the initial system, the dynamics of MC trajectories are not representative of the true system dynamics. On the other hand it depends on the type of the random moves performed. There are several standard Monte Carlo moves that one can use to explore conformational degrees of freedom. One simple example is the *single-particle displacement* that perturbed the position of the atoms within the maximum displacement range. This displacement is computed as a free parameter that can be turned to adjust the efficiency of the moves.

Regarding biomolecular simulations, specially chain molecules such as proteins, a combination of different kind of moves is employed. Commonly a torsional space sampling is used, that sometimes is augmented by sampling of angular degrees of freedom [99] or even the Cartesian coordinates directly [100]. The inclusion of such moves is determined by the force field used for the calculations [97]. For instance in the PROFASI code [95], first a pivot-type rotation about individual backbone bonds is performed, followed by a semi-local backbone update, employing the Biased Gaussian Steps (BGS) method, which rotates up to eight consecutive torsion angles simultaneously [101] and finally a rotation of individual side-chain angles is implemented [102].

3.4.4.4 Replica Exchange / Parallel Tempering

MD and MC simulations are usually carried out at a given temperature, starting from a representative initial configuration that ends reaching the thermal equilibrium. However the system of interest can have two or more potential wells separated by relatively high barriers. This is the case of globular proteins in solution that need a macroscopic time to fold into an specific configuration.

To study IDPs (see section 1.6.1), that are multiconfigurational and not well-folded proteins, Replica Exchange (RE) methods are more suitable than, for instance, MD due to the replicas performed on this kind of simulations prevent the system to fall into an energy minimum from which can not scape, resulting into a bad definition of the conformational landscape; a potential problem due to the extended nature of IDPs. Replica Exchange methods, also called Parallel-Tempering, try to overcome the multiple-minima problem by exchanging the temperature of non-interacting replicas of the system running at several temperatures. Both Monte Carlo, REMC, and Molecular Dynamics, REMD, variants are possible.

Formally the RE methods, applied to biomolecular systems, simulate M replicas of an original system of interest, each at a different temperature. By including an exchange mechanism between different temperatures, a total ensemble is generated encompassing the full range of temperatures. RE methods are ideally suited for run in parallel because each replica runs on a separate processor and there is only communication between processors when exchanges are attempted. The method is not restricted to a range of temperatures, but could involve a range of Hamiltonians, representing different parameters of the system.

In REMC by allowing configuration exchange between different (typically adjacent in temperature) replicas, the systems simulated at lower-temperature can access a representative set of low-energy regions of the energy landscape without be trapped into a local energy minima. *In this thesis we have performed several REMC simulations to study IDPs using the Profasi and Campari force fields.*

3.4.5 Data analysis

The biomolecular simulations can quickly generate very large amounts of complex data. As a consequence of the more available computational power, larger biomolecular simulations could be performed, i.e. researchers can tackle larger systems and simulate for longer time scales, producing more data[103]. These simulations usually generate (time-dependent or not) trajectories and it is not obvious from their direct visualization the relevant properties that one can (or should) extract. However, there are several data analysis techniques able to to overpass this

issue. Each of them use to be designed to an specific purpose (Here we are going to summarize some of these techniques used in this thesis).

3.4.5.1 Principal Components Analysis (PCA)

Principal component analysis is one of the most important techniques to the study of multivariate data. Although one of the earliest multivariate techniques, it continues being one of the most used. It is extremely versatile with applications in many disciplines[104,105]. PCA converts a set of observations (probably correlated variables of the system) into principal components (a set of values of linearly uncorrelated variables), which number is not higher than the number of original variables. The dimension reduction, i.e. the reduced set of new variables, is achieved through an orthogonal and linear combination of the original variables[104]. Focusing on protein dynamics simulations, this can provide a simplified description of the correlations between parameters that could be related somehow to different states or conformations of the studied system.

For instance analyzing QM/MM enzymatic calculations, where usually multiple trajectories (to achieve statistical significance) are generated, it allows the exploration of the diverse conformations of the system (such as reactants, TS and products). By using a set of geometrical parameters, relevant to the enzymatic process, could be found correlations between them related to the variance of the conformations.

Lets us consider the PCA method in more detail. Mathematically PCA is defined as a orthogonal linear transformation assuming all basis vectors are an orthonormal matrix[104]. PCA is oriented to extract the correlations and variances of a dataset finding the eigenvectors and eigenvalues. Thus, the PCA is computed by determining the eigenvectors and eigenvalues of the covariance matrix (that is used to measure how much the dimensions vary from the mean with respect to each other) built as:

$$\mathbf{Q}_{ij} = \langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle \tag{3.48}$$

where r_k is the k component of vector $\mathbf{r} = r_1 ... r_{3N}$ which defines the coordinates of a system of N atoms. $\mathbf{Q}$ is a symmetric N x N matrix, whose diagonal elements represent the variables variations and the off-diagonal elements the correlation between the variables. The eigenvectors of $\mathbf{Q}$ are N-dimensional vectors that indicate the principal components or essential modes and the corresponding eigenvalues the variations of the mode[104,106].

PCA is often used as the first step, reducing dimensionality, before undertaking

another multivariate technique such as partial least square regression(PLSR) or cluster analysis.

3.4.5.2 Partial Least Square Regression (PLSR)

PLSR[107] is used to find the fundamental relations between two matrices (X and Y). Is a well-known multivariate linear regression (chemometrics) method that does not suffer from linear dependencies among measured parameters and avoids noise overfitting problems when the number of correlated variables or parameters (X values) is high. PLSR produces new latent variables (variables that are not directly observed but are rather inferred) that optimally predict changes in the independent variables (Y values) from the observed variance in the dependent variables (X values).

PLSR regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among X values. Thus is a good option after PCA being the X matrix composed by the principal components. Furthermore, since PLSR is based on a bilinear model (X and Y data are projected to new spaces), it is a reliable and robust method for the validation and interpretation of statistical data.

3.4.5.3 Deviation techniques

Another simple but very useful methods employed to analyze biomolecular simulations are based on position measurements. These methods compare differences between two data sets, in our case, the variability of the atomic positions distribution functions.

Root Mean Square Displacement (RMSD)

Root mean square deviation (RMSD) is a measure of how much the protein structure changes along the simulated trajectory, i.e, along the time. It measures the average distance between the atoms (usually the backbone atoms) of the conformational ensembles, for instance comparing the structure of a partially folded protein and the native state.

Furthermore this measure could be used as a control of the simulation, because the RMSD could be measured along the simulated trajectory; if the RMSD is still changing on average at the end of your simulation, probably is due to it is not long enough and is not equilibrated. If the equilibrium is reached the RMSD values should stabilize around a fixed value. The RMSD for the conformation corresponding to the frame x is computes as:

$$RMSD_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i'(t(x)) - r_i(t(ref)))^2} \tag{3.49}$$

where N is the number of atoms in the atom selection, t_{ref} is the reference time, (typically the first frame $t = 0$ and ri is the position of the selected atom in the frame x recorded at time $t(x)$.

3.4.5.4 Reweighting techniques

As we have explained at the Introduction (see section 1.4) the ensembles have to be representative of the simulated system. Despite the development of more accurate force fields, there are still continuing inconsistencies[108]. A good option to overpass the problem is to reweight the simulated ensembles. Reweighting techniques allows to 'expand' the results from the original simulation to fit the 'correct' (generated experimentally or theoretically) ensembles.

Maximum Entropy Principle

A recently proposed way to reweight conformational ensembles is using the maximum entropy (MaxEnt) method[109,110] that is a logically consistent way to fit data to previously known models introducing the minimum possible modifications. Originally the method was introduced by Jaynes in 1957[111] and derives from minimizing the information included in an ensemble to fit certain observables.

Assuming that we have a set of simulated (by MD or MC) N structures $X_{j=1,N}$, for a set of M observables $q = q_{i=1,M}$, according to Pitera and Chodera[109], the application of the MaxEnt principle resulted in a reweighting of the probability of each structure j by a term w^j defined as:

$$w^j = \sum_{i}^{M} exp(\lambda_i q_i^j) \tag{3.50}$$

The form of the reweighting is kept fix, applying a single parameter λ_i to each observable. w^j modifies the weight of the structure to fit the objective (usually experimental) observables. q_i^j represents the value of the observable i in the structure X_j and $\mathbf{q}$ is a M x N matrix. λ_i is a lagrange multiplier that represent the constrained experimental data:

$$c_1 = \int d(\overline{x}q_i(x)p(x) - q_i^{exp} \tag{3.51}$$

where $\mathbf{q}^{exp}$ is the vector that contains the experimental measurements, being the average value of observable q_i for a given reweighting defined as:

$$\langle q_i \rangle = \sum_j^N w^j q_i^j \tag{3.52}$$

In this thesis we have employed all the above mentioned data analysis methods. PCA, PLSR and RMSD to analyze QM/MM an MD simulations and our own implemented MaxEnt to fit RDCs ensembles of IDPs.

3.5 Charge Transfer methods

Charge transfer (CT) reactions play a very important role in a wide range of chemical and biological processes (see section 1.3.2). In most of them, charge transfer occurs between two chemical groups, donor and acceptor, which usually are separated by several angstroms. One useful way to characterize this transfer is by calculating the electronic coupling V_{DA} between them. The strength of this coupling determines whether the process is adiabatic or non-adiabatic. It determines if the system has a certain probability to jump from the initial to the final potential energy curves (non-adiabatic and a weak electronic coupling) or remains on the lower potential energy curve (adiabatic and reasonable high electronic coupling)(see section 1.3.2.1).

3.5.1 Electronic Couplings

Several procedures have proved useful for calculating electronic coupling matrix elements[112]. For the system where donor and acceptor are separated by a bridge, effective coupling can be estimated using Larsson's formula.

$$V_{DA} = V_{D1} V_{nA} \sum_{i=1}^{N} \frac{C_{1i} C_{ni}}{E - \epsilon_i} \tag{3.53}$$

where V_{D1}, V_{nA} are the matrix elements between the bridge and the donor or the acceptor, respectively, and E is the tunnelling energy. The summation extends over all states of the bridge.

From this first attempt other approaches, as Newton and Cave[113], describes the Generalized Mülliken-Hush:

$$V_{DA} = \frac{|(E_2 - E_1)|\mu_{12}}{\sqrt{(\mu 1 - \mu_2)^2 + 4q_{12}^2}} \tag{3.54}$$

The adiabatic states are transformed to the diabatic states using the matrix that diagonalizes the adiabatic dipole moment matrix. For a two state model, the electronic coupling (the off-diagonal matrix element of the non-adiabatic Hamiltonian) can be expressed via the vertical excitation energy $E_2 - E_1$, the difference $\mu_1 - \mu_2$ of the adiabatic dipole moments and the transition dipole moment μ_{12} [112].

From this and using the Fragment Charge Differentiation (FCD):

$$V_{DA} = \frac{|(E_2 - E_1)|\Delta q_{12}}{\sqrt{(\Delta q_1 - \Delta q_2)^2 + 4q_{12}^2}} \tag{3.55}$$

being $\Delta q_1, \Delta q_2$ are the Donor-Acceptor differences in the adiabatic states ψ_1, ψ_2, respectively for the two-state model, and Δq_{12} are the corresponding off-diagonal terms. In the same spirit that Eq. 3.55 we can derive the SFCD (or simplified FCD):

$$V_{DA} = \frac{1}{2}(E_2 - E_1)\sqrt{1 - \Delta q^2} \tag{3.56}$$

where Δq is the difference of the charges on donor and acceptor in the ground state. When donor and acceptor are 'in resonance', $E_D = E_A$, then Eq. 3.56 - 3.54 are reduced to the minimum splitting expression:

$$V_{DA} = \frac{1}{2}(E_2 - E_1) \tag{3.57}$$

The resonance condition imply that $\mu_1 = \mu_2$ in Eq. 3.54, $\Delta q_1 = \Delta q_2 = 0$ in Eq. 3.55 and $\Delta q = 0$ in Eq. 3.56.

The adiabatic splitting can be calculated as the first excitation energy of the radical (cation or anion) $\Delta = E_2 - E_1$ using a configuration interaction (CI) method. Alternatively the Koopman's theorem (which states that in closed-shell HF, the first ionization energy of a given molecular system is equal to the negative of the orbital energy of the highest occupied molecular orbital (HOMO)) can be employed. Doing that Δ can be estimated as the difference of the one-electron energies of the two highest occupied molecular orbitals HOMO and HOMO-1 calculated for the closed-shell neutral dimer. This constitutes the one-electron approximation.

Following the one-electron approximation, the donor and acceptor charges of the first adiabatic state of the neutral dimer can be estimated via the corresponding Mulliken populations of the HOMO of the neutral system. Then the charge on a fragment can be estimated as [112]:

$$q_1(F) = \sum_{i \in F} C_{i,HOMO} \sum_{j=1}^{M} C_{j,HOMO} S_{ij} \tag{3.58}$$

where S_{ij} is the overlap of atomic orbitals i and j; i runs over atomic orbitals associated with the selected fragment F while j runs over all Atomic Orbitals (AOs). The fragment charges of the second adiabatic state are calculated analogously using the coefficients $C_{i,HOMO-1}$ of the molecular orbital HOMO-1 in place of $C_{i,HOMO}$ [112]. In this approximation, the general quantity $q_{mn}(F)$ can be defined by:

$$q_{mn}(F) = \frac{1}{2} * \\ * \left[\sum_{i \in F} C_{i,HOMO+1-m} \sum_{j=1}^{M} C_{j,HOMO+1-n} S_{ij} + \sum_{i \in F} C_{i,HOMO+1-n} \sum_{j=1}^{M} C_{j,HOMO+1-m} S_{ij} \right] \tag{3.59}$$

and having into account that:

$$k_{ET} \propto |V_{DA}|^2 e^{\frac{-\Delta E}{RT}} \tag{3.60}$$

as can be seen at section 1.3.2, we arrive to the formula developed by Rösch and Voityuk [112], the formula that used to calculate the electronic couplings:

$$V_{DA} = \frac{\Delta E_{12} |\mu_{12}|}{|\mu_D - \mu_A|} \tag{3.61}$$

where,

$$\mu_1 - \mu_2 = \sum_{i,j=1}^{M} (C_{i,HOMO} C_{j,HOMO} - C_{i,HOMO-1} C_{j,HOMO-1}) d_{i,j} \tag{3.62}$$

and

$$\mu_{12} = \sum_{i,j=1}^{M} C_{i,HOMO} C_{j,HOMO-1} d_{i,j} \tag{3.63}$$

(In this thesis we have employed the FCD method to analyze CT processes in an ezymatic damage reaction.)

Bibliography

[1] C. J. Cramer. *Essentials of Computational Chesmistry:Theories and models*. Sussex:Wiley,2002

[2] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. New York, Mc Graw-Hill, 1989.

[3] F. Jensen. *Introduction to Computational Chemistry*. John Wiley & Sons, 1999.

[4] J. Andres and J. Bertran, editors. *Theoretical and Computational Chemistry: Foundations, Methods and Techniques*. Publicacions de la Universitat Jaume I, 2007.

[5] A. R. Leach. *Molecular Modeling: Principles and Applications*. Essex: Addisson Wesley Longman, 1996.

[6] D. Frenkel and B. Smit. *Understanding Molecular Simulation. From Algorithms to Applications*. Academic Press, 2002.

[7] H. J. C. Berendsen. *Simulating the physical world. Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*. Cambridge University Press, 2007.

[8] D. M. Zuckerman *Statistical Physics of Biomolecules. An Introduction*. CRC Press, 2010.

*Bibliography references the books on the whole section is based, and References indicates the specific citations to articles or book chapters. The citation numbers within the main the text refers to the References not the Bibliography.

References

[1] C. C. J. Roothaan. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.*, 23:69–89, Apr 1951.

[2] G. Hall. The Molecular Orbital Theory of Chemical Valency. VIII. A Method of Calculating Ionization Potential. *Proc. Roy. Soc.*, A205:541–552, Apr 1951.

[3] S. Grimme. Improved second-order Møller-Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies. *J. Chem. Phys.*, 118(20):9095, 2003.

[4] S. Grimme, L. Goerigk, and R. F. Fink. Spin-component-scaled electron correlation methods. *Wiley Interdiscip. Rev. Comput. Mol. Science.*, 2(6):886–906, 2012.

[5] J. A. Pople, D. P. Santry, and G. A. Segal. Approximate self−consistent molecular orbital theory. I. invariant procedures. *J. Chem. Phys.*, 43(10):S129–S135, 1965.

[6] J. A. Pople and G. A. Segal. Approximate self−consistent molecular orbital theory. II. calculations with complete neglect of differential overlap. *J. Chem. Phys.*, 43(10):S136–S151, 1965.

[7] J. A. Pople, D. I. Beveridge, and P. Debosh. Approximate self-consistent molecular-orbital theory.5.intermediate neglect of differential overlap. *J. Cheical Phys.*, 47:2026–2023, 1997.

[8] M. J. S. Dewar and W. Thiel. Ground states of molecules. 38. The MNDO method. Approximations and parameters. *J. Am. Chem. Soc.*, 99(15):4899–4907, 1977.

[9] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.*, 107(13):3902–3909, 1985.

[10] J. J. P. Stewart. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.*, 10(2):209–220, 1989.

[11] G. B. Rocha, R. O. Freire, A. M. Simas, and J. J. P. Stewart. RM1: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.*, 27(10):1101–1111, 2006.

[12] W. Thiel and A. Voityuk. Extension of the MNDO formalism tod orbitals: Integral approximations and preliminary numerical result. *Theor. chimica acta*, 81(6):391–404, 1992.

[13] W. Thiel and A. Voityuk. Extension of the MNDO formalism to d orbitals: Integral approximations and preliminary numerical results. *Theor. chimica acta*, 93(5):315–315, 1996.

[14] W. Thiel and A. A. Voityuk. Extension of MNDO to d orbitals: Parameters and Results for the Second-Row Elements and for the Zinc Group. *J. Phys. Chem.*, 100(2):616–626, 1996.

[15] A. A. Voityuk and N. Rösch. AM1/d Parameters for Molybdenum. *J. Phys. Chem. A*, 104(17):4089–4094, 2000.

[16] X. Lopez and D. M. York. Parameterization of semiempirical methods to treat nucleophilic attacks to biological phosphates: AM1/d parameters for phosphorus. *Theor. Chem. Accounts*, 109(3):149–159, 2003.

[17] P. Winget, A. H. Horn, C. Selçuki, B. Martin, and T. Clark. AM1* parameters for phosphorus, sulfur and chlorine. *J. Mol. Model.*, 6(6):408–414, 2003.

[18] K. Nam, J. Gao, and D. M. York. Electrostatic interactions in the hairpin ribozyme account for the majority of the rate acceleration without chemical participation by nucleobases. *RNA (New York, N.Y.)*, 14(8):1501–1507, 2008.

[19] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas . *Phys. Rev.*, 136:B864–B871, Nov 1964.

[20] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[21] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.*, 58(8):1200–1211, 1980.

[22] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.*, 98(45):11623–11627, 1994.

[23] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098–3100, 1988.

[24] C. Lee, W. Yang, and R. G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789, Jan 1988.

[25] A. D. Becke. A new mixing of Hartree-Fock and local density-functional theories. *J. Chem. Phys.*, 98:1372–1377, 1993.

[26] C. Adamo and V. Barone. Exchange functionals with improved long-range behavior and adiabatic connection methods without adjustable parameters: The mPW and mPW1PW models. *J. Chem. Phys.*, 108(2):664–675, 1998.

[27] Y. He, J. Gräfenstein, E. Kraka, and D. Cremer. What correlation effects are covered by density functional theory? *Mol. Phys.*, 98(20):1639–1658, 2000.

[28] B. J. Lynch, P. L. Fast, M. Harris, and D. G. Truhlar. Adiabatic Connection for Kinetics. *J. Phys. Chem. A*, 104(21):4811–4815, 2000.

[29] B. L. Kormos and C. J. Cramer. Adiabatic connection method for X−+RX nucleophilic substitution reactions ($X = F, CL$). *J. Phys. Org. Chem.*, 15(10):712–720, 2002.

[30] Y. Zhao and D. G. Truhlar. Density Functionals with Broad Applicability in Chemistry. *Accounts Chem. Res.*, 41(2):157–167, 2008.

[31] S. Grimme. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.*, 27(15):1787–1799, 2006.

[32] S. Grimme. Density functional theory with London dispersion corrections. *Wiley Interdiscip. Rev. Comput. Mol. Science.*, 1(2):211–228, 2011.

[33] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.*, 7(2):230–252, 1986.

[34] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4(2):187–217, 1983.

[35] W. R. P. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger, and W. F. van Gunsteren. The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A*, 103(19):3596–3607, 1999.

[36] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.

[37] P. Cieplak, F.-Y. Dupradeau, Y. Duan, and J. Wang. Polarization effects in molecular mechanical force fields. *J. Physics: Condens. Matter*, 21(33):333102, 2009.

[38] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, 25(13):1656–1676, 2004.

[39] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct. Funct. Bioinforma.*, 78(8):1950–1958, 2010.

[40] H. Berendsen, J. Postma, W. van Gunsteren, and J. Hermans. Interaction Models for Water in Relation to Protein Hydration. In B. Pullman, editor, *Intermolecular Forces*, volume 14 of *The Jerusalem Symposia on Quantum Chemistry and Biochemistry*, pages 331–342. Springer, 1981.

[41] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[42] M. W. Mahoney and W. L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.*, 112(20):8910–8922, 2000.

[43] D. M. York, T. A. Darden, and L. G. Pedersen. 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(10):8345–8348, 1993.

[44] K. Nam, J. Gao, and D. M. York. An Efficient Linear-Scaling Ewald Method for Long-Range Electrostatic Interactions in Combined QM/MM Calculations. *J. Chem. Theory Comput.*, 1(1):2–13, 2005.

[45] P. Sherwood. Hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) approaches. *Mod. methods algorithms quantum chemistry*, 1:257–277, 2000.

[46] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103(2):227 – 249, 1976.

[47] M. J. Field. *A Practical Introduction to the Simulation of Molecular Systems.* Cambridge University Press, 2007.

[48] H. M. Senn and W. Thiel. QM/MM Methods for Biological Systems. In *Atomistic Approaches in Modern Biology*, volume 268, pages 173–290. Springer, 2007.

[49] H. M. Senn and W. Thiel. QM/MM Methods for Biomolecular Systems. *Angewandte Chemie*, 48(7):1198–1229, January 2009.

[50] M. J. Field. The pDynamo program for molecular simulations using hybrid quantum chemical and molecular mechanical potentials. *J. Chem. Theory Comput.*, 4(7):1151–1161, 2008.

[51] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.*, 125(2):24106, 2006.

[52] D. J. Wales and T. V. Bogdan. Potential Energy and Free Energy Landscapes. *J. Phys. Chem. B*, 110(42):20765–20776, 2006.

[53] M. Garcia-Viloca, J. Gao, M. Karplus, and D. G. Truhlar. How enzymes work: analysis by modern rate theory and computer simulations. *Science.*, 303(5655):186–195, 2004.

[54] D. G. Truhlar, B. C. Garrett, and S. J. Klippenstein. Current Status of Transition-State Theory. *J. Phys. Chem.*, 100(31):12771–12800, 1996.

[55] E. Weinan and E. Vanden-Eijnden. Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.*, 61:391–420, 2010.

[56] L. Maragliano, B. Roux, and E. Vanden-Eijnden. A comparison between mean forces and swarms-of-trajectories string methods. *J. Chem. Theory Comput.*, 10(2):524–533, 2014.

[57] E. Marcos, M. Sanchez-Martinez, and R. Crehuet. Interplay between enzyme function and protein dynamics. A multi-scale approach to the study of the NAG kinase family and two class II aldolases. In *Computational Approaches to Protein Dynamics: From Quantum to Coarse-Grained Methods*. Academic Press, In press.

[58] C. Gonzalez and H. B. Schlegel. Improved algorithms for reaction path following: HigherâĂŘorder implicit algorithms. *J. Chem. Phys.*, 95(8):5853–5860, 1991.

[59] Y. S. Lee, S. E. Worthington, M. Krauss, and B. R. Brooks. Reaction Mechanism of Chorismate Mutase Studied by the Combined Potentials of Quantum Mechanics and Molecular Mechanics. *J. Phys. Chem. B*, 106(46):12059–12065, 2002.

[60] W. E, W. Ren, and E. Vanden-Eijnden. String method for the study of rare events. *Phys. Rev. B*, 66(5):052301, 2002.

[61] W. E, W. Ren, and E. Vanden-Eijnden. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.*, 126(16):164103, 2007.

[62] H. Jonsson, G. Mills, and K. W. Jacobsen. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*. World Scientific, 385–404.

[63] G. Henkelman and H. Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.*, 113(22):9978–9985, 2000.

[64] A. C. Pan, D. Sezer, and B. Roux. Finding transition pathways using the string method with swarms of trajectories. *J. Phys. Chem. B*, 112(11):3432–40, 2008.

[65] G. Hummer and I. G. Kevrekidis. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *J. Chem. Phys.*, 118(23):10762–10773, 2003.

[66] W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E. Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide. *J. Chem. Phys.*, 123(13):–, 2005.

[67] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23(3):327 – 341, 1977.

[68] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[69] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.*, 4(3):435–447, 2008.

[70] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole, and S. J. Marrink. The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Science.*, 4(3):225–248, 2014.

[71] V. Tozzini. Coarse-grained Models for Proteins. *Curr. Opin. Struct. Biol.*, 15(2):144–150, 2005.

[72] J. A. McCammon and V. Tozzini. One-bead coarse-grained models for proteins. In G. A. Voth, editor, *Coarse-Graining of Condensed Phase and Biomolecular Systems*, pages 285–298. CRC Press, 2008.

[73] V. Tozzini. Minimalist models for proteins: a comparative analysis. *Q. Rev. Biophys.*, 43:333–371, 8 2010.

[74] J. D. Honeycutt and D. Thirumalai. Metastability of the folded states of globular proteins. *Proc. Natl. Acad. Sci.*, 87(9):3526–3529, 1990.

[75] P. Das, S. Matysiak, and C. Clementi. Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes. *Proc. Natl. Acad. Sci.*, 102(29):10141–10146, 2005.

[76] V. Tozzini and J. A. McCammon. A coarse grained model for the dynamics of flap opening in HIV-1 protease. *Chem. Phys. Lett.*, 413:123 – 128, 2005.

[77] A. Korkut and W. A. Hendrickson. A force field for virtual atom molecular mechanics of proteins. *Proc. Natl. Acad. Sci.*, 2009.

[78] J. A. McCammon, S. H. Northrup, M. Karplus, and R. M. Levy. Helix-coil transitions in a simple polypeptide model. *Biopolym.*, 19(11):2033–2045, 1980.

[79] I. Bahar and R. Jernigan. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.*, 266(1):195 – 214, 1997.

[80] I. Bahar, M. Kaplan, and R. Jernigan. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. *Proteins: Struct. Funct. Bioinforma.*, 29(3):292–308, 1997.

[81] A. Mukherjee and B. Bagchi. Correlation between rate of folding, energy landscape, and topology in the folding of a model protein HP-36. *J. Chem. Phys.*, 118(10):4733–4747, 2003.

[82] P. Májek and R. Elber. A coarse-grained potential for fold recognition and molecular dynamics simulations of proteins. *Proteins: Struct. Funct. Bioinforma.*, 76(4):822–836, 2009.

[83] J. Zhou, I. F. Thorpe, S. Izvekov, and G. A. Voth. Structural equilibrium fluctuations in mesophilic and thermophilic α-amylase. *Biophys. J.*, 92(12):4289–4303, 2007.

[84] M. Zacharias. Protein−protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Science.*, 12(6):1271–1282, 2003.

[85] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.*, 4(5):819–834, 2008.

[86] T. Ha-Duong. Protein Backbone Dynamics Simulations Using Coarse-Grained Bonded Potentials and Simplified Hydrogen Bonds. *J. Chem. Theory Comput.*, 6(3):761–773, 2010.

[87] A. Liwo, S. Ołdziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.*, 18(7):849–873, 1997.

[88] A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, S. Ołdziej, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization. *J. Comput. Chem.*, 18(7):874–887, 1997.

[89] M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, 104(1):59 – 107, 1976.

[90] A. P. Lyubartsev and A. Laaksonen. Calculation of effective interaction potentials from radial distribution functions: A reverse monte carlo approach. *Phys. Rev. E*, 52:3730–3737, 1995.

[91] D. Reith, M. Pütz, and F. Müller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.*, 24(13):1624–1636, 2003.

[92] S. Izvekov and G. A. Voth. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B*, 109(7):2469–2473, 2005.

[93] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.*, 128(24):–, 2008.

[94] E. Brini, E. A. Algaer, P. Ganguly, C. Li, F. Rodriguez-Ropero, and N. F. A. van der Vegt. Systematic coarse-graining methods for soft matter simulations - a review. *Soft Matter*, 9:2108–2119, 2013.

[95] A. Irbäck and S. Mohanty. PROFASI : A Monte Carlo Simulation Package for Protein Folding and Aggregation. *J. Comput. Chem.*, 27:1548–1555, 2006.

[96] A. Vitalis and R. V. Pappu. Absinth: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.*, 30(5):673–699, 2009.

[97] A. Vitalis and R. V. Pappu. Methods for monte carlo simulations of biomacromolecules. In R. A. Wheeler, editor, *Annual Reports in Computational Chemistry*, volume 5, pages 49 – 76. Elsevier, 2009.

[98] D. J. Earl and M. W. Deem. Monte carlo simulations. In *Methods in Molecular Biology*, volume 443. Springer, 2008.

[99] J. P. Ulmschneider, M. B. Ulmschneider, and A. Di Nola. Monte carlo vs molecular dynamics for all-atom polypeptide folding simulations. *J. Phys. Chem. B*, 110(33):16733–16742, 2006.

[100] S. Cahill, M. Cahill, and K. Cahill. On the kinematics of protein folding. *J. Comput. Chem.*, 24(11):1364–1370, 2003.

[101] G. Favrin, A. Irbäck, and F. Sjunnesson. Monte carlo update for chain molecules biased gaussian steps in torsional space. *J. Chem. Phys.*, 114(18):8154–8158, 2001.

[102] A. Irbäck and S. Mohanty. All-atom monte carlo simulations of protein folding and aggregation. In *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes*, pages 433–444. Springer, 2014.

[103] J. C. Thibault, T. E. Cheatham, and J. C. Facelli. Ibiomes Lite: Summarizing Biomolecular Simulation Data in Limited Settings. *J. Chem. Inf. Model.*, 54(6):1810–1819, 2014.

[104] I. Jolliffe. *Principal Component Analysis*. Springer Verlag: Berlin, Germany, 2002.

[105] I. Jolliffe. Principal Component Analysis. In *Encyclopedia of Statistics in Behavioral Science*, pages 1580–1584. John Wiley & Sons, 2002.

[106] D. H. Jeong, C. Ziemkiewicz, W. Ribarsky, and R. Chang. Understanding Principal Component Analysis Using a Visual Analytics Tool. In *Mathematics: Fundamentals and Applications*. UKC 2009 conferences, 2009.

[107] P. Geladi and B. R. Kowalski. Partial least-squares regression: a tutorial . *Anal. Chimica Acta*, 185:1 – 17, 1986.

[108] K. A. Beauchamp, Y.-S. Lin, R. Das, and V. S. Pande. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J. Chem. Theory Comput.*, 8(4):1409–1414, 2012.

[109] J. W. Pitera and J. D. Chodera. On the Use of Experimental Observations to Bias Simulated Ensembles. *J. Chem. Theory Comput.*, 8(10):3445–3451, 2012.

[110] K. A. Beauchamp, V. S. Pande, and R. Das. Bayesian Energy Landscape Tilting: Towards Concordant Models of Molecular Ensembles. *Biophys. J.*, 106:1381–1390, 2014.

[111] E. T. Jaynes. Information Theory and Statistical Mechanics. *Phys. Rev.*, 106:620–630, 1957.

[112] A. A. Voityuk and N. Rösch. Fragment charge difference method for estimating donor-acceptor electronic coupling: Application to dna π-stacks. *J. Chem. Phys.*, 117(12), 2002.

[113] M. Newton and R. Cave. *Molecular Electronics*, chapter Molecular control of electron and hole transfer processes: Theory and applications, pages 73–118. Blackwell, 1997.

Chapter 4

Results

The results of this thesis are divided into two main sections, and several subsections, addressed to highlight the local and global protein movements, encompassing the different studies realized. These studies regarding local motions are related to enzyme catalysis and protein damage and within global movements to the reweighting of protein conformational ensembles and the cooperativity of secondary structure elements over them, specially focused on IDPs. Each subsection presents a brief summary of the study developed to address it, followed by the corresponding manuscript or the draft for *ASAP* publication.

4.1 Local Motions

4.1.1 Catalytic role of protein motions

To try to clarify whether the dynamics at the global level influence the local motions to catalyze the chemical step or not, we decided to use the NAGK enzyme. This enzyme has been widely studied by the group of Professor Rubio (IBV-CSIC) experimentally and theoretically by our group. It constitutes a good example because it presents a high number of accessible crystal structure representing different states of the chemical reaction. Rubio and co-workers showed that those corresponding to transition state analogues had shorter substrate distances than the crystal structures corresponding to reactants, inferring that the 'conformational compression' (O-O distance between substrates ATP and NAG) of the substrates favours catalysis. To investigate this hypothesis we complexed four different representative crystal structures, PDB accession numbers 1GS5, 1OH9, 1OHA, and 2X2W, with the natural substrates of the reaction ATP and NAG (see Fig. 4.1). Then we performed MD followed by QM/MM simulations over the different crystal structures and different trajectory snapshots for each of them,

respectively. Among all crystal structures considered, we found that two of them (1OHA and 1OH9) were particularly useful to address this hypothesis.

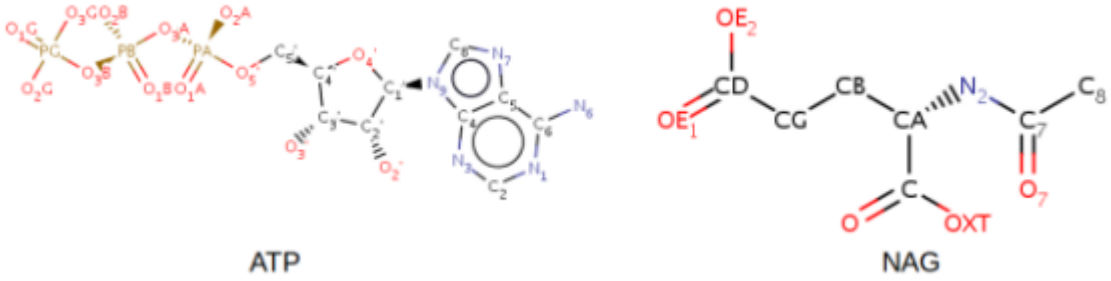


Figure 4.1: An schematic picture of the natural substrates of NAGK.

The variety of energy profiles obtained (see Fig. 4.2), even over the same crystal structure, indicate that the energy barrier is not determined by the change in the distance (between nucleophile and leaving group, i.e. O-O distance) along the reaction process, i.e. it does not depend on how much compression is needed from reactants to transition state. Instead, there is a noticeable correlation between the substrate distance in the reactants state and the energy barrier. The lower the O-O distance the lower the energy barrier. In this sense, the statistical analysis (PCA + PLSR) reveals that each reactant conformation proceeds through its own reaction valley with a transition state whose instability (represented by a high energy barrier) will increase as the reactants be afar.

Additionally, we found that the energy barrier is not only determined by the reactants compression distance, but also by its spatial distribution, i.e, the linear angle of the transferring phosphoryl with the nucleophile and leaving groups (O-P-O angle). The higher the O-P-O angle plus the shorter the O-O distance, the lower the energy barrier. Furthermore the role of water in the active site was also found extremely important. Overall, the structure of the pre-reactive complex contains relevant predictive information on the energy barrier.

The calculated energy barrier for the chemical step for all our conformations is significantly lower than the apparent (experimental) energy barrier ($\Delta E^{\ddagger} = 67$ kJ/mol). This energy barrier corresponds to a free energy which implicitly incorporates dynamical and tunnelling corrections. All the energy values are based on approaches: to estimate the average experimental energy barrier from thermodynamical data (that underestimates the dynamical and tunnelling corrections) and to perform MD and QM/MM simulations (ranging from the choice of the QM region to the limiting sampling of conformations) that may induce uncertain-

ties. Furthermore, at the end, our QM/MM calculations return potential energy instead of free energy.

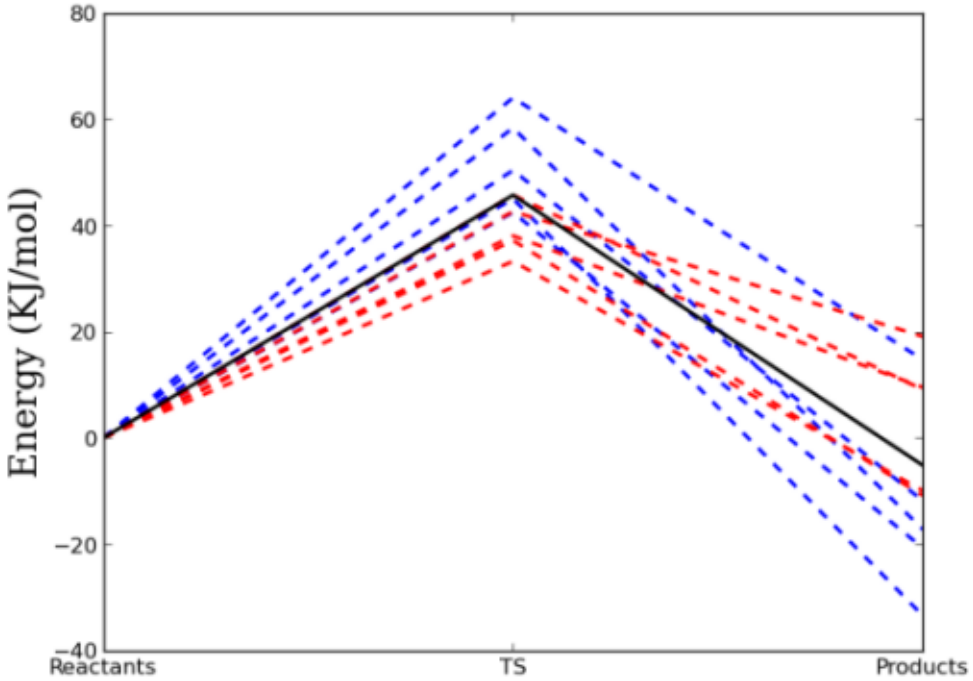


Figure 4.2: Reactants, transition state and product energies of the five 1OH9 (blue) and 1OHA (red) structures. The dispersion of energy values is large, even for snapshots coming from the same crystal structure. The black line represents the average energie values.

However, the experimental energy barrier almost doubles (constituting a too much big difference to be cancelled by methodological errors) the calculated average energy barrier for the chemical step (39 kJ/mol), suggesting that the chemical step is not the rate-limiting step for this enzyme and that conformational motions (associated with the lid opening and closing) can be slower than the chemical reaction. This hypothesis has also been put forward for other enzymes, such as adenylate kinase, cyclophilin A or dihydrofolate reductase.

Summarizing, our results indicate that the catalytic proficiency of the enzyme lies in collective motions accessing properly oriented and highly compressed active site conformations, thus supporting the 'conformational compression' hypothesis inferred by Rubio and coworkers. Besides the fact that the energy barrier depends too much on the reactants conformation, indicates that maybe the chemical step is not rate-limiting and, instead, the protein motions leading to catalytic compressed

conformations (to the extent that the resultant chemical barrier is close to the conformational barrier) are the limiting process.

THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Conformational Compression and Barrier Height Heterogeneity in the *N*-Acetylglutamate Kinase

Melchor Sanchez-Martinez,[†] Enrique Marcos,[†,‡] Romà Tauler,[∥] Martin Field,[§] and Ramon Crehuet*[,†]

[†]Institute of Advanced Chemistry of Catalonia (IQAC), CSIC, Jordi Girona 18-26, 08034, Barcelona, Spain

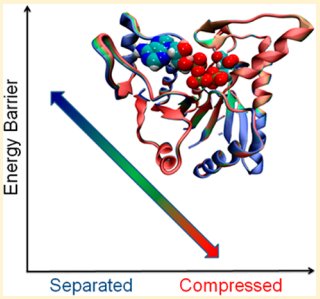
[‡]Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States

[§]Institut de Biologie Structurale Jean-Pierre Ebel (CEA, CNRS UMR5075, Université Joseph Fourier - Grenoble I), 41 rue Jules Horowitz, 38027 Grenoble, France

[∥]Institute of Environmental Assessment and Water Research (IDAEA), CSIC, Jordi Girona 18-26, 08034, Barcelona, Spain

Supporting Information

ABSTRACT: The role of motions in the catalytic cycle of an enzyme is the subject of much debate. Crystallographic results for the enzyme *N*-acetyl-L-glutamate kinase (NAGK), which is a suitable target for antibacterial drugs, suggest that a conformational compression of the active site favors catalysis. We have used a QM/MM scheme to compute energy profiles of the phosphoryl transfer reaction for 20 conformations of NAGK, starting from four crystal structures that represent different stages of the catalytic process. All paths show a common associative mechanism but with a wide range of barrier heights. The position of several active site residues and water molecules are found to determine the energetic barrier of each conformation, as revealed by principal component and partial least-squares chemometric analyses. In particular, conformations in which the two substrates have a shorter distance separation and a more linear mutual orientation tend to have lower energetic barriers, thus supporting the putative role of conformational compressive motions in catalysis. Interestingly, these motions are the same that lead to opening of the active site, which molecular dynamics simulations indicate is a fast process when the enzyme is free of substrates. Despite the lack of extended sampling, the energy barrier we calculate for the chemical step lies significantly below the apparent energetic barrier derived from experiment. Although not conclusive, this result supports a previous hypothesis, also derived from experiment, that conformational motions, rather than the chemical step, are rate limiting.



INTRODUCTION

The flexibility of proteins has been widely studied both experimentally and computationally. Enzymes, being proteins, are also flexible, but the role of flexibility in catalysis is still widely debated.[1] NMR studies have proved that proteins have access to an ensemble of conformations, encoded into their 3D structure.[2,3] That dynamics occur during a catalytic cycle is accepted by all scientists, but some argue that the term "dynamical effects" should only be used to assess deviations from Transition State Theory, which is an equilibrium theory. Some experimental[4] and computational[5−7] studies suggest that these dynamical effects are small or negligible in enzymes. By contrast other studies indicate that fast dynamics are implicated in the enzymatic cycle[8] via promoting vibrations that are coupled to the catalytic reaction coordinate.[9−11]

When slower conformational motions are present during the catalytic cycle, they can become the rate-limiting step, and, for some enzymes, NMR experiments indicate that this is indeed the case.[12] These motions are often associated with ligand binding processes, although they have also been observed in the free enzyme, pointing to an intrinsic functional dynamics.[13] It is unclear, however, whether dynamics at these millisecond time-scales help catalyze the chemical step[14] or not.[6,15]

The present study focuses on the amino acid kinase (AAK) family, in particular, on *N*-acetyl-L-glutamate (NAG) kinase (NAGK). The amino acid kinase family of enzymes comprises a series of enzymes that catalyze a phosphorylation reaction and have a high similarity in terms of sequence and structure. NAGK catalyzes the phosphorylation of NAG, which is the controlling step in arginine biosynthesis. This biosynthetic route in bacteria proceeds through N-acetylated intermediates, whereas in mammals nonacetylated intermediates are produced. Consequently, NAGK is a potential target for drugs that selectively inhibit the bacterial enzymes. The NAGK form of *Escherichia coli* (*Ec*NAGK) has been extensively characterized by biochemical and crystallographic methods[16−21] and is regarded as the structural paradigm of the AAK family of enzymes. Focusing on *Ec*NAGK, crystallographic studies by Rubio and co-workers[16,17,20] have provided insights into its mechanisms of binding and catalysis. *Ec*NAGK is a homodimer of 258 residues, each monomer being folded into an $\alpha\beta\alpha$ sandwich. The N-domain of each subunit makes intersubunit contacts and hosts the NAG binding site (NAG lid), whereas

Received: July 16, 2013
Revised: October 21, 2013
Published: October 22, 2013

   dx.doi.org/10.1021/jp407016v | *J. Phys. Chem. B* 2013, 117, 14261−14272

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 in *Ec*NAGK,[18] and so our study focuses on the monomer.

The diverse crystallographic structures solved for this enzyme indicate two types of functional motions:[17]

(1) X-ray structures of *Ec*NAGK complexed with either ADP or with the inert ATP analogue AMPPNP (PDB codes 1GS5, 1OH9, 1OHA, and 2X2W) have active sites that are too narrow to let the substrates bind directly, whereas structures with an unoccupied ATP site (PDB code 2WXB) have a more open active site that does allow the substrates to enter. This suggests that the C-domain and NAG lid undergo a conformational closure that is likely to be triggered by nucleotide binding (see Figure 2). According to the terminology coined by Gora and co-workers,[22] both flexible domains can be regarded as a "double drawbridge" gate, which is rare in active site entrances.

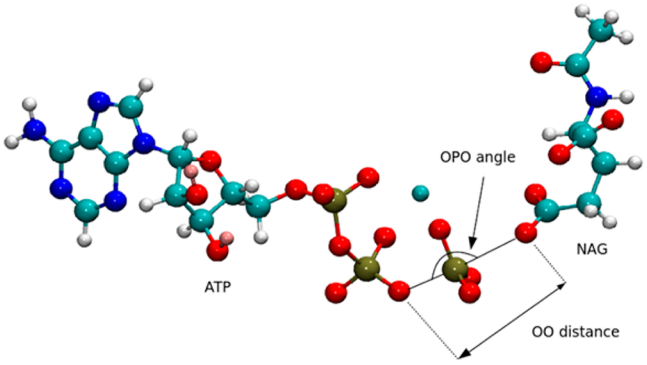


Figure 1. Structure of reactants with some relevant geometrical parameters.

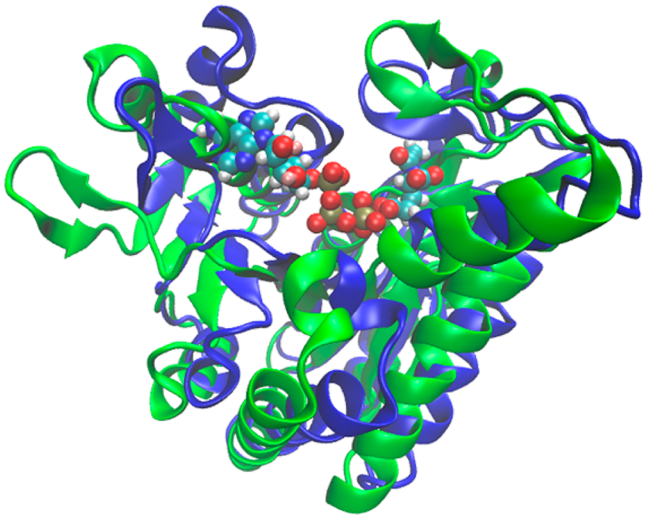


Figure 2. Open (green) and closed (blue) conformations of NAGK. The substrates, in van der Waals spheres, correspond to the closed conformation.

(2) Rubio and co-workers[17] hypothesize that in the closed form of *Ec*NAGK the narrowness of the active site exerts a "conformational compression" on the substrates (the O−O distance in Figure 1) that favors catalysis. This conclusion is deduced solely from the crystal structures.

Our recent computational studies on the AAK family[23,24] showed that the large-amplitude motions of *Ec*NAGK are intrinsic to the enzyme, and are shared among other family members, thereby pointing to a common mechanism of action. We also found that the oligomeric assembly enhances both intra- and intersubunit collective motions. The latter are especially important for AAK members with allosteric regulation as such cooperative motions between subunits are ultimately responsible for regulating substrate binding events.[24]

In this work we use computational methods to evaluate the role of conformational motions in the chemical step of the reaction catalyzed by NAGK. More specifically our aims were to (1) study the reactivity of the different crystal structures of *Ec*NAGK; (2) estimate the significance of "conformational compression"; (3) determine whether induced fit is a plausible mechanism for catalysis; and (4) identify to what extent conformational motions determine the overall turnover of the enzyme.

METHODS

We investigated the catalytic mechanism of *Ec*NAGK using hybrid quantum mechanical (QM)/molecular mechanical (MM) potentials in combination with reaction path calculations. All QM/MM simulations, including system setup, were done with the pDynamo[25] program. Calculations with QM/MM potentials that employed density functional theory (DFT) methods were performed using pDynamo and its interface to the ORCA[26−28] quantum chemistry package. We also carried out some MM molecular dynamics (MD) simulations using the Gromacs[29] program. Figures 1 and 2 have been generated with the VMD code[30] and Figures 3−10 have been generated with Matplotlib.[31] Figures in the Supporting Information use VMD, matplotlib, LigPlot+,[32] and Matlab.[33]

System Setup. We employed four crystal structures of *Ec*NAGK, with PDB entries 1GS5, 1OH9, 1OHA, and 2X2W. For each of these crystal structures, we built a model of the enzyme complexed with its natural substrates, ATP and NAG.

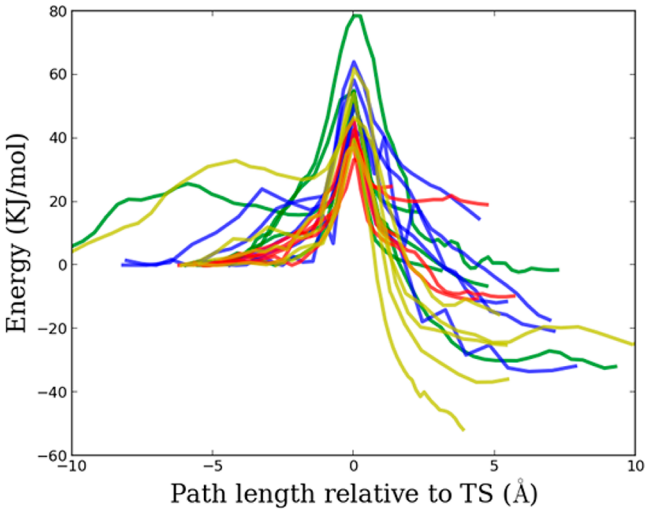


Figure 3. Energy vs reaction coordinate for the optimized NEB reaction paths. Each color represents a different crystal structure: green, 1GS5 snapshots; blue, 1OH9 snapshots; red, 1OHA snapshots; yellow, 2X2W snapshots.

dx.doi.org/10.1021/jp407016v | J. Phys. Chem. B 2013, 117, 14261−14272

Structure 1GS5 contains the ATP analogue AMPPNP and NAG, so we replaced the NH group linking P_γ and P_β of AMPPNP with an oxygen atom. For 1OH9, we replaced the AlF_4 moiety of the ADP-AlF_4-NAG complex with a PO_3 phosphoryl group, whose coordinates were obtained by superimposing the ATP structure from 1GS5 into the ADP moiety of 1OH9. The same procedure was used to add the transferring PO_3 group to structure 1OHA, which only contains ADP and NAG. Structure 2X2W chain A represents the product complex of the enzyme, which contains phosphorylated NAG (NAGP), using the same procedure as before we superimposed the Mg and ADP moiety from 1GS5.

Each simulation system consisted of a single monomer of the corresponding crystal structure complexed with the natural substrates. The protein structure was immersed in an orthorhombic water box. To simulate conditions equivalent to those of kinetic experiments carried out on NAGK,[16] protonation states were assigned at pH = 7.0 with PropKa,[34] and K^+ and Cl^- ions were added to achieve a salt concentration of 100 mM. The OPLS/AA force field[35] was used to describe protein atoms and the TIP3P[36] for water molecules.

To generate suitable starting structures for the simulations, the structure of the solvated protein was first energy minimized. This was followed by a short molecular dynamics simulation at a temperature of 300 K with position restraints on the heavy atoms of the protein to equilibrate the solvent. Finally, in a second MD simulation, the position restraints on the protein atoms were progressively relaxed until all atoms were free to move.

The aim of this protocol was to obtain structures for the protein that were relaxed, but as close to the initial crystal structure as possible. Nevertheless, in all the runs, we found significant distortion of the active site geometry. In particular, Asp162 coordinated with the Mg cation in the active site, although it is known that it coordinates two neighboring lysines, Lys8 and Lys217, which have an important role in anchoring the substrates.[18] Likewise, we noted that NAG coordinates to the Mg cation, which is again something that is not observed in the crystal structures. The situation was not improved by employing semiempirical QM/MM potentials, although DFT QM/MM approaches provided better results (see below for a fuller discussion). This sensitivity of the active site geometry to the energy function was highly unsatisfactory given that our aim was to compare the reactivity of different crystal conformations. As a result, we decided to restrict our equilibrating dynamics to the solvent and the hydrogens of the protein and substrates, thereby preserving the initial crystal structure and preventing inappropriate conformations of the system. For each crystal structure, we performed a MD simulation of 100 ns, and selected five snapshots at 20 ns intervals as starting points for our QM/MM calculations.

QM/MM Potentials. Our initial choice for the QM method to use for the atoms in the QM region was the semiempirical AM1/d-PhoT,[37] as it outperforms the original AM1 model in the description of phosphoryl transfer reactions. AM1/d-PhoT includes d-orbitals and incorporates a scaling factor in the core−core term that attenuates the artificially attractive interactions involving P atoms. Our previous calculations[38] showed the ability of AM1d-PhoT to describe the phosphoryl transfer, but it has important shortcomings for the present system. First, it severely underestimates the exothermicity of the reaction and, second, carboxylic oxygen shows a too strong interaction with the Mg cation, thereby causing NAG to coordinate the metal very rapidly during a simulation. Both these failures can be ascribed to the following limitations in the parametrization of the AM1d-PhoT method: (i) carboxylic acids were not included as nucleophiles in the parametrization training set; and (ii) the Mg cation was not reparameterized, even though it frequently accompanies ATP.

A further problem that we observed for the semiempirical QM/MM method was the tendency of MM waters to substitute for QM ones in the coordination shell of the Mg cation. In our system, one needs to treat the Mg coordinating water molecules in the QM region (see below) to get a correct exothermicity, but the exchange of QM and MM waters was difficult to prevent, even with the use of constraints. Given that the main advantage of a semiempirical method is the possibility to perform molecular dynamics, and the fact that the AM1d-PhoT method cannot be used for our system, we decided to employ a DFT QM/MM method instead.

We used the *m*PWPW91[39] DFT functional, as this has been shown to reliably describe the geometry and energetics of pentacoordinated phosphorus species[40] and of enzymatic phosphoryl transfer reactions.[41]

The QM region used had 34 atoms and comprised the three phosphates of ATP, the acetyl fragment of N-acetylglutamate (CH_2-COO^-), the Mg cation, and the three water molecules that coordinate it (see Figures S1 and S5, Supporting Information). The inclusion of Lys8 and Lys217 was also considered (see Figures S3 and S4), as suggested by one of the reviewers, but their inclusion does not change the shape of the profile, and if anything, the energy barrier decreases. The basis set used in all geometry optimizations and reaction path calculations was the Ahlrichs split valence plus polarization SV(d) for C, Mg, and N, and SV for hydrogens, and a SV(p)+ for P and O. The need for a diffuse function on P was based on our previous experience,[38] and on O, because most of them bear a negative charge. Single point calculations were performed with Ahlrich's triple-ζ TZV(2d) basis set (hereafter denoted "large basis set").

For the DFT QM/MM simulations, the systems setup in the previous section were pruned by removing all residues of the system that had no atoms that were less than 25 Å away from the γ-phosphate of ATP. The final systems contained approximately 7200 atoms, with the exact number depending on the starting crystal structure. In the geometry optimizations and reaction path calculations, only atoms within 20 Å of the γ-phosphate of ATP (approximately 4200 atoms) were allowed to move, with the positions of those outside of this remaining fixed. In the QM/MM Hamiltonian, all electrostatic interactions between QM and MM atoms were evaluated with no cutoff, irrespective of whether they were mobile or not.

Geometry optimizations were carried out with the double-ζ basis set, followed by single point calculations with the large basis set. The latter are the final energies reported in the manuscript. The larger basis set increases the energy barrier, but does not change the position of the minima or the TS (see Figure S2), thus, making single point calculations a valid approach for improving the accuracy of our calculations.

To evaluate the influence of the solvent configuration on the total energy, we have performed Poisson−Boltzmann calculations with the APBS software.[42] We report the details in the Supporting Information.

Calculation of Reaction Paths. The higher computational demand of DFT QM/MM potentials makes free energy calculations difficult and so we optimized reaction paths

between different conformations of the systems using the Nudged Elastic Band (NEB) method.[43−46]

Five different reaction paths were optimized for each system using snapshots taken from the initial MD trajectories. These snapshots corresponded to the reactants for the 1GS5, 1OH9, and 1OHA systems and to the products for the 2X2W system. To obtain the corresponding products or reactants (the other ends of the paths), we geometry optimized by constraining the P−O bonds on ATP and NAG and then releasing the constraint. The NEB calculations themselves were started off with a small number of intermediate structures but gradually increased until the energy profile converged. The number of structures per path depended on the path length and ruggedness, and it ranged from 19 to 57 images, with an average of 33.

Statistical Analysis. To evaluate the diversity of the investigated configurations, we employed a principal component analysis (PCA)[47] of various geometrical parameters that have been postulated to be important in the catalytic cycle of the enzyme. PCA defines a reduced set of new variables, as an orthogonal and linear combination of the original variables that provides a simplified description of the correlations between parameters and the variance in conformations.

As a follow up to the PCA, we used a partial least-squares regression (PLSR)[48] to analyze the conformations. PLSR is a well-known multivariate linear regression chemometrics method that does not suffer from linear dependencies among measured parameters and avoids noise overfitting problems when the number of correlated variables or parameters is high. PLSR produces new latent variables that optimally predict changes in the independent variables (here the energy barriers or y block values) from the observed variance in the dependent variables (here the geometrical parameters or x block values). Since PLS is based on a bilinear model, it is a reliable and robust method for the validation and interpretation of statistical data. In this work, the PLS Toolbox package[49] was used within the MATLAB computer and visualization environment.[33]

Molecular Dynamics. We carried out MD simulations for analysis purposes on the 1GS5, 1OH9, and 1OHA systems using the Gromacs[29] program. Two were performed with the OPLS[35] forcefield and one with the Ambers99SB-ILDN[50] parameter set, for 1GS5. For 1OH9 and 1OHA, one with OPLS and one with Amber force field were performed. The SPC[51] water model was employed in each case, with the imposition of periodic boundary conditions via a rhombic dodecahedral box. The temperature was kept at 300 K, using a Berendsen thermostat[52] with time constant τ = 0.1 ps, whereas the pressure was kept at 1 bar using the Berendsen barostat[52] with an isotropic compressibility of 4.5×10^{-5} bar^{-1} and time constant of 0.5 ps. The integration time step was 2 fs, and bond lengths to hydrogen were constrained with the LINCS algorithm.[53,54] Electrostatics were computed via PME[55,56] using a grid of 1 Å, and van der Waals interactions with a switch function between 0.8 and 0.9 nm.

Before production runs were started, the structure of the solvated enzyme 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 heating simulation was performed from 0 to 300 K, followed by a 40 ns simulation at 300 K, with snapshots being saved every 50 ps for subsequent analysis. For the production run, all position restraints were removed. This was possible because the constraints used to sample the QM/MM starting structures were needed to keep the substrates in their experimental position, whereas these simulations were for the apo enzyme, free of substrates.

Average Energy Barrier and Experimental Energy Barrier. The comparison of kinetic data to calculations is indirect. The correct way to obtain an activation barrier is to calculate the variation of the rate constant with temperature. When only a single rate constant is known, one can calculate an apparent free activation energy from transition state theory using the Eyring equation. Under this approximation, all dynamical, tunneling, and temperature effects are included in the free energy, as this is the single variable that is fit to the rate constant, although they should in principle form part of the pre-exponential factor to the rate. Nevertheless, in the majority of cases, the true free energy of activation will remain the main contribution to the apparent free energy.

The experimental catalytic rate constant (k_{cat} = 40 s^{-1}) obtained by Rubio and co-workers[16] corresponds to an energy of activation of 66 kJ/mol, according to the Eyring equation,

$$\Delta G_i^{\ddagger} = -RT \ln\left(\frac{k_{cat}}{k_B T/h}\right)$$

Here k_{cat} is the rate constant, k_B is the Boltzmann constant, T is the temperature (310 K), h is the Planck constant, and $\Delta G_i^{\ddagger}$ is the free energy barrier. To compare this value with the ensemble of values obtained from the different snapshots, one has to take into account the exponential weight of the barrier to the rate constant. This means one has to average the rate constants and then obtain an apparent barrier that would give the average rate:[57,58]

$$\Delta E_{avg}^{\ddagger} = -RT \ln\left\{\frac{1}{n}\sum_{i=1}^{n} \exp\left(\frac{-\Delta E_i^{\ddagger}}{RT}\right)\right\}$$

Here $\Delta E_{avg}^{\ddagger}$ is the average barrier height, R is the gas constant, n is the number of energy profiles considered, $\Delta E_i^{\ddagger}$ is the barrier height of each snapshot, and T is the temperature (310 K). Because NEB calculations neglect entropy contributions, we assume that $\Delta E_i^{\ddagger}$ is a good approximation to $\Delta G_i^{\ddagger}$, as has been done in previous studies.[58,59]

RESULTS AND DISCUSSION

Heterogeneity of Energy Profiles Across Different Conformations. Figure 3 shows the energy profile for the 20 reaction paths that were optimized, and the most relevant energetic parameters are detailed in Table 1. There is considerable heterogeneity in the profiles, but the figure conveys two messages. First, the phosphoryl transfer from ATP to NAG takes place through a one-step mechanism for all the conformations. Second, the reaction is exothermic in most of the conformations, as expected from the course of the reaction in the enzyme's metabolic pathway. Interestingly, there is a wide range of variation in the activation energies (33−78 kJ/mol), as well as in exothermicity. This range, although large, has been observed for other enzymes, including chorismate mutase[57,60−62] and a fatty acid amide hydrolase.[63]

The barriers are lower than the apparent free energy of activation of 66 kJ/mol for all structures, except for one. We have checked that this discrepancy does not stem from a too small QM region. We have recomputed the NEB profiles with

dx.doi.org/10.1021/jp407016v | J. Phys. Chem. B 2013, 117, 14261−14272

Table 1. Energies (in kJ/mol), with Respect to Reactants, of the TS and Product Structures Obtained from the Optimized NEB Reaction Paths

NEB	TS (kJ/mol)	products (kJ/mol)
1GS5		
1	55.0	−6.5
2	51.4	−32.4
3	41.0	−1.7
4	78.6	−2.2
5	54.6	16.8
1OH9		
6	45.0	−33.5
7	42.5	−20.7
8	58.4	−17.2
9	50.4	−11.9
10	64.1	14.7
1OHA		
11	45.7	9.1
12	38.1	9.5
13	41.7	19.1
14	33.3	−10.0
15	37.2	−10.9
2X2W		
16	39.5	−36.9
17	54.1	−24.9
18	46.5	−15.5
19	61.9	−25.1
20	36.9	−51.7

an expanded QM region including two lysines of the active site (Lys8 and Lys217) that interact with ATP and found that the shape of the profiles does not change (see Figures S3 and S4) and that the barriers do not increase.

In all reaction paths there is a single transition state in which the transferring phosphoryl adopts a bipyramidal geometry and the axial bonds are the ones that are being formed and cleaved. As shown in Figure 4, the sum of these two bond distances (the O−O distance) is shorter in the TS than in reactants and products, and in almost all profiles the minimum O−O distance corresponds to the TS. Qualitatively this provides a geometric argument in support of the conformational compression hypothesis.

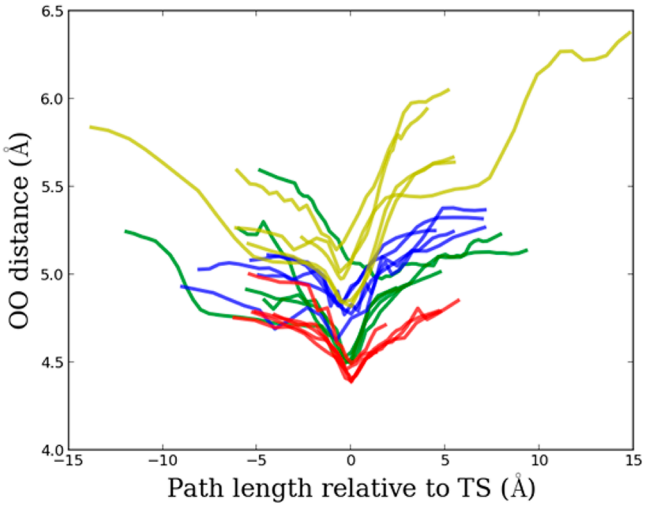


Figure 4. O−O distances along the optimized NEB reaction paths. The colors correspond to those of Figure 3

A comparison of the P−O distances in reactants and products shows that the acceptor, NAG, approaches the P atom before the phosphoryl moiety is cleaved, thereby resulting in an associative mechanism. We note that a carboxylic acid is not a good nucleophile and thus tends to favor dissociative mechanisms, which suggests that it is the configuration of the enzyme active site in this case that drives the associative process. As this is the first theoretical study on acyl-kinases, we wonder whether other members of this and related families with different tertiary and quaternary structures, such as acetate kinase, phosphoglycerate kinase, and biotin carboxylase, share the same associative mechanism.

Figure 5 reveals that when approaching the TS, the O−P−O angles are close to linearity, as expected for a S_N2 reaction. The

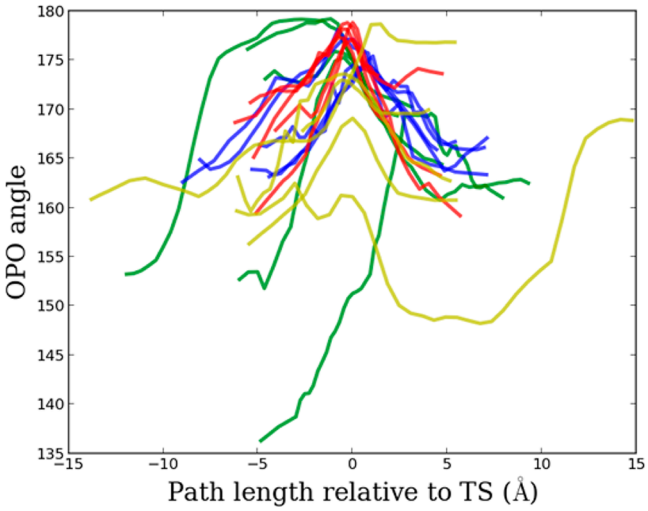


Figure 5. Change of the O−P−O angles along the reaction path. The colors correspond to those of Figure 3. Except for some ill-behaved cases discussed in the text, the O−P−O angle approaches linearity at the TS.

only snapshot where this does not happen is the fourth snapshot of 1GS5, which has a very high barrier. As pointed out below this is due to the absence of a bridging water molecule that induces the formation of a strong salt bridge interaction between Lys61 and NAG. It is interesting to emphasize the idea that enzymes do not need to bind their substrates very tightly, because otherwise, the cost of bringing them to the TS structure increases.

Figures 4 and 5 illustrate that there is a significant variability in the TS O−O distance and O−P−O angle, respectively, and the same is evident when other structural parameters of the TS are analyzed, including the O−P and P−O distances, and hydrogen bonds with active site residues (data not shown). This variability is the source of the different energy barriers, but it does not arise from bringing different reactant conformations to a single rigid TS geometry, but from different reactant conformations having different reaction valleys with different energy barriers.

Geometrical Characterization of Conformational Diversity in the Active Site. *Principal Component Analysis.* To disentangle the information coded in the different configurations and the source of the different energy barriers, we have performed a statistical analysis of 16 geometrical

dx.doi.org/10.1021/jp407016v | J. Phys. Chem. B 2013, 117, 14261−14272

parameters that are detailed in Table 2 using an approach similar to Lodola et al. The parameters are considered at both

Table 2. Geometrical Parameters Used to Characterize the 20 Conformations[a]

parameter description	parameter index (reactants and TS)
N(Lys217)−O2B(ATP)	1,2
N(Lys8)−O1G(PO3)	3,4
N(Gly44)−O3G(PO3)	5,6
N(Gly11)−O3G(PO3)	7,8
N(Gly45)−OE2(NAG)	9,10
O3B(ATP)−OE2(NAG)	11,12
N(Lys61)−OE2(NAG)	13,14
OPO angle	15,16

[a]The residues selected have been identified as being relevant for binding or catalysis in previous work.[17,20] The values of each parameter at the reactant (odd indexes) and TS (even indexes) structures are considered.

the reactant and TS geometries and are the distances between several active site residues and the PO_3 moiety; the O−O distance between the ATP and NAG oxygen atoms that transfer the PO_3 moiety; and the O−P−O angle. The structure of the active site and the positions of the selected residues in the parameters are depicted in Table 2 and Figures S1 and S5.

We started by performing a principal component analysis (PCA) of the geometrical parameters to evaluate the diversity of the configurations and to simplify their characterization. The results are given in Table 3 from which it can be seen that the first two principal components (PCs) explain 80% of the parameter variance.

Table 3. PCA Analysis of the Geometrical Parameters[a]

principal component	eigenvalue	% of variance captured	% of accumulated variance
1	9.92	58.1	58.1
2	3.51	21.9	80.0
3	1.39	8.7	88.7
4	0.726	4.5	93.2
5	0.399	2.5	95.7
6	0.280	1.8	97.4

[a]The first two PCs account for 80% of the variance of these parameters.

Figure 6 displays the projection of the 20 conformations on the first two PCs. PC1 distinguishes 2X2W structures from the others, whereas PC2 mainly differentiates the fourth 1GS5 structure, which has the highest energy barrier (see below). As shown in Figure 7, PC1 has large contributions from all parameters except for the O−P−O angle and the Gly11 distance (15,16 and 7,8, respectively). This is probably due to the loss of a large number of hydrogen bonds in the 2X2W structures and also because the O−P−O conformation angle is less linear in this case.

PC2 is mainly loaded by the Lys61 distance and the OPO angle (Figure 7). The interaction between the lysine and the carboxylate moiety of NAG leads to an orientation of NAG with respect to ATP that is much less favorable for nucleophilic attack. A closer analysis of the 1GS5 snapshots reveals that the short contact between Lys61 and NAG is due to a missing water molecule that bridges Lys61 and NAG in the crystal (and

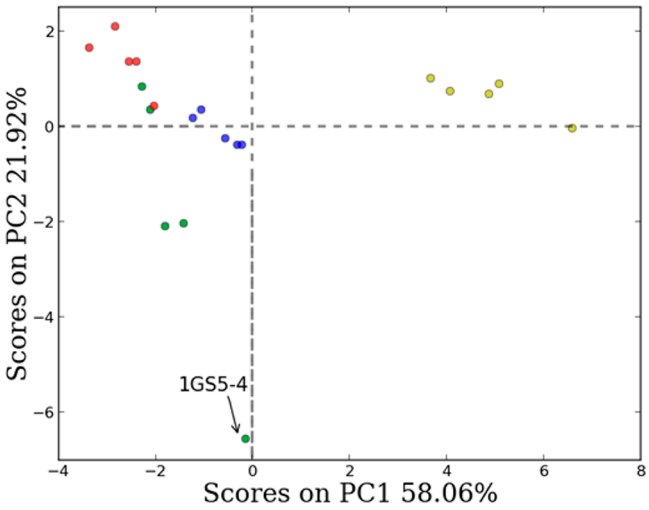


Figure 6. Projection of the parameters on the first two principal components. The first PC differentiates structures 16−20 (from 2X2W), whereas the second mainly differentiates structure 4 (from 1GS5). The colors correspond to those of Figures 3−5.

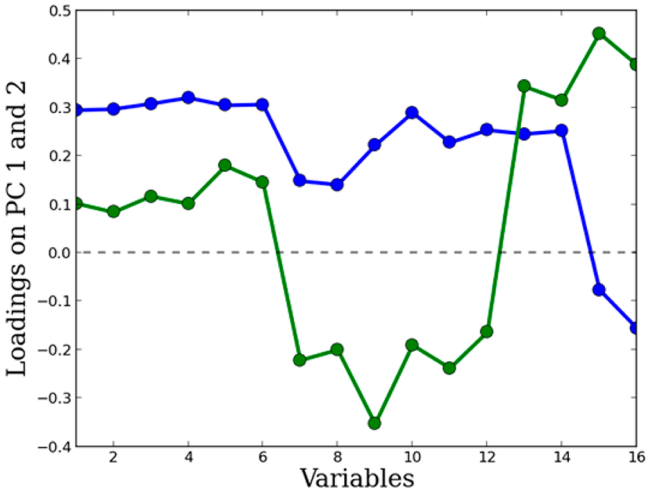


Figure 7. PC1 loadings in blue. All parameters, except for distances to Gly11 (7,8) and O−P−O angles (15,16), contribute significantly to this PC, which serves to differentiate 2X2W conformations from others in Figure 6. PC2 loadings in green: Gly45(9), Lys61(13,14), and the OPO(15,16) are the main contributors to this PC, which distinguishes the fourth 1GS5 snapshot.

all of the other snapshots). The absence of this water in 1GS5 allows the formation of a strong salt bridge between Lys61 and the carboxylate of NAG in the reactants, which must be disrupted if the PO_3 moiety is to be accepted from ATP, thereby adding a substantial energetic cost to the phosphoryl transfer.

All 1GS5 structures lack this water molecule and, among them, the fourth snapshot, which has the shortest Lys61−NAG distance, is the one with the highest energy barrier in this study (80 kJ/mol). Although other structures have longer Lys61−NAG distances and lower barriers, it appears that the bridging water is necessary to have the most proficient catalytic conformation. In general, a strong interaction between the substrates and the enzyme decreases K_m, but may also decrease k_{cat} if that interaction needs to be weakened to reach the products. This has been highlighted, for example, by Lluch and

dx.doi.org/10.1021/jp407016v | J. Phys. Chem. B 2013, 117, 14261−14272

co-workers who showed that the presence of a tyrosine in the active site of ferredoxin-NADP+ reductase destabilizes reactants and thereby reduces the energy barrier.[64,65]

Partial Least Squares Regression Analysis. Table 4 shows the PLSR results for our data and gives a picture that is different from that given by the PCA. Although two components still explain about 80% of the variance in the parameters, these components only explain 56% of the variance in the energy barrier. A third component, with a minor 5% weight in the parameters, has a 21% weight in the variance of the barrier. The first PLSR component consists of principally the Gly41, O−O, and Lys61 (mainly at the TS) distances, and the O−P−O angle (Figure S6).The second component, which has a minor influence on the barrier, is described by the Lys61 and, to a lesser extent, Gly44 distances, and the O−P−O angle at the TS (Figure S7). The third component is again determined by Lys61 and the change of the O−P−O angle from reactants to TS (Figure S8).

Table 4. Results of the PLSR[a]

	percent variance captured by regression model			
principal component	X block, this component	X block, total	Y block, this component	Y block, total
1	33.6	33.6	47.4	47.4
2	45.8	79.3	9.1	56.5
3	4.8	84.1	20.1	77.5
4	5.7	89.8	9.3	86.8
5	4.5	94.2	3.8	90.6
6	1.8	96.0	1.2	92.6

[a]As before, two components (though not with the same composition) capture almost 80% of the variance in the parameters (X block). However, the second component has little impact on describing the barrier (Y block), and one needs a third component to account for almost 80% of the Y block variance. Interestingly, this component has a minor effect (4.8%) on the description of the X block.

For a clearer interpretation of the role of each geometric parameter in the energy barrier we have calculated the variable importance on projection (VIP) score for each parameter in the first three components of the PLSR.

The results are displayed in Figure 8, with scores above unity being regarded as significant.[66] These show the importance of the O−P−O angles, the O−O distance in the reactants, and to a lesser extent in the TS, and the Lys61−NAG interaction at the TS on the prediction of energy barrier values. They also highlight the role of the Gly45−NAG hydrogen bond, which is a strong interaction in 1OHA, a weaker one in 1OH9 and 1GS5, but is absent in 2X2W.[17] Our analysis reveals its role in stabilizing the TS and thus reducing the energy barrier.

There is a correlation between both the O−O distance in the reactants and in the TS since lower energy barriers mean shorter distances. On the contrary, there is no correlation between the *change* of the O−O distance between reactants and TS and the energy barrier. That is, it is not the compression *along* the reaction that plays a role. In fact, this value is relatively constant. This makes sense, because the motions that cause the compression have a much slower time scale than the chemical reaction, and cannot be coupled to them. Therefore, there is not a dynamical effect during the chemical reaction, at least for the variables that we have considered. However, slow large amplitude motion will bring the enzyme to conformations that are more compressed and reactive than others.

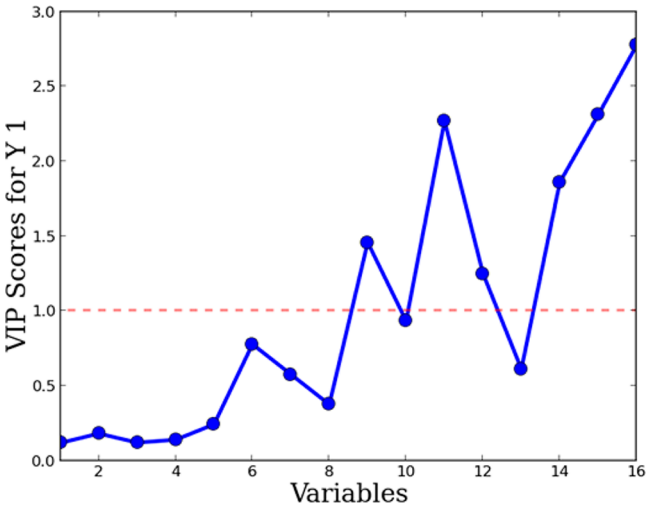


Figure 8. VIP scores using three PLSR components. Scores above one are considered to be the more significant ones. Hydrogen bond of Gly45 at the reactants (9), O−O distance (11,12), interaction of Lys61 at the TS (14), and O−P−O angle (15,16) are the most influential parameters in the prediction of energy barrier values of the 20 snapshots.

These slow motions on the time scale of the reactions should not be confused with the still slower on−off states detected in fluorescence spectroscopy.[67,68] In our case, the compressive motions, associated to the open and closed states, will take place at least once during each catalytic cycle, as the open conformation is necessary for product release and entry of reactants. If the motions are slow enough that they only take place once, they will be rate limiting. By contrast, the on−off states seen in other enzymes last for several catalytic cycles and so the enzymes have memory whether they are in an active or an inactive conformation.

The PLSR results also showed the relevance of several residues in determining the barrier height. However, it is intriguing that the 2X2W profiles, whose structures lack interaction with many catalytic residues, do not have very high energy barriers. This can be rationalized by noting that the 2X2W reactant structures were prepared from a conformation that contained products in the active site using a local minimization procedure. This was not sufficient to permit the enzyme geometry to fully adapt to the reactants, and means that they are less stabilized by the enzyme than the products. This explains why the reactant and TS energies can be so low and also why the 2X2W profiles are more exothermic than those for the other structures (−31 kJ/mol compared to −7 kJ/mol for the average of the other 15 structures). The high similarity between the reactant-like crystal structures (1GS5) and the TS-like crystal structure (1OH9) shows the lack of large conformational motions from reactants to TS and thus makes the calculations of the energy barrier from reactants more reliable.

A full adaptation of the enzyme geometry to the structures along the reaction path can, in principle, be achieved with a number of techniques, including potential of mean force (PMF) calculations as a function of a set of reaction coordinates. However, such calculations often require simulation lengths to converge that render them impractical with QM/MM methods, and these effects will only be exacerbated if sampling of slow motions of the enzyme is also necessary. The

fact that different enzyme conformations can lead to different energy profiles has already been pointed out by Mulholland and co-workers.[1,63,69] An important difference with our work, however, is that they used only a single initial structure, whereas we employ starting configurations from four different crystal structures, thereby allowing a more in-depth of the conformational dependence of the energy profile.

The "energized" reactant structures in 2X2W shows that the active site is not fully preorganized to stabilize only the TS,[5,7,70] but that it can fluctuate and stabilize preferentially the reactants and products. Despite being present, this reorganization has been shown to be smaller in enzymes than in water,[71] and it is one of the sources of the catalytic effect.

Overall, these results show that the variability of energy barriers within and among crystal structures is similar. Although all snapshots show a one-step reaction, the "energized" reactants of 2X2W do not represent the most stable conformation. Likewise, the strong interaction between Lys61 and NAG in the 1GS5 structures leads to some structures with high energy barriers and highlights the role of water coordination in the active site.

Water Dynamics and Magnesium Coordination. In the 1GS5, 1OH9, and 1OHA crystal structures, there are two water molecules that bind the PO_3 moiety. In the corresponding reaction paths, these molecules follow the PO_3 along the reaction path, thereby supporting the suggestion by Rubio and co-workers that they aid catalysis by stabilizing the charge of PO_3.[17]

We have also analyzed the coordination sphere of the Mg cation. In the snapshots from 1OHA and 1OH9, as expected, this ion remains hexacoordinated throughout all the reaction profiles. The 2X2W snapshots show a pentacoordinated Mg cation, whereas for 1GS5, coordination varies, both among snapshots and along the reaction path. We could not correlate these coordination changes to higher or lower energy barriers, and we believe it also proves the incomplete relaxation of reactants in these two sets of structures.

An active role for solvation might provide a simple explanation for the diversity of energy profiles computed for the chemical step, that is, different configurations of active site water molecules lead to different energy barriers, as the NEB reaction path calculations do not allow solvent equilibration. Although we have shown an example of the large extent to which the absence (or presence) of a water bridging NAG and Lys61 can alter the computed energy barriers, we note that active site solvation is not the only source of catalytic hetereogeneity. To make an estimation of the effect of a truly equilibrated solvent on the energy barriers we have performed Poisson−Boltzmann (PB) calculations of reactants, transition state, and products. Despite the oversimplified description of the PB approximation, we find a variety of energy barriers that is even broader than that obtained from the QM/MM explicit description of solvent. The variability of the energy barriers comes, in part, from the solvent contribution, but to a larger extent, from the protein conformation (see Table S1). This indicates that besides active site solvation, other factors also play an important role in modulating the energy barrier, mainly conformational compression of the active site.

Conformational compression in 1OHA and 1OH9. We now turn our attention to evaluating the effect of conformational compression. Given that the water structure in the active site of 1GS5 disagrees with the crystallographic evidence, and that 2X2W structures have energized reactants, we have excluded these two structures from the following analysis and focused solely on the 1OHA and 1OH9 structures. When considering all 20 structures, several geometrical parameters are needed to explain the heterogeneous energy barriers, but a new picture emerges if we only examine the group of 10 structures from 1OHA and 1OH9 that have the most probable reactant conformations.

For the snapshots obtained from these two structures, we found a striking correlation of the energy barrier with two geometric parameters directly linked to the conformational compression hypothesis: the O−O distance and the O−P−O angle at the TS (see Figure 9). The correlation is valid across a wide range of energy values, 35−65 kJ/mol, with a Pearson correlation coefficient r^2 = 0.93, and is shared by snapshots of the two crystal structures. The lower barriers of the 1OHA structures, compared to those of 1OH9, correlate with shorter average O−O distances (4.43 vs 4.84 Å) and more linear average O−P−O angles (177° vs 175°). Overall, this correlation underscores the fact that both compression and proper mutual orientation of the active are important for lowering the energy barrier.

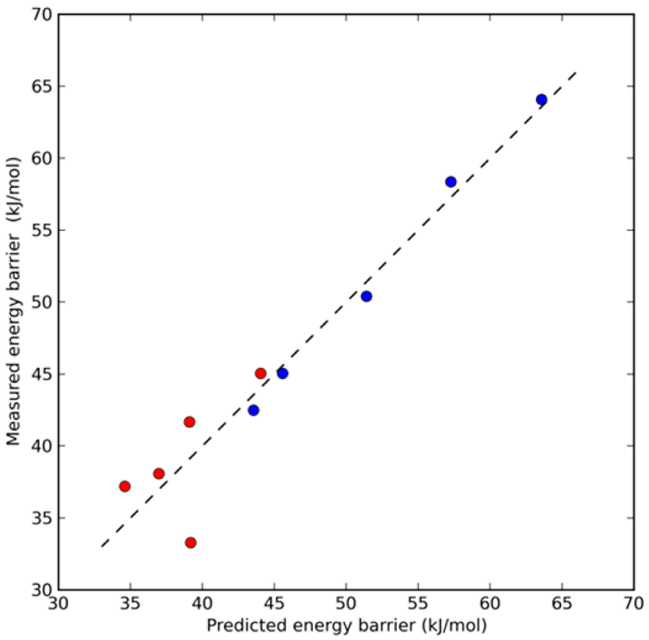


Figure 9. Measured vs predicted energy barrier values using a linear regression model with with two variables: the OO distance and the OPO angle at the TS. The equation is y = 12.3 $OO_{distance}$ − 2.68 OPO_{angle} + 460.5. Red dots: 1OHA structures; blue dots: 1OH9 structures.

The reason that 1OHA has longer O−O distances (4.40 Å) than 1OH9 (4.21 Å) is that the 1OH9 crystal structure contains AlF_4 as a mimic for PO_3. Despite the TS-like geometry of AlF_4, Waltho and co-workers[72] showed that it is the anionic charge of AlF_4, rather than its geometry, that mainly determines its tight binding to phosphoryl transfer enzymes. Indeed, the geometries of the AlF_4 TS analogue and the actual TS are quite different as AlF_4 is not bipyramidal and the Al−O and P−O distances are also not the same. This highlights the danger of extrapolating a detailed geometrical analysis of TS analogues to the true TS.

The average apparent energy barrier in the 1OHA and 1OH9 structures is 39 kJ/mol, which is 28 kJ/mol below the reference value. Part of this discrepancy arises from the reduced entropy

dx.doi.org/10.1021/jp407016v | J. Phys. Chem. B 2013, 117, 14261−14272

of the TS, that we do not account for in our calculations. Nevertheless, a difference of 28 kJ/mol is large. The presence of large-scale conformational motions has been demonstrated for this enzyme with several crystal structures. In other enzymes these kinds of motions are the rate-limiting step for the catalytic turnover. Very recently, an NMR study of another enzyme transferring a phosphoryl group (a phosphatase) showed that loop motions determine the enzyme turnover.[73] Our identification of a low chemical energy barrier can be explained if the chemical step is not rate limiting, but the conformational motions are, as also suggested by Rubio and co-workers.[16] The identification of a partly closed crystal structure containing the products (2X2W) might indicate that product release is the slow potentially rate-limiting step.

One could argue that this compressive motion is non-catalytic, because if the enzyme were always in a rigid compressed structure, the barrier for catalysis would be lower. Although that is true, it neglects the fact that reactants and products need to get in and out of the active site, something that would be extremely slow in a closed conformation. By oscillating between these two conformations, the enzyme can both have the cake (the substrates) and eat it (do catalysis). We finish by emphasizing that these large scale dynamic motions, although relevant for catalysis, are completely in equilibrium and so do not violate Transition State Theory

Stability of the Closed Conformation. We have seen that conformational motions are relevant to catalysis and we have suggested that the open form is needed for binding. Either under an induced fit or a conformational selection scenario, the open form should be lower in energy in the unbound enzyme. To test this last hypothesis, we performed seven Molecular Dynamics simulations, two of a closed conformation 1GS5, two of 1OH9, and two of 1OHA without the substrates. As Figure 10 shows, the closed form opens after less than 10 ns of simulation when the substrates are missing, which implies that the open form of the apoenzyme is more stable. For the other structures (Figures S9−S14), the opening time changes, but the behavior is the same. We have checked that this observation is consistent in two different force fields (OPLS and AMBERSB99-ILDN)

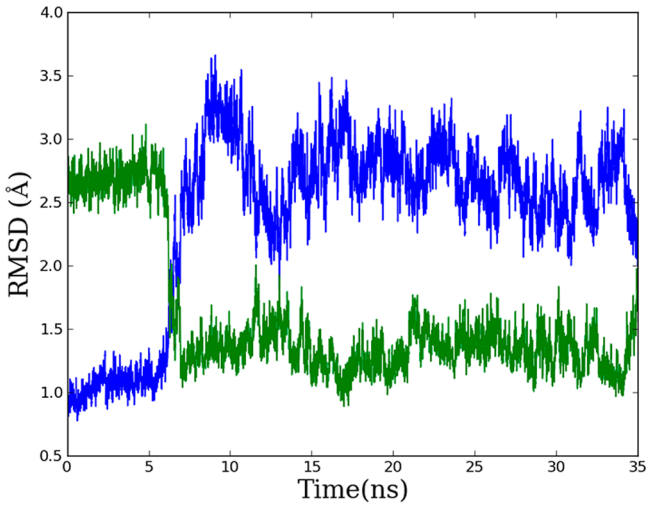


Figure 10. RMSD with respect to the closed 1GS5 structure (blue) and the open 2X2W chain A structure (green) for the a MD simulation of 1OHA with the AMBERSB99-ILDN force field. In less than 10 ns, the apoenzyme opens its lid spontaneously. The results for the other simulations are plotted in Figures S9−S14, and reproduce the behavior of this simulation. The 1GS5 structure has been taken as the reference as this structure is the one used by Rubio and co-workers as the model for the closed structure. Its RMSD with 1OHA is 0.36 Å.

The fact that this motion is fast agrees with our previous studies that reported the opening of the enzyme as one of the most easily accessible collective motions in this protein family.[23] This fast opening of the active site, however, contrasts to the closed form found for the crystal structure of the products (PDB code 2X2W) and to the suggestion that lid-opening is the rate limiting step. It is worth pointing out, however, that the opening event will probably be much rarer when the substrates are present as both the reactants, ATP and NAG, and products, ADP and phosphorylated NAG, are likely to stabilize the closed conformation. This is because the main role of the lid is not only to compress the reactants, but also to avoid compressing (and thus phosphorylating) unwanted substrates, in particular the more abundant glutamate. The role of gates in selecting substrates has been reviewed in ref 22, but we leave the study of selectivity in the case of *Ec*NAGK for future work.

What our simulations do show is that, once the products are released, the enzyme will preferentially remain in the open conformation, ready for a new catalytic cycle. This favors the induced fit mechanism, rather than the conformational selection hypothesis, and agrees with the results of previous work on other enzymes with lid-gated active sites, such as phosphoenolpyruvate carboxykinase.[74] The induced fit mechanism does not rule out the possibility that the apoenzyme samples the closed conformation, but to maximize binding this sampling should be as rare as possible, as the closed conformation cannot bind the substrates. The submillisecond opening of the structure in our simulations, without sampling the closed conformation again, and the lack of a closed apoenzyme crystal structure for *Ec*NAGK, supports the idea that an efficient enzyme with lid-gated active sites should not waste time sampling the closed conformation without substrates.

CONCLUSIONS

In this work we have analyzed the catalytic mechanism of NAG kinase. We have used information from several crystal structures to trace the course of phosphoryl transfer in the catalytic cycle.

Our results show that the phosphoryl transfer is an associative one-step mechanism. We found that the TS of the reaction is more compact (short O−O distance) than reactants and products and that those conformations with more compact reactants also have lower energy barriers for phosphoryl transfer. Rubio and co-workers coined the term conformational compression[16] to indicate this relation between transition state compactness and catalysis and our results introduce energetic considerations to the purely geometric description given by Rubio et al.

The correct alignment of the reactants (reflected in the O−P−O angle) turns out to be also necessary, in addition to compression of the O−O distance. Thus, motions that access conformations that shorten the O−O distance and increase the O−P−O angles tend to enhance the catalytic power of NAGK.

Even changes that occur on a short-time scale generate local conformations with a wide range of energy barriers. The fact that 2X2W snapshots have low energy barriers, despite being in a product conformation, also points to the difficulty of

generating plausible conformations for the truly equilibrated reactants. This is likely to be a challenge for other enzymes where fewer structures are available, or when some substrates are absent. However, it is pertinent to remind the reader that these observations emerge from a diverse, but still limited, set of conformations sampled by the enzyme (20 in total obtained from 4 different X-ray structures). It is the high computational demands of the DFT/MM method, which is required for an accurate description of the electronic structure of the molecular species involved, that makes consideration of a larger set of reaction pathways or the use of PMF calculations prohibitive. Other computational approaches, such as free energy perturbation, can be valuable to enhance the statistical significance of the variety of energy barriers and thus provide a more accurate estimation of the catalytic role of compressive motions. We consider it more appropriate for future work.

The role of water in the active site is also found to be crucial for this enzyme. Water plays different roles. One molecule prevents the formation of a salt bridge between Lys61 and the substrate NAG which overstabilizes the reactants and hinders catalysis, whereas others reduce the cost of product release, by accompanying the unbound products.

The energy barrier of the conformations studied is 28 kJ/mol lower than the apparent experimental energy barrier. This, together with the experimental observation that the 2X2W crystal structure contains the products of the reaction, can be explained by the idea suggested by Rubio and co-workers[16] that lid opening and product release are the rate-limiting step of this enzyme.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Details on the atoms present in the QM region; data on Poisson−Boltzmann calculations; plots showing the active site components; plots showing the energy profiles from different basis sets and QM regions; plots showing PLSR results and plots showing the open−close dynamics with different force fields and structures 1GS5, 1OHA, and 1OH9. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ramon.crehuet@iqac.csic.es.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge financial support from the Ministerio de Innovación y Competitividad (CTQ2009-08223 and CTQ2012-33324) and the Generalitat de Catalunya (2009SGR01472). M.S-M. thanks the Ministerio de Economia y Competitividad for a predoctoral fellowship. E.M. acknowledges a postdoctoral fellowship from the European Union (FP7-PEOPLE-2011-IOF 298976). Part of the calculations described in this work were carried out at the CESCA.

■ REFERENCES

(1) McGeagh, J. D.; Ranaghan, K. E.; Mulholland, A. J. Protein dynamics and enzyme catalysis: Insights from simulations. *Biochim. Biophys. Acta* **2011**, *1814*, 1077−1092.

(2) Eisenmesser, E. Z.; Bosco, D. a.; Akke, M.; Kern, D. Enzyme dynamics during catalysis. *Science* **2002**, *295*, 1520−3.

(3) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. a.; Skalicky, J. J.; Kay, L. E.; Kern, D. Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **2005**, *438*, 117−21.

(4) Doshi, U.; McGowan, L. C.; Ladani, S. T.; Hamelberg, D. Resolving the complex role of enzyme conformational dynamics in catalytic function. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 5699−5704.

(5) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. Electrostatic basis for enzyme catalysis. *Chem. Rev.* **2006**, *106*, 3210−3235.

(6) Pisliakov, A. V.; Cao, J.; Kamerlin, S. C. L.; Warshel, A. Enzyme millisecond conformational dynamics do not catalyze the chemical step. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 17359−64.

(7) Kamerlin, S. C. L.; Warshel, A. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1339−1375.

(8) Bhabha, G.; Lee, J.; Ekiert, D. C.; Gam, J.; Wilson, I. a.; Dyson, H. J.; Benkovic, S. J.; Wright, P. E. A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science* **2011**, *332*, 234−238.

(9) Agarwal, P. K.; Billeter, S. R.; Rajagopalan, P. T. R.; Benkovic, S. J.; Hammes-Schiffer, S. Network of coupled promoting motions in enzyme catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2794−2799.

(10) Saen-Oon, S.; Ghanem, M.; Schramm, V. L.; Schwartz, S. D. Remote mutations and active site dynamics correlate with catalytic properties of purine nucleoside phosphorylase. *Biophys. J.* **2008**, *94*, 4078−4088.

(11) Antoniou, D.; Basner, J.; Núñez, S.; Schwartz, S. D. Computational and theoretical methods to explore the relation between enzyme dynamics and catalysis. *Chem. Rev.* **2006**, *106*, 3170−3187.

(12) Wolf-Watz, M.; Thai, V.; Henzler-Wildman, K.; Hadjipavlou, G.; Eisenmesser, E. Z.; Kern, D. Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat. Struct. Mol. Biol.* **2004**, *11*, 945−9.

(13) Tobi, D.; Bahar, I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 18908−18913.

(14) Karplus, M. Role of conformation transitions in adenylate kinase. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, E71 author reply E72..

(15) Kamerlin, S. C. L.; Warshel, a. Reply to Karplus: Conformational dynamics have no role in the chemical step. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, E72−E72.

(16) Gil-Ortiz, F.; Ramón-Maiques, S.; Fernández-Murga, M. L.; Fita, I.; Rubio, V. Two crystal structures of *Escherichia coli* *N*-acetyl-L-glutamate kinase demonstrate the cycling between open and closed conformations. *J. Mol. Biol.* **2010**, *399*, 476−490.

(17) Gil-Ortiz, F.; Ramón-Maiques, S.; Fita, I.; Rubio, V. 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 AlF^{4-} transition state mimic. *J. Mol. Biol.* **2003**, *331*, 231−244.

(18) Marco-Marín, C.; Ramón-Maiques, S.; Tavárez, S.; Rubio, V. 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.* **2003**, *334*, 459−476.

(19) Ramón-Maiques, S.; Fernández-Murga, M. L.; Gil-Ortiz, F.; Vagin, A.; Fita, I.; Rubio, V. 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.* **2006**, *356*, 695−713.

(20) Ramón-Maiques, S.; Marina, A.; Gil-Ortiz, F.; Fita, I.; Rubio, V. Structure of acetylglutamate kinase, a key enzyme for arginine biosynthesis and a prototype for the amino acid kinase enzyme family, during catalysis. *Structure* **2002**, *10*, 329−42.

(21) Fernández-Murga, M. L.; Rubio, V. 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.* **2008**, *190*, 3018−3025.
(22) Gora, A.; Brezovsky, J.; Damborsky, J. Gates of eEnzymes. *Chem. Rev.* **2013**.
(23) Marcos, E.; Crehuet, R.; Bahar, I. On the conservation of the slow conformational dynamics within the amino acid kinase family: NAGK the paradigm. *PLoS Comput. Biol.* **2010**, *6*, e1000738.
(24) Marcos, E.; Crehuet, R.; Bahar, I. Changes in dynamics upon oligomerization regulate substrate binding and allostery in amino acid kinase family members. *PLoS Comput. Biol.* **2011**, *7*, e1002201.
(25) Field, M. J. The pDynamo program for molecular simulations using hybrid quantum chemical and molecular mechanical potentials. *J. Chem. Theory Comput.* **2008**, *4*, 1151−1161.
(26) Neese, F. *ORCA - an ab initio, density functional and semiempirical program package*, Version 2.6.; University of Bonn: Bonn, Germany, 2008.
(27) 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.
(28) 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.
(29) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(30) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(31) Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90−95.
(32) Laskowski, R. A.; Swindells, M. B. LigPlot+: Multiple ligand−protein interaction diagrams for drug discovery. *J. Chem. Inf. Model.* **2011**, *51*, 2778−2786.
(33) *MATLAB and Statistics Toolbox Release 2012b*; The MathWorks Inc.: Natick, Massachussets, United States, 2012.
(34) Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pK_a values. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 704−721.
(35) Jorgensen, W. L.; Maxwell, D. S.; 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.
(36) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(37) Nam, K.; Cui, Q.; Gao, J.; York, D. M. Specific reaction parametrization of the AM1/d Hamiltonian for phosphoryl transfer reactions: H, O, and P atoms. *J. Chem. Theory Comput.* **2007**, *3*, 486−504.
(38) Marcos, E.; Anglada, J. M.; 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.
(39) 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.
(40) Marcos, E.; Crehuet, R.; Anglada, J. M. Inductive and external electric field effects in pentacoordinated phosphorus compounds. *J. Chem. Theory Comput.* **2008**, *4*, 49−63.
(41) Marcos, E.; Field, M. J.; Crehuet, R. Pentacoordinated phosphorus revisited by high-level QM/MM calculations. *Proteins* **2010**, *78*, 2405−2411.
(42) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037−10041.
(43) Jónsson, H.; Mills, G.; Jacobsen, K. W. *Nudged elastic band method for finding minimum energy paths of transitions*; World Scientific: Singapore, 1998.
(44) 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.
(45) Crehuet, R.; Thomas, A.; Field, M. J. An implementation of the nudged elastic band algorithm and application to the reaction mechanism of HGXPRTase from *Plasmodium falciparum*. *J. Mol. Graph. Model.* **2005**, *24*, 102−10.
(46) Galvan, I. F.; Field, M. J. Improving the efficiency of the NEB reaction path finding algorithm. *J. Comput. Chem.* **2008**, *29*, 139−143.
(47) Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer Verlag: Berlin, Germany, 2002.
(48) Geladi, P.; Kowalski, B. R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(49) *PLS Toolbox v 7.0*. Eigenvector Research: Manson, WA, U.S.A., 2012.
(50) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950−1958.
(51) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Hermans, J. In *Interaction models for water in relation to protein hydration*; Pullman, B., Ed.; D. Reidel Publishing: Dordrecht, The Netherlands, 1981; pp 331−338.
(52) Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. v.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(53) Hess, B. P-LINCS: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* **2007**, *4*, 116−122.
(54) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(55) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(56) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An $N\cdot\log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(57) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. A practical guide to modelling enzyme-catalysed reactions. *Chem. Soc. Rev.* **2012**, 3025−3038.
(58) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. Compound I reactivity defines alkene oxidation selectivity in cytochrome P450cam. *J. Phys. Chem. B* **2010**, *114*, 1156−62.
(59) Claeyssens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. a.; Mulholland, A. J.; Ranaghan, K. E.; Schütz, M.; Thiel, S.; Thiel, W.; Werner, H.-J. High-accuracy computation of reaction barriers in enzymes. *Angew. Chem., Int. Ed.* **2006**, *45*, 6856−6859.
(60) Steinmann, C.; Fedorov, D. G.; Jensen, J. H. Mapping enzymatic catalysis using the effective fragment molecular orbital method: Towards all ab initio biochemistry. *PLoS One* **2013**, *8*, e60602.
(61) Claeyssens, F.; Ranaghan, K. E.; Lawan, N.; Macrae, S. J.; Manby, F. R.; Harvey, J. N.; Mulholland, A. J. Analysis of chorismate mutase catalysis by QM/MM modelling of enzyme-catalysed and uncatalysed reactions. *Org. Biomol. Chem.* **2011**, *9*, 1578−1590.
(62) Ranaghan, K. E.; Ridder, L.; Szefczyk, B.; Sokalski, W. A.; Hermann, J. C.; Mulholland, A. J. Transition state stabilization and substrate strain in enzyme catalysis: ab initio QM/MM modelling of the chorismate mutase reaction. *Org. Biomol. Chem.* **2004**, *2*, 968−980.
(63) Lodola, A.; Sirirak, J.; Fey, N.; Rivara, S.; Mor, M.; Mulholland, A. J. Structural fluctuations in enzyme-catalyzed reactions: determinants of reactivity in fatty acid amide hydrolase from multivariate statistical analysis of quantum mechanics/molecular mechanics paths. *J. Chem. Theory Comput.* **2010**, *6*, 2948−2960.
(64) Lans, I.; Peregrina, J. R. n.; Medina, M.; Garcia-Viloca, M.; González-Lafont, A. n.; Lluch, J. M. Mechanism of the hydride transfer between anabaena Tyr303Ser FNRrd/FNRox and NADP+/H. A combined pre-steady-state kinetic/ensemble-averaged transition-state theory with multidimensional tunneling study. *J. Phys. Chem. B* **2010**, *114*, 3368−3379.

(65) Lans, I.; Medina, M.; Rosta, E.; Hummer, G.; Garcia-Viloca, M.; Lluch, J. M.; González-Lafont, À. Theoretical study of the mechanism of the hydride transfer between ferredoxin−NADP+ reductase and NADP+: The role of Tyr303. *J. Am. Chem. Soc.* **2012**, *134*, 20544−20553.
(66) Chong, I.-G.; Jun, C.-H. Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103−112.
(67) Min, W.; English, B. P.; Luo, G.; Cherayil, B. J.; Kou, S. C.; Xie, X. S. Fluctuating enzymes: lessons from single-molecule studies. *Acc. Chem. Res.* **2005**, *38*, 923−31.
(68) Engelkamp, H.; Hatzakis, N. S.; Hofkens, J.; De Schryver, F. C.; Nolte, R. J.; Rowan, A. E. Do enzymes sleep and work? *Chem. Commun.* **2006**, 935−40.
(69) Lodola, A.; Mor, M.; Zurek, J.; Tarzia, G.; Piomelli, D.; Harvey, J. N.; Mulholland, A. J. Conformational effects in enzyme catalysis: reaction via a high energy conformation in fatty acid amide hydrolase. *Biophys. J.* **2007**, *92*, L20−L22.
(70) Warshel, A. Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *J. Biol. Chem.* **1998**, *273*, 27035−27038.
(71) García-Meseguer, R.; Martí, S.; Ruiz-Pernía, J. J.; Moliner, V.; Tuñón, I. Studying the role of protein dynamics in an S_N2 enzyme reaction using free-energy surfaces and solvent coordinates. *Nat. Chem.* **2013**, *5*, 566−571.
(72) Xiaoxia, L.; Marston, J. P.; Baxter, N. J.; Hounslow, A. M.; Yufen, Z.; Blackburn, G. M.; Cliff, M. J.; Waltho, J. P. Prioritization of charge over geometry in transition state analogues of a dual specificity protein kinase. *J. Am. Chem. Soc.* **2011**, *133*, 3989−3994.
(73) Whittier, S. K.; Hengge, A. C.; Loria, J. P. Conformational motions regulate phosphoryl transfer in related protein tyrosine phosphatases. *Science* **2013**, *341*, 899−903.
(74) Sullivan, S. M.; Holyoak, T. Enzymes with lid-gated active sites must operate by an induced fit mechanism instead of conformational selection. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13829−13834.

4.1.2 Swarms of Trajectories applied to enzyme catalysis

One of the major concerns of enzyme catalysis lies on the description of the reaction path and the associated energy barrier. As we commented in previous sections, enzymes present multidimensional and rough landscapes that difficult the calculation of minimum free energy paths. However, as it has been also highlighted previously, to directly compare with experimental results free energies are needed. One example of this problem is our previous NAGK work, in that only potential energies were obtained preventing a direct comparison with the experimental free energy energy barrier.

Due to this, in close collaboration with Professor Field in Grenoble, we decide to implement the Swarms of Trajectories (SoT) method into the pDynamo library, the library we usually employ to perform QM/MM calculations, aimed to apply it to enzyme catalysis. We wanted to have easily accessible a method to calculate minimum free energy paths and profiles. The SoT method has never been applied to this purpose before. So, we needed to devise its correct implementation to simulate enzyme catalyzed reactions. To check our implementation, we employed the Chorismate Mutase (CM) and Isochorismate Pyruvate Lyase (IPL), as test cases.

The SoT method was developed by Roux and coworkers (*J. Phys. Chem. B, 2008, 112 (11), pp 3432 - 3440*) based on the String method with collective variables. The method was tested and used with a couple of proteic systems but it was never employed to study enzyme catalysis. In fact one of the conclusions of the study in that SoT was first introduced, was that modifications refining formal and practical aspects would be required to use the method with real biomolecular applications. An important problem of the method regarding enzyme catalysis was that it works in a non-inertial, almost overdamped, regime and chemical reactions work on inertial regimes. However, Maragliano and coworkers in a recent paper (*J. Chem. Theory Comput. 2014, 10 (2), pp 524 - 533*) that compared the Sot and the String method with CV, formulated a hypothesis stating that the SoT could work at inertial short time regimes being the CV evolution independent of the dynamics.

Within our implementation we confirmed this hypothesis for the first time, showing that the results of the method do not depend on the dynamical evolution of the system, only depends on the value of the CVs. To do that we simulate our test systems at inertial and non-inertial regimes (changing the dynamical parameters accordingly). The way in that the reaction path is determined does not matter. Furthermore, an important issue of our implementation is that in the way we did it, it is possible to calculate the contribution of each collective variable to the free energy profile.

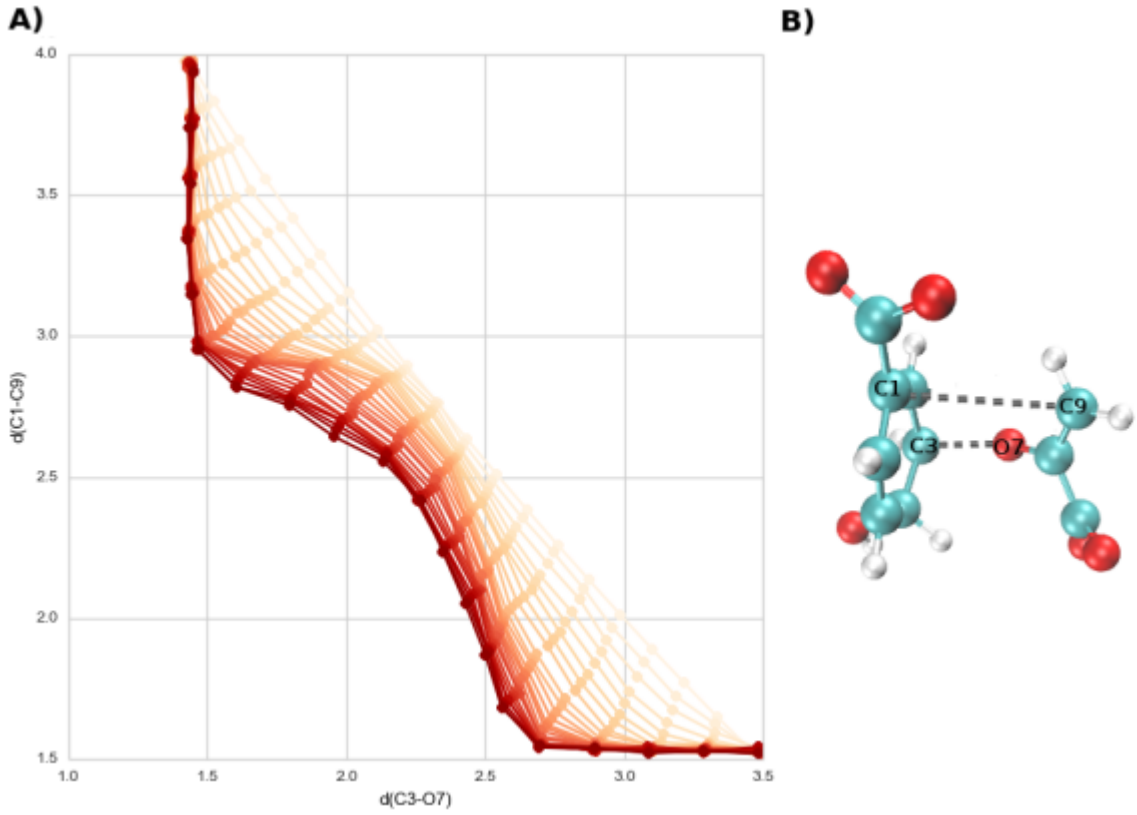


Figure 4.3: A) Evolution of the MFEP CVs calculated with SoT in the CM. The evolution is from light to dark colours. B) Schematic represenattion of the CM. The CVs are represented by dotted lines.

As a method based on CVs, its selection has to be done carefully. Testing our implementation we found that if one CV, important to describe the reaction process, is not taken into account the energy barrier will be underestimated. On the other hand we also found that if one use a CV that does not play a role in the description of the chemical reaction, it remains invariant along the energy path without affecting the energy profile. Furthermore the computational cost is not affected. Thus this leads us to conclude that if there is a CV whose effect describing the reaction process is not clear, is better to add it and then check its contribution.

Summarizing, we have devised the most suitable setup of the SoT method showing that the results do not depend on the dynamical evolution of the CV. Besides we have suggested the use of the Minimum Energy Path (MEP) to accelerate the optimization as well as to ease convergence problems.

THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Enzymatic Minimum Free Energy Path Calculations Using Swarms of Trajectories

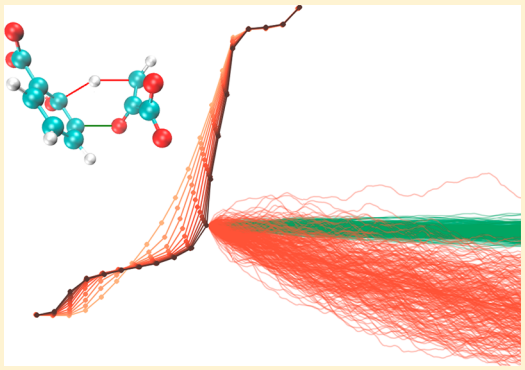
Melchor Sanchez-Martinez,[†] Martin Field,[‡] and Ramon Crehuet*[,†]

[†]Institute of Advanced Chemistry of Catalonia (IQAC), CSIC, Jordi Girona 18-26, 08034, Barcelona, Spain

[‡]Institut de Biologie Structurale (CEA, CNRS UMR5075, Université Joseph Fourier - Grenoble I), 71 Avenue des Martyrs, CS 10090, 38044 Grenoble Cedex 9, France

Supporting Information

ABSTRACT: The development of approaches for simulating rare events in complex molecular systems is a central concern in chemical physics. In recent work, Roux and co-workers proposed a novel, swarms of trajectories (SoT) method for determining the transition paths of such events. It consists of the dynamical refinement on the system's free energy surface of a putative transition path that is parametrized in terms of a set of collective variables (CVs) that are identified as being important for the transition. In this work, we have implemented the SoT method and used it to investigate the catalytic mechanisms of two enzymatic reactions using hybrid QM/MM potentials. Our aim has been to test the performance of SoT for enzyme systems and to devise robust simulation protocols that can be employed in future studies of this type. We identify the conditions under which converged results can be obtained using inertial and Brownian dynamical evolutions of the CVs, show that the inclusion of several CVs can give significant additional insight into the mechanisms of the reactions, and show that the use of minimum energy paths as starting guesses can greatly accelerate path refinement.



INTRODUCTION

The theoretical description of enzymatic mechanisms is based on free energy profiles, and the calculation of these profiles has become an important problem in computational biochemistry.[1−4] The free energy profile describes the chemical mechanism, and the resulting energy barrier allows the estimation of the rate of the process.[5] These profiles are defined along a hypothetical reaction coordinate whose finding is highly nontrivial.

There have been different approaches to the description of these profiles. On the one hand, one can define a set of presumably relevant collective variables (CVs) and calculate a free energy surface (aka potential of mean force) depending on these variables. Once a free energy surface is determined, it is usually projected in two dimensions and visually inspected to determine minimum free energy paths (MFEPs) connecting the different basins. These paths give a one-dimensional representation of the surface, and the value of the free energy along these paths produces a free energy profile (see the Methods section for a mathematical definition).[6] Methods such as adaptive biased force (ABF)[7] sampling, metadynamics,[8,9] and umbrella sampling[10] require a precise choice of a few CVs. As these methods describe the full free energy surface, they scale exponentially with the number of variables and rapidly become impractical due to the computational expense and difficulty of exploring multidimensional energy surfaces. Unfortunately, enzymatic reactions are complex and often need many CVs to be completely described.

On the other hand, the computational burden would be highly reduced if one could directly trace the paths in the free energy surface without having to fully determine that surface. Chain-of-states methods, that include the zero-temperature string[11,12] and nudged elastic band (NEB)[13,14] methods, can directly determine reaction paths. However, in their basic versions, these methods produce only minimum (potential) energy paths (MEPs), as they omit sampling and entropic contributions.[15] Nevertheless, they can be extended or generalized so that the determination of free energies is, in principle, possible.[11,16,17]

In one example of this type, the string method was modified to produce MFEPs by permitting sampling among a set of CVs.[15] In a related development, Roux and co-workers[18] proposed a novel method that employed swarms of trajectories (SoT) to evolve the string and to estimate its average displacement in CV space. Subsequently, they applied it to the study of a large biomolecular transformation.[19] Chemical reactions catalyzed by enzymes, however, have very different

Special Issue: William L. Jorgensen Festschrift

Received: July 2, 2014
Revised: October 1, 2014

   dx.doi.org/10.1021/jp506593t | J. Phys. Chem. B XXXX, XXX, XXX−XXX

dynamics than the latter types of processes, as they take place on the time scale of bond vibrations and involve only small-amplitude motions of a few atoms. All of these effects make the relaxation from the transition state an inertial process, far from the diffusive process that was assumed in the original SoT formulation. This would suggest that SoT is not a good framework to study enzyme catalysis, but in a recent comprehensive study, Maragliano et al. showed that SoT gives the correct converged MFEP *independent* of the dynamics of the system, as long as the time scale chosen to calculate the changes of the CVs is short enough.[20]

Alternative approaches to calculate free energies based on CVs have also been proposed,[21] but their application to large systems such as enzymes is rare. An exception is the work of Tuñon and co-workers who adapted the method of Branduardi et al.[21] to study enzyme catalysis and applied it to the mechanism of isochorismate pyruvate lyase (IPL).[22] This method does not seek to optimize a MFEP but instead calculates the free energy profile associated with the curve followed by the minimum potential energy path.

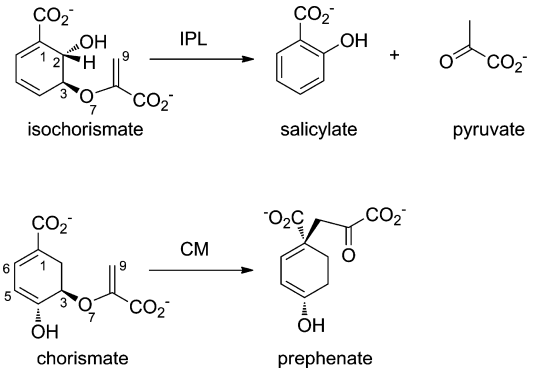
In this work, we have implemented the SoT method in the pDynamo molecular modeling library[23] and applied it to the study of the IPL and chorismate mutase (CM) enzymatic reaction mechanisms. These are both realistic test cases about which much information concerning the mechanisms is already known.[22,24−32] Our aim has been to investigate the performance of SoT in enzymes and to evaluate the conditions under which it works. We show that using several CVs can give significant additional insight into enzymatic mechanisms and also find that large increases in performance can be obtained if one initiates the SoT calculation from MEPs calculated with a hybrid string/NEB chain-of-states method.[4,33,34]

METHODS

Computational Details. All methodological development and simulations, including system setup, were done with the pDynamo[23] program, version 1.8.0. As preliminary work, we implemented a version of the SoT method in pDynamo and ensured that it reproduced the results of tests, including those on the blocked alanine dipeptide, that were published in the original papers.[15,18] Subsequently, to test the validity of our implementation of the SoT method for the investigation of enzymatic reactions, we chose two different enzyme systems, IPL and CM. We studied the reactions using hybrid quantum mechanical (QM) / molecular mechanical (MM) potentials in combination with chain-of-states reaction path calculations. As the QM method in our QM/MM potentials, we employed the AM1[35] semiempirical Hamiltonian. Although the latter is less precise than, say, density functional theory (DFT) methods, it is much less computationally demanding and thus makes possible a thorough analysis of the SoT approach to our test systems. In any case, we intend to apply these higher level potentials in our future studies using the SoT method.

Isochorismate Pyruvate Lyase. IPL transforms isochorismate into salicylate and pyruvate in a pericyclic reaction (Scheme 1). Our simulation model of IPL was derived from the X-ray crystallographic structure with PDB entry 2H9D.[36] This contains the pyruvate-bound IPL from *Pseudomonas aerugionsa* (PchB) with two pyruvate molecules in the active site. The latter were removed and replaced by an isochorismate molecule which represents the reactant state. The positions of hydrogens were then built and the whole system solvated in an orthorhombic water box of dimension 68 × 46 × 36 Å^3, with

Scheme 1. Schemes of the Reactions Catalyzed by Isochorismate Pyruvate Lyase (IPL) and Chorismate Mutase (CM)



an appropriate number of Na^+ counterions added to neutralize the overall charge of the system, giving ∼11000 atoms in total. We used the OPLS/AA[37] force field to describe the protein, together with the TIP3P[38] model for the water solvent, and periodic boundary conditions for the long-range interactions. To generate a suitable starting structure for the simulation, the structure of the solvated protein was first energy minimized, followed by an equilibrating molecular dynamics (MD) simulation at a temperature of 300 K. For the QM/MM simulations, the QM region contained just the substrate, isochorismate, numbering 24 atoms. In the QM/MM calculations, only atoms within 20 Å from the O7 oxygen of the substrate were allowed to move (Figure 1).

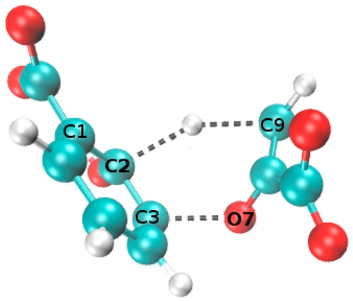


Figure 1. Transition state structure for the IPL transformation. The bonds that are being broken or formed are represented with dotted lines.

Chorismate Mutase. CM catalyzes the Claisen rearrangement from chorismate to pyruvate (Scheme 1). We modeled our CM simulation system from the X-ray crystallographic structure of the *Baciilus subtilis* enzyme with PDB entry 1COM.[39] The latter contained four homotrimers and one prephenate (PRE) of which we retained one homotrimer together with the PRE molecule. The remaining setup was similar to that we employed for the IPL system, except that we used K^+ and Cl^- counterions to neutralize the overall system charge. The final system had ∼17700 atoms with box dimension 60 × 50 × 60 Å^3. In the QM/MM calculations, the QM region contained only the substrate, PRE, numbering 24 atoms, and in the reaction path calculations, only atoms within 12 Å from the C1 carbon of the PRE were allowed to move (Figure 2).

Initial Reaction Paths. Initial paths are required to perform SoT calculations. We employed MEPs that were obtained by

dx.doi.org/10.1021/jp506593t | J. Phys. Chem. B XXXX, XXX, XXX−XXX

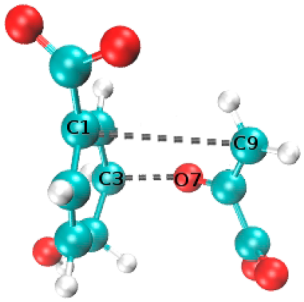


Figure 2. Transition state structure for the CM transformation. The bonds that are being broken or formed are represented with dotted lines.

carrying out chain-of-states reaction path calculations using the hybrid NEB/string method that is implemented in the pDynamo program.[4,33,34] For each of the test cases, we started from a small number of path structures and gradually increased them until the energy profile converged. The number of structures per path depended on the path length and ruggedness, ranging from 19 to 37 in IPL and from 18 to 28 in CM.

The MFEP and SoT Methods. Here we present a brief summary of the theoretical background behind the MFEP and SoT calculations. Readers are referred to the original papers for a full discussion.[15,18,20]

The Minimum Free Energy Path. Consider a system described by the Cartesian coordinates $\mathbf{x} \in R^n$ with a standard equilibrium distribution

$$p(\mathbf{x}) = Z^{-1} e^{-\beta V(\mathbf{x})} \tag{1}$$

where $\beta = 1/k_B T$ is the inverse temperature, $V(\mathbf{x})$ is the potential energy, and $Z = \int d\mathbf{x} e^{-\beta V(\mathbf{x})}$. We assume that there are no constraints in the system, and that the part of the density arising from the momenta has been integrated out.[15] We now introduce M CVs that are functions of $\mathbf{x}$ and that can distinguish distinct reacting configurations of the system

$$\tilde{\mathbf{z}}(\mathbf{x}) = \{\tilde{z}_1(\mathbf{x}), \tilde{z}_2(\mathbf{x}), ..., \tilde{z}_M(\mathbf{x})\} \tag{2}$$

The free energy, also known as the potential of mean force, associated with $\tilde{z}(\mathbf{x})$ is a function that depends on $z = (z_1, ..., z_M)$ and is calculated as

$$F(\mathbf{z}) = -k_B T \ln(Z^{-1} \int_{R^n} e^{-\beta V(\mathbf{x})} \delta(z_1 - \tilde{z}_1(\mathbf{x})), ..., \delta(z_M - \tilde{z}_M(\mathbf{x}))\, d\mathbf{x}) \tag{3}$$

This is the M-dimensional free energy surface described in the Introduction. A MFEP is defined on a free energy surface of eq 3 in the same way as a MEP is defined on a potential energy surface. Thus, it is the path between two minima on the surface such that the following condition holds:

$$[\mathbf{M}(\mathbf{z}) \cdot \nabla F(\mathbf{z})]^{\perp} = 0 \tag{4}$$

In this equation, $\nabla F(\mathbf{z})$ is the gradient of the free energy, $\mathbf{M}(\mathbf{z})$ is a metric tensor that accounts for the curvilinear nature of the CVs, and $\perp$ indicates projection in the direction perpendicular to the curve. Full details, including an expression for the metric tensor, may be found in the work of Maragliano et al.[15,20] We note that the metric tensor also appears when calculating minimum potential energy paths in terms of CVs, so it is not exclusive to the free energy. The main advantage of the SoT method is that this tensor is never calculated, as its influence is implicitly taken into account with the swarms of trajectories (see below).

As long as the energy barriers are high compared to the thermal energy and that the reaction trajectories cluster around reaction tubes, the MFEP corresponds to the path with highest likelihood, and its maximum corresponds to structures that have the same chance of falling into either the reactant or product basins.[15]

The SoT Reaction Path: Evolution and Reparameterization. Equation 3 can be used to determine the MFEP if a convenient representation of the path as a function of the variables, z, is available. In the SoT formalism, this is done by parametrizing the path as $z(\alpha)$, with $\alpha \in [0, 1]$, and where $\alpha = 0$ represents the initial (reactant) state and $\alpha = 1$ the final (product) conformation.[15,18] It is then assumed that the CVs evolve according to a noninertial Brownian dynamics over some time step, Δt, according to

$$z_i(\Delta t) = z_i(0) + \sum_j \left(-\beta D_{ij}[\mathbf{z}(0)] \partial_j F[\mathbf{z}(0)] + \partial_{z_j} D_{ij}[\mathbf{z}(0)]\right)\Delta t + R_i(0) \tag{5}$$

where D_{ij} is the diffusion tensor, which is equal to $k_\beta T \mathbf{M}(\mathbf{z})$ and $R_i(t)$ is a Gaussian thermal noise with a mean of zero. Once eq 5 has been defined, it can be employed to locate the most probable transition path,[15,18] which is the path such that a system anywhere along it will have the highest probability of remaining on it as it evolves. This is so because the most probable value of the Gaussian noise is zero. The key finding of ref 20 was to prove that when δT is small eq 5 also defines a path that satisfies

$$[-\mathbf{M}(\mathbf{z}) \nabla F(\mathbf{z}) + k_B T \nabla M(\mathbf{z})]^{\perp} = 0 \tag{6}$$

The extra term in this equation compared to eq 3 defining the MFEP was shown to be negligible (at least for the molecular system studied),[22] and from now on, we will consider that the SoT method converges to the MFEP. The equivalence of the SoT method—originally defined to locate the most probable transition paths—and the MFEP can be understood and expected because the MFEP corresponds to the path with the maximum likelihood for a system described with a given set of CVs.[15]

To evolve an initial path toward the MPTP, an approximation to eq 5 is needed. A way to accomplish this is using the so-called average drift[40] (or average displacement) evaluated from an ensemble of unbiased trajectories of length Δt initiated from each image of the path

$$\overline{\Delta z_i(\Delta t)} = \overline{z_i(\Delta t) - z_i(0)} \equiv \sum_j \left(-\beta D_{ij}[\mathbf{z}(0)] \partial_j F[\mathbf{z}(0)] + \partial_{z_j} D_{ij}[\mathbf{z}(0)]\right)\Delta t \tag{7}$$

where the thermal noise in eq 5 is averaged to zero. Thus, by calculating how the CVs evolve as a function of Δt, we need not calculate the terms on the right-hand side of eq 7.

An important insight of the string method is that of reparameterization,[11,12,15,16] which consists of the imposition of a constraint, normally a Euclidean distance,[15] between neighboring path images after each iteration. In practice, this is done by interpolating a curve through the path image structures and then redistributing them along the interpolated path. This is essential because it avoids the problem of the path

dx.doi.org/10.1021/jp506593t | J. Phys. Chem. B XXXX, XXX, XXX−XXX

images congregating in regions of low free energy after repeated application of eq 7.

Implementation of the SoT Method. Implementation of the SoT method is quite straightforward and consists of the following steps:

(i) Generate a path of N images with M CVs that describes the reacting system.

(ii) Perform thermalized molecular dynamics simulation for each image with the values of the CVs for each image restrained about their starting (reference) values from the preceding step—either (i) or (v). This step consists of a short equilibrium trajectory followed by the generation of a larger production trajectory for each image. We employed Langevin dynamics for these simulations using a collision frequency of 25 ps^{-1}. The CVs were restrained using potentials of harmonic form, with force constants of 8000 kJ mol^{-1} $Å^{-2}$, when a single distance was being restrained, and of 4000 kJ mol^{-1} $Å^{-2}$ for the sum or difference of two distances. In the case of IPL, we also tested a larger single-distance force constant of 12000 kJ mol^{-1} $Å^{-2}$ but obtained the same results as with the smaller value (data not shown).

(iii) Run multiple short unbiased trajectories for each image using configurations from the trajectories generated in step (ii). We explored different lengths and types of dynamics for these unbiased trajectories, the details of which are discussed later.

(iv) Calculate the average displacements of the CVs for each image arising from the unbiased trajectories using eq 7.

(v) Determine if the differences between the current average displacements and those of the previous iteration fall below a certain tolerance level. If so, convergence of the SoT calculation has been achieved and the simulation stops. If not, the path is reparameterized by ensuring that the images are redistributed in CV space and the simulation returns to step (ii).

Free Energy Calculations. As we discussed in the Introduction, the advantage of finding a MFEP is that we can directly calculate a free energy profile from it. Thanks to the metric tensor present in eq 4 and using eq 3, we calculate the free energy profile from[15]

$$F(\mathbf{z}(\alpha)) - F(\mathbf{z}(0)) = \int_0^{\alpha} \sum_{i=0}^{M} \frac{\mathrm{d}z_i(\alpha')}{\mathrm{d}\alpha'} \frac{\partial F(\mathbf{z}(\alpha'))}{\partial z_i} \,\mathrm{d}\alpha' \tag{8}$$

where M is the number of CVs and α is a scalar that parametrizes the path curve. In our implementation, we have parametrized $z(\alpha)$ as a cubic spline which means that we can calculate its derivatives with respect to α analytically. The derivatives of the free energy with respect to the CVs are obtained from the constrained dynamics simulations of step (ii) by averaging over the constraint forces applied to each of the CVs. Once these averages have been determined, we also parametrize them with a spline so that the integral of eq 8 and, hence, the free energy profile can be evaluated with an arbitrary number of points. We calculated confidence intervals for the profiles using a bootstrap method in which 1000 resamples of the raw data were generated for each constrained dynamics.

Evolution of the SoT and Dynamics of the Unbiased Trajectories. The free energy is a thermodynamic property of a system that does not depend on its dynamical evolution (eq 8). However, the definition of the MFEP is based on the evolution of the CVs as a function of a time increment (eq 7).

$$\begin{aligned}\tilde{\mathbf{z}}(\mathbf{x}(\Delta t)) = \tilde{\mathbf{z}}(\mathbf{x}(0)) &+ \Delta t \sum_i v_i \frac{\partial \tilde{z}_\alpha(\mathbf{x}(0))}{\partial x_i} \\ &+ \frac{1}{2}\Delta t^2 \sum_i \frac{1}{m_i}\left(-\frac{\partial V}{\partial x_i} - \gamma_i v_i\right)\frac{\partial \tilde{z}_\alpha(\mathbf{x}(0))}{\partial x_i} \\ &+ \frac{1}{2}\Delta t^2 \sum_{i,j} v_i v_j \frac{\partial^2 \tilde{z}_\alpha(\mathbf{x}(0))}{\partial x_i \partial x_j} + O(\Delta t^3)\end{aligned} \tag{9}$$

This formula was deduced assuming that z evolves in the Brownian regime.[18] Later, Maragliano et al. showed that this formula is valid irrespective of the dynamics of the system,[20] as long as it is in the limit of short-time evolution. In that regime, the evolution of the CVs depends on Δt^2 because the velocities average to zero. This result is based on Langevin dynamics for a system (which include inertial dynamics for zero friction). However, in the overdamped regime (Brownian dynamics), the velocity does not average to zero but is proportional to the force:

$$\langle v_i \rangle = \left\langle -\frac{\partial V}{\gamma_i \partial x_i} \right\rangle \tag{10}$$

where γ_i is the friction coefficient. Thus, in a Brownian, noninertial regime, the evolution of the CV is linear with time.

For the IPL system, we have explored three collision frequencies for the evolution of the unbiased trajectories. A very low collision frequency of 25 ps^{-1}, close to the inertial regime, a high collision frequency of 2500 ps^{-1}, close to the Brownian regime, and an intermediate frequency of 250 ps^{-1}.

RESULTS AND DISCUSSION

Minimum Energy Path for IPL. We calculated MEPs with different numbers of structures for IPL, and these are shown in Figure S1 (Supporting Information). The potential energy barrier is high and narrow, which needs a fairly high number of points to be well-defined. The barrier height is 38 kcal/mol, which agrees with the value obtained by Tuñon and co-workers.[22]

The MEPs in Figure S1 (Supporting Information) are functions of the coordinates of all the movable atoms in the system. This is wasteful, as an important part of the path is devoted to describe relaxations that do not involve any energy change or motions relevant to reaction. Instead, it is advantageous to reparameterize these paths in terms of pertinent CVs, as this permits a more compact description and the use of less points because the transformation is better defined. IPL catalyzes the transformation of isochorismate into salycilate and pyruvate in a one-step process in which a proton is transferred from C2 to C9 and the C3−O7 bond is cleaved (Scheme 1, Figure 1, and Figure S2, Supporting Information). In this section, we have chosen the three most obvious CVs, namely, the C2−H, C9−H, and C3−O7 distances. In later sections, we explore the effects of using different sets of reaction coordinates. Figure S2 (Supporting Information) shows the energy profile for the reparameterized path and the evolution of these three CVs along it. It can be seen that there is an error of approximately 2.5 kcal/mol in the barrier height for the path with 19 points in comparison to that of 37 points, whereas the latter is in almost exact agreement with the 73-point barrier of Figure S1 (Supporting Information).

We started the SoT simulations from the reparameterized MEP. This has two advantages. First, the MEP is much closer

dx.doi.org/10.1021/jp506593t | J. Phys. Chem. B XXXX, XXX, XXX−XXX

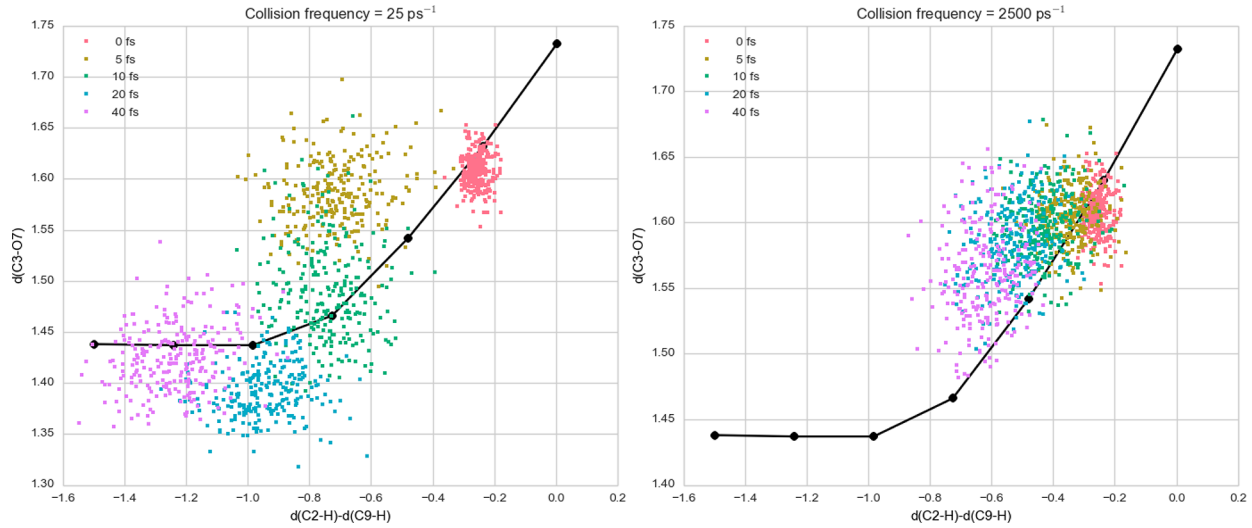


Figure 3. Positions of 250 unbiased trajectories after different simulation times for two different collision frequencies using the IPL system. For the low collision frequency (left), structures are almost reactant-like after only 40 fs of simulation. The MEP structures (used as a starting curve for the evolution of the MFEP) are plotted as black dots.

to the MFEP than any arbitrary initial guess. And second, we avoid the generation of strained geometries that could result in convergence problems. In fact, for this system, we wanted to compare convergence starting from a guess obtained by linear interpolation between the reactant and product structures. This, however, proved impossible, as sampling one of the frames of the linear guess resulted in a proton transfer from C2 to C1, which is both unrealistic and precluded any further optimization of the path.

As a final remark, we note that, around the barrier, the values of the three CVs change smoothly along the MEP with no oscillations. We expect the MFEP to display similar behavior, although oscillations have been observed in free energy paths based on the definition of Branduardi et al.[21] for many combinations of the algorithm's parameters.[22]

The Dynamics of the IPL System. Figure 3 shows that, when the friction is low, the CVs relax very fast, whereas, with high friction, as expected, the evolution of the system is slower. Thus, even though the ensemble average of eq 9 is not friction dependent,[20] the validity of the expression does depend on the friction, as the limiting case of eq 10 shows. Even with high frictions, the trajectories do not fall along the MEP, which indicates that the MEP and the MFEP will be different, as the results in the following section show.

It is clear that with a collision frequency of 25 ps^{-1} we have to take a much shorter time increment, Δt, than with a collision frequency of 2500 ps^{-1}. A short Δt is important not only to remain in the quadratic regime but also to remain in a region where the curvature of the underlying free energy surface is negligible compared to the evolution of the CV. Considering that at low frequency the CVs relax to the reactant values in less than 100 fs, this time has to be very short. At these short time scales, the evolution of the CVs cannot be considered Brownian or diffusive. Figure 4 shows that inertial oscillations remain after 100 fs. Therefore, it is obvious that one cannot find a Δt where the CVs evolve in a Brownian regime and, at the same time, the curvature of the free energy surface is not apparent. To alleviate this, we have two possibilities, both of which we explore. First, increase the friction coefficient in an artificial manner, or second, remain in an inertial regime using a short Δt. Although Vanden-Eijnden showed that the validity of the SoT evolution

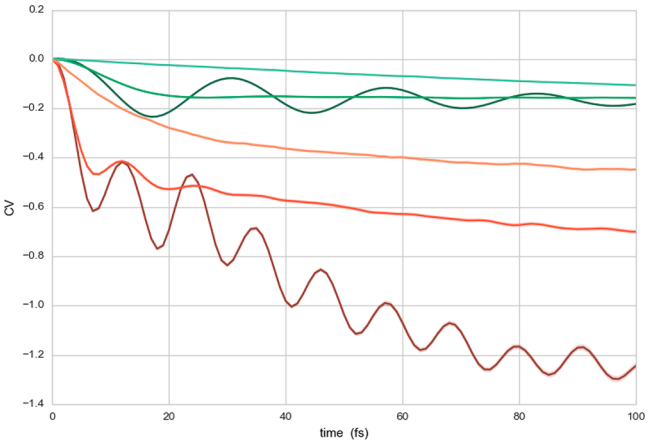


Figure 4. Time evolution of two CVs with different friction coefficients for the IPL system: d(C3−O7) with friction coefficients of 25, 250, and 2500 ps^{-1} (green from dark to light) and d(C2−H) − d(C9−H) with coefficients of 25, 250, and 2500 ps^{-1} (orange, from dark to light). Although we plot the difference of d(C2−H) and d(C9−H), they were treated as independent variables.

does not require diffusive dynamics, in the numerical example he used, the situation was diffusion-like (see Figure 3 in ref 20), and thus, our work is the first that studies SoT for inertial systems.

Figure 4 shows the values of the CVs as a function of time after averaging over 250 trajectories. It can be seen that the fast evolution of the CVs is also linked to an inertial evolution, as oscillations for the low collision frequency remain even after averaging. These oscillations are almost absent when the collision frequency is 250 ps^{-1} and disappear completely for a frequency of 2500 ps^{-1}. These results confirm that our simulations with different collision frequencies cover both the inertial and Brownian regimes.

To confirm these findings, Figure 5 shows the initial evolution of the CVs in the low and high collision frequency cases. As expected, at low friction, the evolution is quadratic for short time scales, so that for times <20 fs the dominant term in eq 9 is quadratic. By contrast, at high friction, the evolution is

dx.doi.org/10.1021/jp506593t | J. Phys. Chem. B XXXX, XXX, XXX−XXX

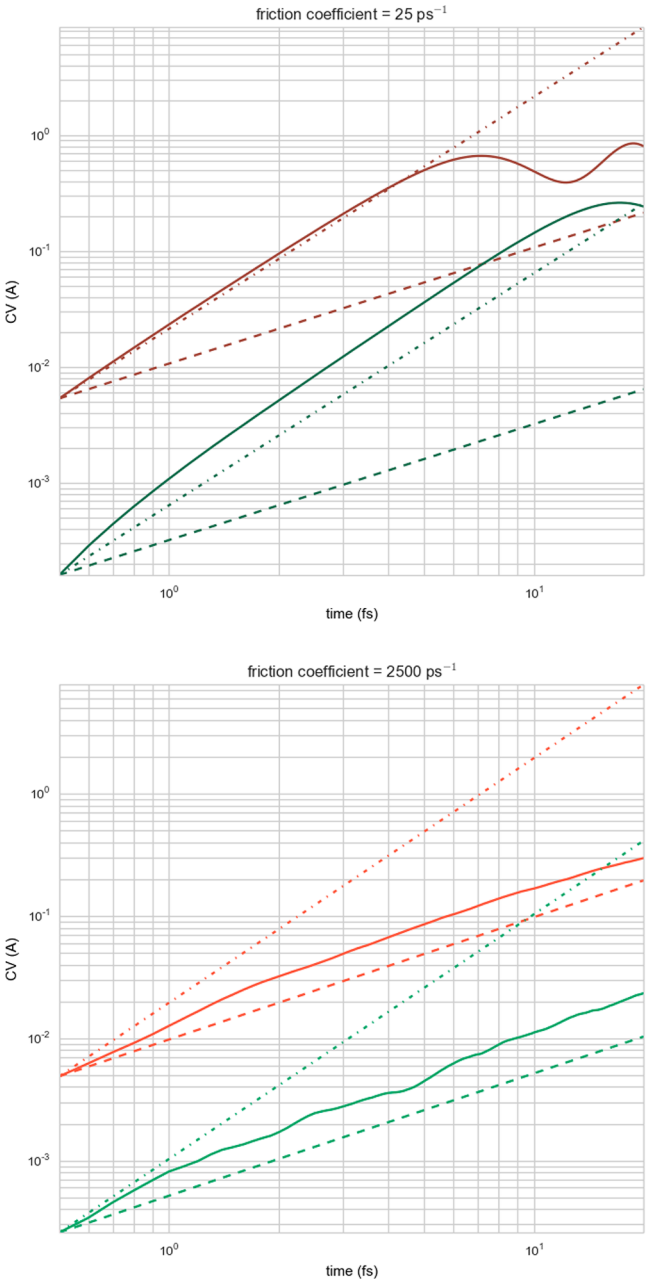


Figure 5. Short time evolution of the CVs for the IPL system (log scale) for the lowest and highest collision frequency dynamics. The color code is the same as in Figure 4. The dashed line corresponds to a linear regime and the dashed−dotted line to a quadratic regime.

initially quadratic but reverts quickly to a near linear form, as expected for Brownian motion.

On the basis of these results, we decided to try two different evolutions of the system. For the low collision frequency of 25 ps^{-1}, the unbiased trajectories were performed with 10 steps and a time step of 0.1 fs, resulting in Δt = 1 fs which assured we remained in the quadratic regime.

Because the change in **z** was small, we scaled $\langle \Delta z(\Delta t) \rangle_{\tilde{z}(\mathbf{x}(0))=z}$ in eq 7 by a factor of 4 so as to obtain similar displacements as those in the Brownian regime. The second setting corresponded to the Brownian regime, with a collision frequency of 2500 ps^{-1}, for which we performed 10 steps with a time step of 1 fs, which resulted in Δt = 10 fs.

Minimum Free Energy Paths. Figure 6 shows the evolution of the MFEPs using three CVs and the two

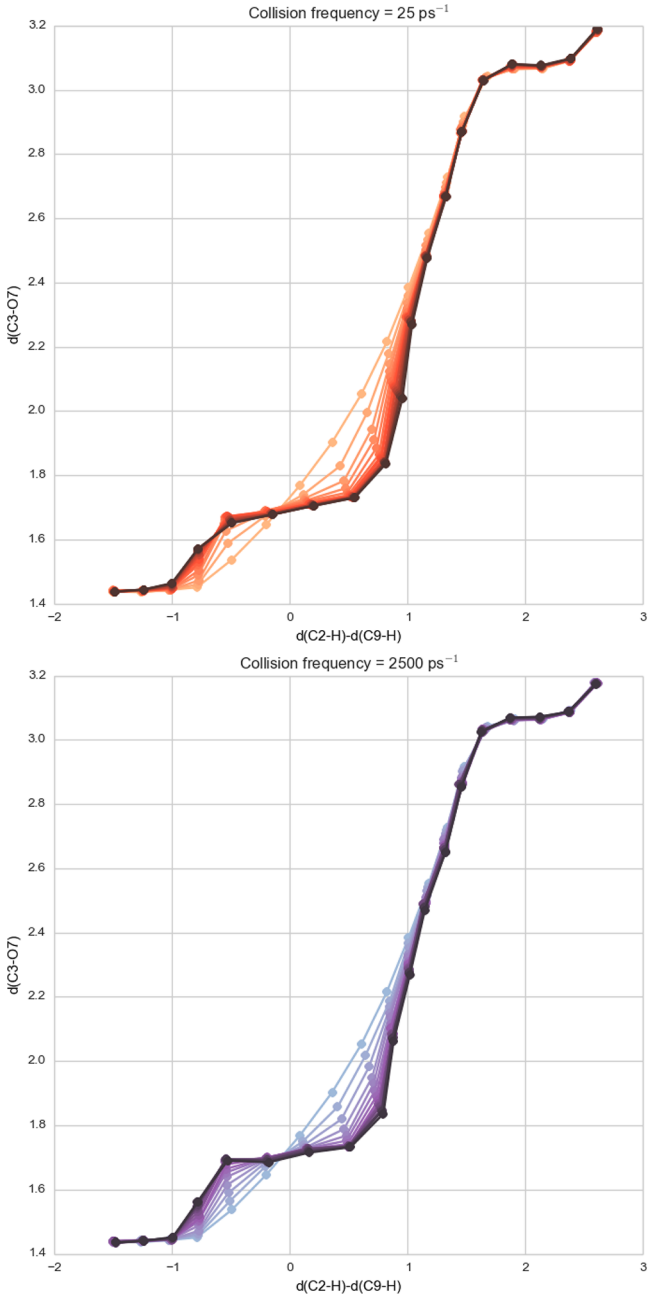


Figure 6. Evolution of the MFEP CVs for the IPL system calculated with SoT and in two different dynamical regimes (low friction, top; high friction, bottom). In both cases, the evolution is from light to dark colors. Although we plot the difference of d(C2−H) and d(C9−H), they were both treated as independent variables.

dynamical regimes described in the previous section. One can see that the MFEP is close to the MEP and that the evolution for both dynamical regimes is also very similar. We plot the difference of two of the CVs because of the difficulty of plotting three CVs separately—nevertheless, in the simulations, they were treated independently. The values of the CVs along the paths are compared in Figure 7. These paths are slightly different from the ones found by Tuñon and co-workers,[22] as ours are smoother and closer to the MEP. In addition, ours

dx.doi.org/10.1021/jp506593t | J. Phys. Chem. B XXXX, XXX, XXX−XXX

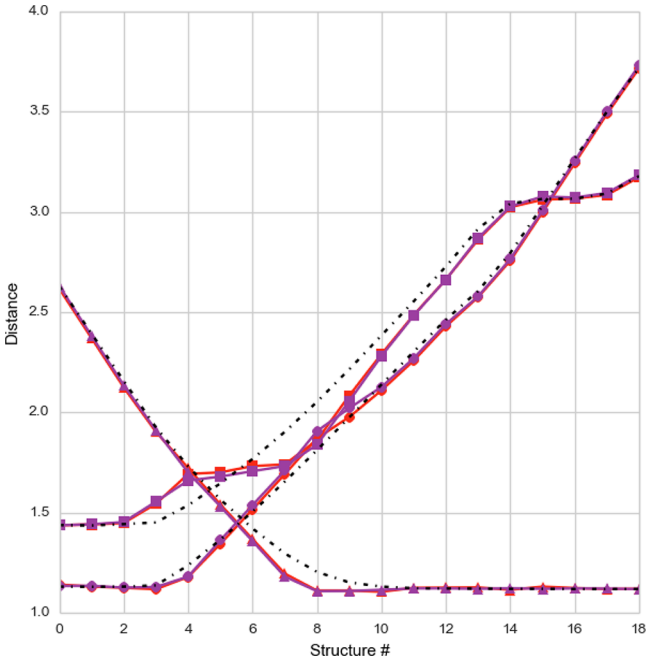


Figure 7. Values of the three CVs for the IPL system in the final path (same color code as Figure 6). The dashed black values are the MEP values. d(C3−O7), squares; d(C2−H), circles; d(C9−H), triangles.

seem to better characterize the products as they extend further and we find that the products have a larger C2−H distance than C3−O7 distance.

As the final paths are similar in both regimes, one expects that the free energies will also be similar. The free energy calculation is independent of the way the path is calculated, and is determined uniquely by the values of the CVs. As detailed in the Methods, we calculate the mean force contribution to the free energy integral with the same settings and the same underlying dynamics. The resulting profiles are shown in Figure 8, from which it can be seen that the barrier increases slightly during path optimization. As we start from the MEP, this seems a reasonable result, because the MFEP will have a better transition vector and thus a higher free energy.[15] Although Tuñon and co-workers find a free energy barrier lower than the potential energy barrier, our free energy barrier is above the potential energy barriers. In both cases, the differences are only 1−2 kcal/mol and thus the discrepancies are small. There are two reasons to expect a higher free energy barrier than a potential energy barrier. First, the interactions with the transition state are stronger than with reactants or products which will tend to rigidify the ensemble of transition state structures compared to reactants or products,[24] although this effect is often small.[41] Second, the transition state of this reaction is a cyclic species that is more constrained than either reactants or products. Figure S4 (Supporting Information) plots the confidence intervals of the final profile of Figure 8 (bottom). We can see that the oscillations of the path and the confidence interval are of the same order. The errors tend to accumulate at the end of the path because the profile is calculated by integrating the mean force, via eq 8.

Number of Degrees of Freedom. An important difference between the MEP and the MFEP is that the MFEP involves sampling along all the nonconstrained degrees of freedom. Some sets of reaction coordinates that are adequate for MEPs are not sufficient for MFEPs, as they do not include

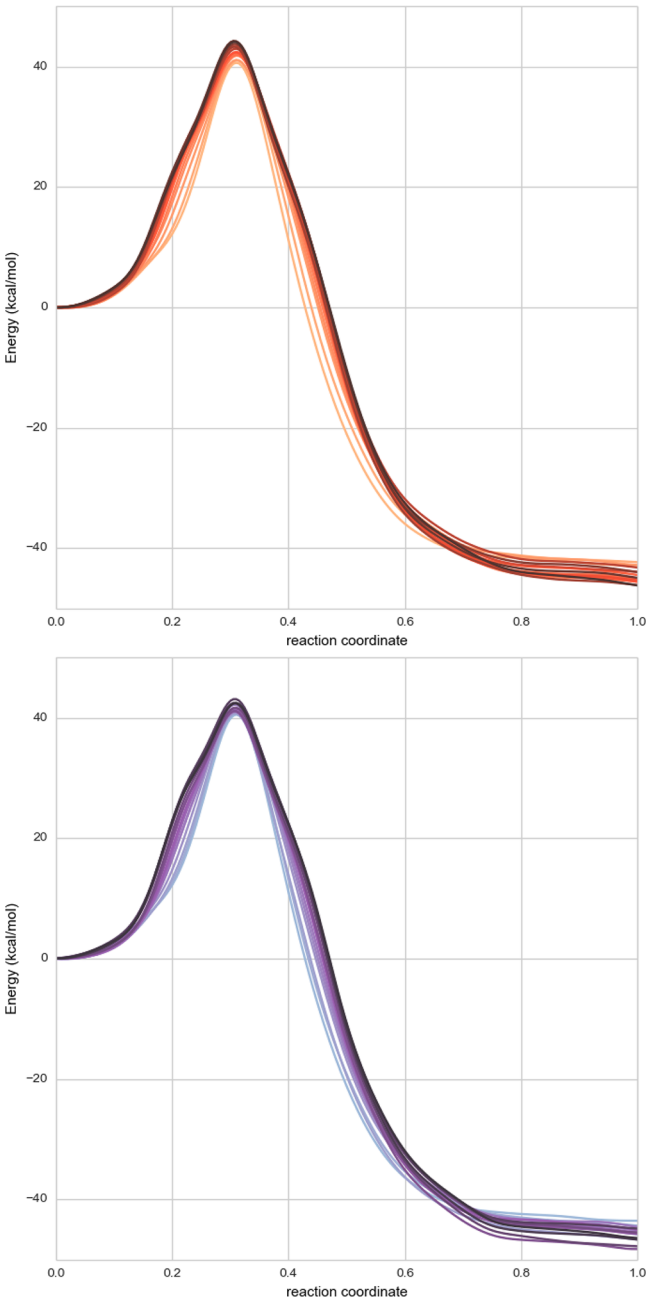


Figure 8. Evolution of the free energy profiles calculated from the MFEP for the IPL system calculated with SoT and in two different dynamical regimes (same color code as Figure 6).

all the variables involved in the transition vector at the transition state hypersurface. When this is the case, the free-energy barrier is too low. Thus, finding a lower free energy barrier does not necessarily indicate a more favorable path; it could also be due to a poor choice of reaction coordinates.[15,42,43] The use of SoT frees us from the use of a small number of degrees of freedom to determine the free energy profile. This is in contrast to more traditional methods that calculate free energies as a function of only one or two coordinates. Employing more coordinates in these schemes involves an exponential increase in computational cost and also produces surfaces in more than two dimensions, from which the extraction of the MFEP is nontrivial because gradients of these surfaces are not available.

dx.doi.org/10.1021/jp506593t | J. Phys. Chem. B XXXX, XXX, XXX−XXX

A typical approximation that is made when an atom is transferred is to use the difference of bond distances between donor and acceptor atoms, as was done in previous work on the IPL system.[25] We recalculated the IPL MFEP with two CVs defined as $z_1 = d(\mathrm{C2-H}) - d(\mathrm{C9-H})$ and $z_2 = d(\mathrm{C3-O7})$. The resulting path and its evolution is shown in Figure 9, and is similar to the three-CV results plotted in Figure 6, thereby confirming that this is a reasonable simplification in this system. The MFEP (Figure 10) gives also a barrier equal to the one obtained with three CVs.

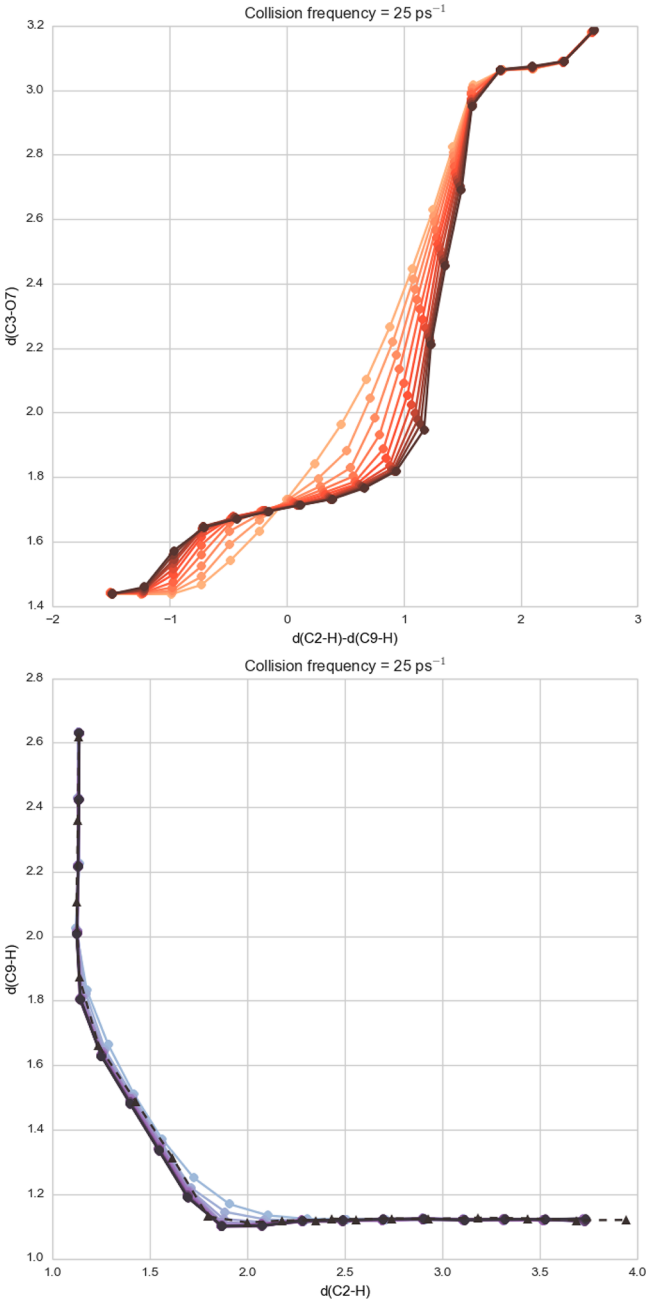


Figure 9. Evolution of the CVs when using two CVs for the IPL system. Top: a difference of distances, $d(\mathrm{C2-H}) - d(\mathrm{C9-H})$, and a distance, $d(\mathrm{C3-O7})$. Bottom: the distances $d(\mathrm{C2-H})$ and $d(\mathrm{C9-H})$ but with an apparently important CV, $d(\mathrm{C3-O7})$, neglected. The optimized values of $d(\mathrm{C2-H})$ and $d(\mathrm{C9-H})$ when using the full three CVs are plotted with red triangles.

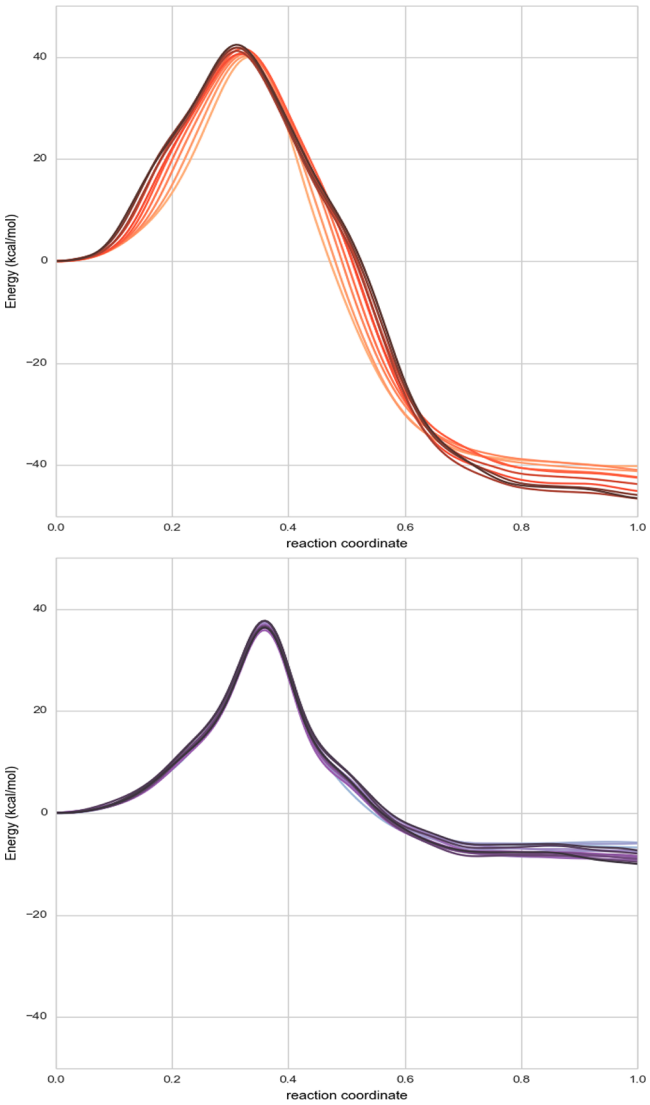


Figure 10. Free energy profiles calculated from the MFEP when using two CVs for the IPL system. Top: $d(\mathrm{C2-H}) - d(\mathrm{C9-H})$ and $d(\mathrm{C3-O7})$ correctly represent the transition vector, and thus give a correct energy profile. Bottom: $d(\mathrm{C2-H})$ and $d(\mathrm{C9-H})$. The relevant $d(\mathrm{C3-O7})$ CV is underestimated, so the free energy barrier is underestimated.

As a next test case, we disregard $d(\mathrm{C3-O7})$ and use only $d(\mathrm{C2-H})$ and $d(\mathrm{C9-H})$ as our CVs. Although unrealistic here, this is a situation that could arise because we inadvertently miss a relevant CV or because we do a traditional free energy calculation method and cannot afford to include one more reaction coordinate. Figure 9 shows that the evolution of the path is correct, and that both distances end with values equivalent to those of the three-CV optimized path. In that sense, the path is correctly described. However, if the missing reaction coordinate is necessary to describe the transition vector—as we expect—this will show up in the resulting free energy profile, which is precisely what Figure 10 shows. When we miss a relevant CV, the free energy barrier is lower than expected. This missing CV can be found with an analysis of the committor probability at the top of the barrier,[15,42] but this is an expensive calculation. When in doubt, one can always include an extra CV to the total set of CVs at no extra cost, and if that CV is not relevant, the results will not be affected. In the

dx.doi.org/10.1021/jp506593t | J. Phys. Chem. B XXXX, XXX, XXX−XXX

CM sections that follow, we test the case of including an irrelevant CV, and show that this is indeed the case.

Lowering the Cost of the SoT. The systems on which SoT has been used so far[18,19] have energies that are computationally fast to calculate, and thus, there has been little effort in analyzing the performance of SoT when sampling is expensive. For QM/MM calculation, the cost of each calculation is considerable when using semiempirical methods, as we do here, and would be much higher if DFT or other *ab initio* methods were to be employed. In the simulations that we have reported, we performed 250 × 10 = 2500 steps for the unbiased dynamics and 1000 + 5000 = 6000 steps for the constrained dynamics. We have also tried different numbers of molecular dynamics steps that are displayed in Table 1.

Table 1. Different Settings Used in the SoT Simulations[a]

equilibration steps	sampling steps	unbiased steps
1000	5000	250 × 10
1000	1000	250 × 10
500	500	250 × 10
500	500	250 × 5
200	200	100 × 5
100	100	50 × 5

[a]As the time step is 1 fs, these values also correspond to the total time length span of the simulations in fs.

We were surprised to find that even the least expensive settings gave a good convergence of the path, as is shown in Figure S3 (Supporting Information). The free energy profiles had errors of several kcal/mol for the total exothermicity but had energy barriers within 1 kcal/mol of the converged value. In any case, an optimized MFEP can always be recalculated with larger equilibration and sampling times in a final single iteration. These results suggest that SoT would be a good approach for calculating MFEPs in enzymes using DFT methods, as one can limit the number of steps that are required. In addition, the method is readily parallelizable, as the calculations for each point along the path are independent.

Chorismate Mutase. CM catalyzes a Claisen rearrangement from chorismate to prephanate (Scheme 1). It is probably the most studied enzyme with QM/MM methods,[24,27−32,44] but we will use it to explore two aspects of the SoT method, namely, the convergence of SoT when starting far from the MFEP and how linear combinations of CVs behave.

We first compare the evolution of the MFEP using two different sets of CVs and starting from an initial linear guess for the pathway structures in which intermediate images are linearly interpolated from the reactant and product structures. We note that there were no convergence problems using a linear guess for this system in contrast to the problems we experienced with the equivalent IPL simulations. The first CV set used $z_1 = d(\mathrm{C1-C9})$ and $z_2 = d(\mathrm{C3-O7})$ (see Figure 2). The second used the sum and difference of these two CVs: $z_1 = d(\mathrm{C1-C9}) - d(\mathrm{C3-O7})$ and $z_2 = d(\mathrm{C1-C9}) + d(\mathrm{C3-O7})$. Because they describe exactly the same subspace, they should produce the same free energy profile and they should have the same evolution. Figure 11 shows that this is indeed the case. Although previous work has shown that $z_1 = d(\mathrm{C1-C9}) - d(\mathrm{C3-O7})$ is a sufficient CV to obtain a good free energy profile, this information is usually only known *a posteriori* after the transition state structures or MEP have been determined and free energy calculations carried out. Being able to include both distances as independent CVs from the start gives the SoT approach a great deal of flexibility.

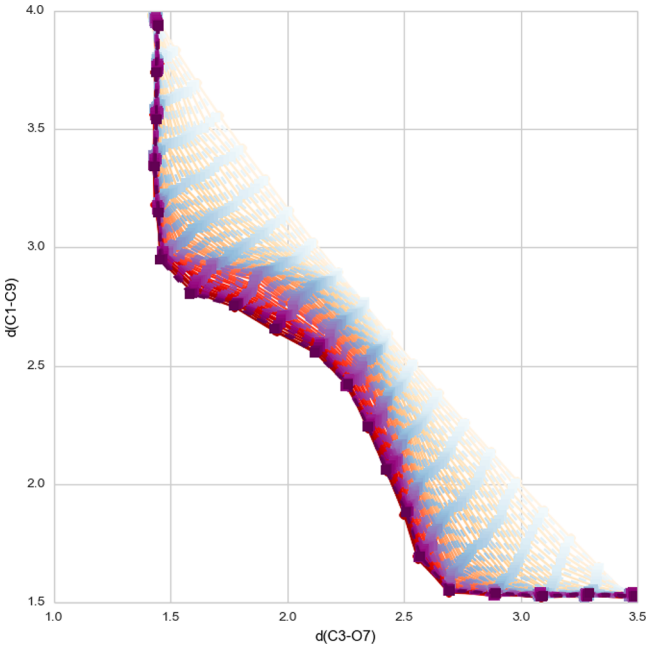


Figure 11. Evolution of two distances when using two different sets of CVs for the CM system. Red (light to dark), the $d(\mathrm{C3-O7})$ and $d(\mathrm{C1-C})$ set; purple (light to dark), the $d(\mathrm{C3-O7}) - d(\mathrm{C1-C9})$ and $d(\mathrm{C3-O7}) + d(\mathrm{C1-C9})$ set. One can see that the evolution and final path shape for both sets of CVs are essentially equivalent.

The calculation of the free energy from eq 8 gives the contribution to the profile as a sum for each of the CVs. In the previous section, we indicated that the use of irrelevant CVs did not affect the calculation of the MFEP or the profile. We have calculated the profile for the CM reaction using the three CVs, $z_1 = d(\mathrm{C1-C9})$, $z_2 = d(\mathrm{C3-O7})$, and $z_3 = d(\mathrm{C5-C6})$. Figure 12 indeed shows that z_3 does not affect the free energy profile, as was to be expected since this distance is not involved in the reaction. We note that the cost of including an extra CV is essentially zero, both in the dynamics and in the free energy calculation. It is possible, though, that extra CVs could slow convergence to the final MFEP, although starting with a

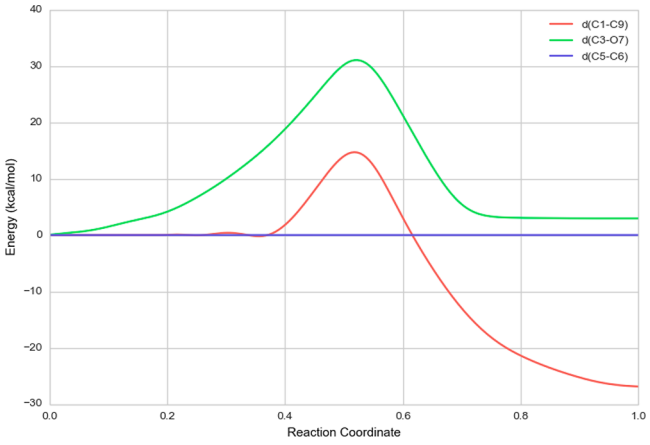


Figure 12. Decomposition of the free energy profile for the CM system into the contributions of the three CVs used. As $d(\mathrm{C5-C6})$ is not involved in this reaction, its contribution is zero.

dx.doi.org/10.1021/jp506593t | J. Phys. Chem. B XXXX, XXX, XXX−XXX

reasonable guess, such as the MEP, should alleviate this problem. Returning to Figure 12, we see that the free energy decomposition in terms of CVs provides insight into their contributions to the MFEP. Thus, the initial free energy cost arises from stretching the C3−O7 bond, which is to be cleaved. In the TS region, both C3−O7 and C1−C9 contribute, whereas the exothermicity that arises in the descent to products comes largely from the formation of the new C1−C9 bond. On the other hand, as the C5−C6 bond remains in its equilibrium position along the MFEP, its contribution is essentially zero. Stretching this bond has a free energy cost, but in this particular reaction, it is not removed from its minimum value at any point, no mean force acts on it, and, thus, its contribution is null. This decomposition of the free energy is mathematically sound in an area where different approaches have proven complex and controversial.[45−48]

CONCLUSIONS

In this work, we have implemented the SoT method[18] and used it to study two enzyme catalytic mechanisms with QM/MM potentials. We have devised a suitable SoT simulation protocol for these types of systems and have shown that the results do not depend on the dynamical evolution of the CVs used to describe the reaction: both inertial and Brownian regimes lead to the same MFEP evolution and final curve. We have also suggested the use of the MEP as an initial starting guess to accelerate the optimization and to reduce convergence problems.

The study of enzyme mechanisms via MFEPs obtained as a function of a set of CVs has several advantages over studies based on more traditional methods. First, the computational cost does not increase with the number of CVs used. Second, the path can be easily visualized and the variation of several CVs analyzed independently. Finally, the calculation of free energies based on these CVs is robust with respect to the set of CVs and can give insights into their respective contributions to the free energy barrier and reaction exo- or endothermicity.

ASSOCIATED CONTENT

Supporting Information

Figures showing minimum energy paths, the evolution of the IPL MFEP, and a free energy profile. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: ramon.crehuet@iqac.csic.es. Phone: +0034 934006116. Fax: +0034 932 045 904.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We acknowledge financial support from the Ministerio de Innovación y Competitividad (CTQ2012-33324) and the Generalitat de Catalunya (2009SGR01472). We also acknowledge access to supercomputing time from the CSUC. M.S-M. thanks the Ministerio de Economia y Competitividad for a predoctoral fellowship. We also thank Sergi Martí for kindly sharing IPL structures with us.

REFERENCES

(1) Field, M. J. Simulating Enzyme Reactions: Challenges and Perspectives. *J. Comput. Chem.* **2002**, *23*, 48−58.
(2) Field, M. J. *A Practical Introduction to the Simulation of Molecular Systems*; Cambridge University Press: Cambridge, U.K., 2007.
(3) E, W.; Vanden-Eijnden, E. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *Annu. Rev. Phys. Chem.* **2010**, *61*, 391−420.
(4) Aleksandrov, A.; Field, M. A Hybrid Elastic Band String Algorithm for Studies of Enzymatic Reactions. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12544−12553.
(5) Sheppard, D.; Terrell, R.; Henkelman, G. Optimization Methods for Finding Minimum Energy Paths. *J. Chem. Phys.* **2008**, *128*, 134106.
(6) Weinan, E.; Vanden-Eijnden, E. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *Annu. Rev. Phys. Chem.* **2010**, *61*, 391−420.
(7) Darve, E.; Pohorille, A. Calculating Free Energies Using Average Force. *J. Chem. Phys.* **2001**, *115*, 9169.
(8) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.
(9) Laio, A.; Rodriguez-Fortea, A.; Gervasio, F. L.; Ceccarelli, M.; Parrinello, M. Assessing the Accuracy of Metadynamics. *J. Phys. Chem. B* **2005**, *109*, 6714−6721.
(10) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.
(11) E, W.; Ren, W.; Vanden-Eijnden, E. String Method for the Study of Rare Events. *Phys. Rev. B* **2002**, *66*, 052301.
(12) E, W.; Ren, W.; Vanden-Eijnden, E. Simplified and Improved String Method for Computing the Minimum Energy Paths in Barrier-Crossing Events. *J. Chem. Phys.* **2007**, *126*, 164103.
(13) Johnson, M. E.; Hummer, G. Characterization of a Dynamic String Method for the Construction of Transition Pathways in Molecular Reactions. *J. Phys. Chem. B* **2012**, *116*, 8573−8583.
(14) Henkelman, G.; Uberuaga, B.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113*, 9978−9985.
(15) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String Method in Collective Variables: Minimum Free Energy Paths and Isocommittor Surfaces. *J. Chem. Phys.* **2006**, *125*, 24106.
(16) Ren, W.; Vanden-Eijnden, E.; Maragakis, P.; E, W. Transition Pathways in Complex Systems: Application of the Finite-Temperature String Method to the Alanine Dipeptide. *J. Chem. Phys.* **2005**, *123*, 134109.
(17) Vanden-Eijnden, E.; Venturoli, M. Revisiting the Finite Temperature String Method for the Calculation of Reaction Tubes and Free Energies. *J. Chem. Phys.* **2009**, *130*, 194103.
(18) Pan, A. C.; Sezer, D.; Roux, B. Finding Transition Pathways Using the String Method with Swarms of Trajectories. *J. Phys. Chem. B* **2008**, *112*, 3432−3440.
(19) Gan, W.; Yang, S.; Roux, B. Atomistic View of the Conformational Activation of Src Kinase Using the String Method with Swarms-of-Trajectories. *Biophys. J.* **2009**, *97*, L8−L10.
(20) Maragliano, L.; Roux, B.; Vanden-Eijnden, E. A Comparison between Mean Forces and Swarms-of-Trajectories String Methods. *J. Chem. Theory Comput.* **2014**, *10*, 524−533.
(21) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in Free Energy Space. *J. Chem. Phys.* **2007**, *126*, 054103.
(22) Zinovjev, K.; Martí, S.; Tuñon, I. A Collective Coordinate to Obtain Free Energy Profiles for Complex Reactions in Condensed Phases. *J. Chem. Theory Comput.* **2012**, *8*, 1795−1801.
(23) Field, M. J. The pDynamo Program for Molecular Simulations Using Hybrid Quantum Chemical and Molecular Mechanical Potentials. *J. Chem. Theory Comput.* **2008**, *4*, 1151−1161.
(24) Crehuet, R.; Field, M. J. A Transition Path Sampling Study of the Reaction Catalyzed by the Enzyme Chorismate Mutase. *J. Phys. Chem. B* **2007**, *111*, 5708−5718.

(25) Martí, S.; Andrés, J.; Moliner, V.; Silla, E.; Tuñón, I.; Bertrán, J. Mechanism and Plasticity of Isochorismate Pyruvate Lyase: A Computational Study. *J. Am. Chem. Soc.* **2009**, *131*, 16156−16161.
(26) Martí, S.; Andrés, J.; Moliner, V.; Silla, E.; Tuñón, I.; Bertrán, J.; Field, M. J. A Hybrid Potential Reaction Path and Free Energy Study of the Chorismate Mutase Reaction. *J. Am. Chem. Soc.* **2001**, *123*, 1709−1712.
(27) Martí, S.; Andrés, J.; Moliner, V.; Silla, E.; Tuñón, I.; Bertrán, J. A Comparative Study of Claisen and Cope Rearrangements Catalyzed by Chorismate Mutase. An Insight into Enzymatic Efficiency: Transition State Stabilization or Substrate Preorganization? *J. Am. Chem. Soc.* **2004**, *126*, 311−319.
(28) Szefczyk, B.; Claeyssens, F.; Mulholland, A. J.; Sokalski, W. A. Quantum Chemical Analysis of Reaction Paths in Chorismate Mutase: Conformational Effects and Electrostatic Stabilization. *Int. J. Quantum Chem.* **2007**, *107*, 2274−2285.
(29) Claeyssens, F.; Ranaghan, K. E.; Lawan, N.; Macrae, S. J.; Manby, F. R.; Harvey, J. N.; Mulholland, A. J. Analysis of Chorismate Mutase Catalysis by QM/MM Modelling of Enzyme-Catalysed and Uncatalysed Reactions. *Org. Biomol. Chem.* **2011**, *9*, 1578−1590.
(30) Repasky, M. P.; Guimarães, C. R. W.; Chandrasekhar, J.; Tirado-Rives, J.; Jorgensen, W. L. Investigation of Solvent Effects for the Claisen Rearrangement of Chorismate to Prephenate: Mechanistic Interpretation via near Attack Conformations. *J. Am. Chem. Soc.* **2003**, *125*, 6663−6672.
(31) Guimarães, C. R. W.; Repasky, M. P.; Chandrasekhar, J.; Tirado-Rives, J.; Jorgensen, W. L. Contributions of Conformational Compression and Preferential Transition State Stabilisation to the Rate Enhancement by Chorismate Mutase. *J. Am. Chem. Soc.* **2003**, *125*, 6892−6899.
(32) Guimarães, C. R. W.; Udier-Blagović, M.; Tubert-Brohman, I.; Jorgensen, W. L. Effects of Arg90 Neutralization on the Enzyme-Catalyzed Rearrangement of Chorismate to Prephanate. *J. Chem. Theory Comput.* **2005**, *1*, 617−625.
(33) Crehuet, R.; Thomas, A.; Field, M. J. An Implementation of the Nudged Elastic Band Algorithm and Application to the Reaction Mechanism of HGXPRTase from Plasmodium Falciparum. *J. Mol. Graphics Modell.* **2005**, *24*, 102−110.
(34) Galvan, I. F.; Field, M. J.; Galván, I. Improving the Efficiency of the NEB Reaction Path Finding Algorithm. *J. Comput. Chem.* **2008**, *29*, 134−143.
(35) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(36) Zaitseva, J.; Lu, J.; Olechoski, K. L.; Lamb, A. L. Two Crystal Structures of the Isochorismate Pyruvate Lyase from Pseudomonas Aeruginosa. *J. Biol. Chem.* **2006**, *281*, 33441−33449.
(37) Jorgensen, W. L.; Maxwell, D. S.; 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.
(38) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(39) Chook, Y. M.; Gray, J. V.; Hengming, K.; Lipscomb, W. N. The Monofunctional Chorismate Mutase from Bacillus Subtilis. Structure Determination of Chorismate Mutase and Its Complexes with a Transition State Analog and Prephenate, and Implications for the Mechanism of the Enzymatic Reaction. *J. Mol. Biol.* **1994**, *240*, 476−500.
(40) Hummer, G.; Kevrekidis, I. G. Coarse Molecular Dynamics of a Peptide Fragment: Free Energy, Kinetics, and Long-Time Dynamics Computations. *J. Chem. Phys.* **2003**, *118*, 10762.
(41) Villa, J.; Strajbl, M.; Glennon, T. M.; Sham, Y. Y.; Chu, Z. T.; Warshel, A. How Important Are Entropic Contributions to Enzyme Catalysis? *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 11899−11904.
(42) Bolhuis, P. G.; Dellago, C.; Chandler, D. Reaction Coordinates of Biomolecular Isomerization. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 5877−5882.
(43) Bonella, S.; Meloni, S.; Ciccotti, G. Theory and Methods for Rare Events. *Eur. Phys. J. B* **2012**, *85*, 1−19.
(44) Ranaghan, K. E.; Ridder, L.; Szefczyk, B.; Sokalski, W. A.; Hermann, J. C.; Mulholland, A. J. Transition State Stabilisation and Substrate Strain in Enzyme Catalysis: Ab Initio QM/MM Modelling of the Chorismate Mutase Reaction. *Org. Biomol. Chem.* **2004**, *2*, 968−980.
(45) Brady, G. P.; Szabo, A.; Sharp, K. A. On the Decomposition of Free Energies. *J. Mol. Biol.* **1996**, *263*, 123−125.
(46) Brady, G. P.; Sharp, K. A. Decomposition of Interaction Free Energies in Proteins and Other Complex Systems. *J. Mol. Biol.* **1995**, *254*, 77−85.
(47) Boresch, S.; Karplus, M. The Meaning of Component Analysis: Decomposition of the Free Energy in Terms of Specific Interactions. *J. Mol. Biol.* **1995**, *254*, 801−807.
(48) Mark, A. E.; van Gunsteren, W. F. Decomposition of the Free Energy of a System in Terms of Specific Interactions. Implications for Theoretical and Experimental Studies. *J. Mol. Biol.* **1994**, *240*, 167−176.

4.1.3 'In silico' enzymatic reactions induced by high radiation damage

X-ray chrystallography is a technique used to study protein structures, being the different substates of the chemical reaction (reactants analogs, transiton state analogs and products) trapped into crystals. It is well known that the X-ray crystallographic methods based on synchrotron radiation techniques provoke undesirable radiation damage effects in the protein crystal structure. This damage is due to the source of radiation employed.

In collaboration with Professor Weik and coworkers, in Grenoble, we studied two crystal structures representing the apo and the holo forms of the Lacatate Dehydrogenase (LDH). They found a tryptophan decarboxylation in the apo form (not in the holo form) that is not found *in vivo*. Weik and coworkers concluded that an electron hole is created in tryptophan Trp62, that is transferred to Asp70 in the apo form via Arg64, but not in the Holo form.

test

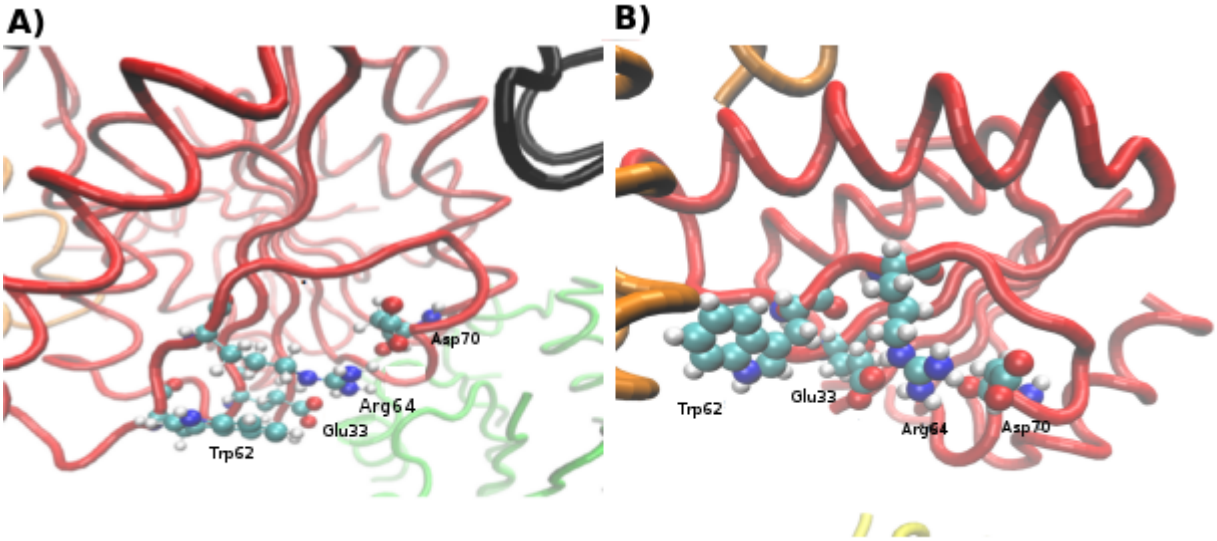


Figure 4.4: A) Active center of the Apo form. B) Active center of the Holo form. In both, there are represented Trp62, Glu33, Arg64 and Asp70. The blue balls are N, the white H, the red O and the turquoise C. The tapes represents chains of proteins that have been taken into account in the calculations. Here have been only represented some of them, concretely, four of six.

After carefully revising this hypothesis, some questions emerged. It seems clear that there is an electron transfer (ET) from Asp70 to Trp62 (Or a hole transfer (HT) in the opposite direction) in the apo form whereas in the holo form, this

ET is not possible. However there are in principle a couple of possible answers. 1) It could be impossible because it is thermodynamically inaccessible. That is, because the electron on the Asp70 is much less stable in the holo form relative to the apo form. 2) It could be because it is kinetically forbidden: the HT rate could be much smaller for the holo than for the apo form. Furthermore, there is another question once it is accepted that the transfer process takes place: why the proximal Glu33 does not decarboxylate whereas Asp does?.

To check Weik's hypothesis and to answer these open questions, we performed MD simulations over the apo and holo crystal structures, followed by QM/MM calculations (reaction coordinate scans) over different trajectory snapshots. Within the calculated paths we performed electron couplings calculations using the Fragment Charge Differentiation (FCD) method developed by Voityuk and Rösch and extensively used in biological charge transfer calculations.

First of all we confirmed that it is a HT process and we located the Hole. By calculating ionization potentials (IP) we confirmed that Glu33 IP is higher than Asp70 both in the holo and the apo form and thus the radical created can not be transferred from Glu33 to Asp70. Also we observed that the residue more easy to ionize is Trp62 (lower IP) followed by Asp70 in agreement with Weik's hypothesis. Furthermore we confirmed that thermodynamically the process is possible both in the apo and the holo form. Thus, why it is not observed in the holo form?

Then, we tested the most plausible biological HT mechanism: Direct and Bridge assisted and within this last one, Superexchage and Hopping (also called sequential) mechanisms. First of all we discarded the direct mechanism because even the shortest distance between Asp and Trp over all the simulations is too large in both apo and holo forms (7.50 and 12.40Å) to expect an overlap between the donor and acceptor orbitals. This fact results in electronic couplings negligibles (10^{-13}-10^{-15}) eV within the apo mechanism and zero for the holo form (see Fig. 4.5). The Hopping mechanism was also discarded because the IP calculations showed that Glu33 and Arg64 exhibited higher IPs, so the transfer is impossible. Thus the process proceeds through a superexchange mechanism, i.e., the charge travels from Trp to Asp, through the orbitals of Glu and Arg which influence the reaction indirectly creating the suitable environment in which the charge transfer is produced. The Electronic couplings calculations revealed that within this mechanism the apo form has a coupling around 10^{-6} eV. This value is small compared to other biological processes (10^{-2}-10^{-4} eV), but this difference is normal because the process we observe is slower than other biological events and takes place at the scale of hours. Furthermore the value of the coupling in the holo form is even lower (around two orders of magnitude). The HT is so slow that is not seen in the X-ray crystallography time scale. This is the reason why it is not observed in the holo form.

Summarizing, we have devised the suitable computational setup to study radiation damage decarboxylation (at least in LDH proteins), showing that the process is due to a HT instead of an ET. In addition we found the mechanism through it proceeds (superexchange) and the reason why it take place in the apo and not in the holo form.

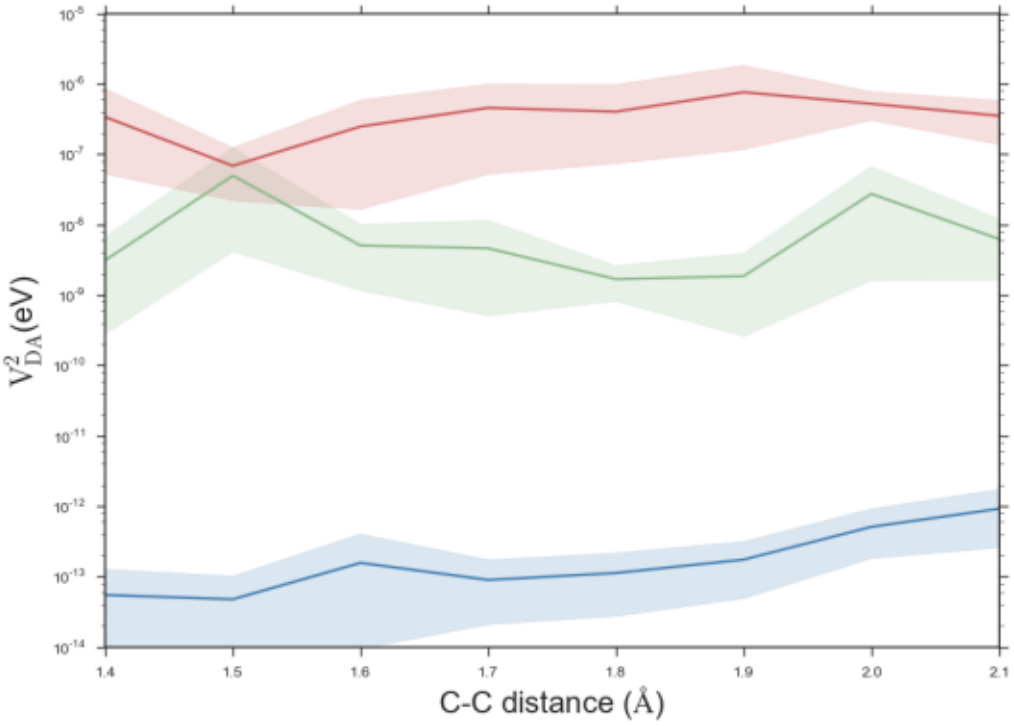


Figure 4.5: Average coupling term for the superexchange and direct mechanism obtained from the 16 profiles calculated. The coupling was calculated at the optimized geometries with the hole on the Trp and on the Asp. Red: coupling for the apo form and Green: coupling for the holo form within the superexchange mechanism, Blue: Coupling for the apo form and direct mechanism. For the Holo form the coupling is zero and thus is not shown.

Tryptophan-mediated decarboxylation in proteins: evidence from X-ray crystallography and QM/MM simulations

Melchor Sanchez-Martinez[4‡], Nicolas Coquelle[1,2,3‡], Alexander Voityuk[5,6], Dominique Madern[1,2,3], Ramon Crehuet[4*], Martin Weik[1,2,3*]

[1]Univ. Grenoble Alpes, Institut de Biologie Structurale (IBS), F-38027 Grenoble, France

[2]CEA, DSV, IBS, F-38027 Grenoble, France

[3]CNRS, IBS, F-38027 Grenoble, France

[4]Institute of Advanced Chemistry of Catalonia (IQAC-CSIC), E-08034 Barcelona, Spain

[5]Institut of Computational Chemistry and Catalysis, University of Girona, E-17071 Girona, Spain

[6]Catalan institution of research and advanced studies (ICREA), E-08010 Barcelona

‡These authors contributed equally

* corresponding authors

Abstract

Decarboxylation of acidic residues in proteins is one of the most prominent signatures of radiation damage in macromolecular X-ray crystallography. So far, structural features that might rationalize the broad distribution of decarboxylation probabilities of chemically identical groups in a protein have remained elusive. Here, we provide evidence that hole transfer from a nearby tryptophan residue can cause decarboxylation as shown by QM/MM simulations confirming the finding by dose-dependent X-ray crystallography (Data not published yet). Furthermore by electronic couplings calculations we explain how this transfer take place discriminating between biologically possible scenarios.

Introduction

Intense X-ray beams from synchrotron sources create specific chemical and structural damage in proteins during crystallographic data collection. This damage even occurs at 100 K, the temperature at which the vast majority of crystallographic data is collected[1-3]. Among the most prominent manifestations in protein crystallographic radiation damage studies are breakage of disulfide bonds and decarboxylation of glutamic and aspartic amino acid residues. This specific damage does not result from direct absorption of an X-ray photon by one of the atoms in the radiation-sensitive group (primary damage), but rather is the manifestation of damage inflicted by secondary radicals created after primary photoabsorption elsewhere in the protein or the surrounding solvent (secondary damage) [4]. Chemically identical groups in the same protein display differential radiation sensitivities. Differences in the chemical and structural environment must be at the origin of the differential sensitivities, yet so far,

they have remained largely elusive[5].

Decarboxylation of acidic residues[2] is initiated by the capture of a secondary hole on the side chain, resulting in the generation of CO_2 and a carbon-centered radical[6]. Protein crystallographic radiation damage studies did not provide evidence for a correlation between the radiation-sensitivity of an acidic residue and its solvent exposure[1,7] or distance to the protein surface[8], at least not at 100 K. A possible relation between the pKa of a carboxyl group and its radiation sensitivity has been controversially discussed[2,7].

Lactate dehydrogenases (LDH) catalyze the conversion of pyruvate to lactate, requiring NADH as a cofactor[9]. Binding of NADH and the substrate analogue oxamate to LDH from *Thermus thermophilus* (*Tt*LDH) results in conformational changes with residues moving as far as 10 Å[10]. These changes alter the chemical and structural environments of certain radiation-sensitive groups, such as carboxylates. In this work, we have invistigateed *Tt*LDH (apo and holo) crystals that consists in a homotetramer similar to the PDB accession numbers 2V6M and 2V7P. The enzyme was purified and crystallized as described earlier[10], either in its apo form, or in complex with NADH and oxamate (holo form). Experimentally, by radiation damage studies, was found that the apo form it is highly damaged, whereas the holo form it is less, as indicated by differences in the size of the $F_o^4 - F_o^1$ difference Fourier maps (Figure 1). Radiation-induced decarboxylation takes thus place to a much larger extend in the apo than in the holo form.

Tryptophan-containing di- and tripeptides have been studied after UV photolysis by electron spin resonance at 77 K[11]. An electron hole was initially created on the aromatic ring to form an aromatic pi-cation radical. The electron hole was then transferred from the photionized aromatic ring to a nearby carboxyl group, followed by decarboxylation. The authors suggested that a similar process could take place in irradiated proteins. Have been hypothesized that the structural results in Figure 1 of crystalline *Tt*LDH irradiated at 100 K in an X-ray crystallographic experiment could be explained by hole transfer (HT) from Trp62 to Asp70 that takes place in the apo but not in the holo form. Furthermroe, in the apo form, there is an Arg64 in between the aromatic Trp62 and the acidic Asp70 residues and its linked to the latter via a hydrogen-bond assisted salt bridge. Such a salt bridge has been shown to increase the efficiency of electron transfer in proteins[12]. Also that a radical preferentially forms on a Trp is a known fact over the literature[13,14,15,16].

In this work, by Quantum Mechanical/Molecular Mechanical (QM/MM) hybrid simulations we analysed specific radiation damage in *Tt*LDH crystals in its apo form and in ternary complex with a substrate analogue and a co-factor, holo form. The radiation sensitivity of one acidic residue halves when the distance to a nearby tryptophan doubles due to conformational changes that accompany formation of the ternary complex. This observation is in line with a mechanism in which an X-ray induced hole, initially localized on the tryptophan residue, is transferred to the acidic residue, thereby triggering its decarboxylation (experimental observations given by X-ray dose crystallography not

already published). The aim of this study is four fold: 1) Confirm that the charge transfer (CT) is happening, 2) Check whether a HT (Trp to Asp) or instead an ET in the opposite direction is occurred, 3) Find the biologically plausible mechanism in that this Charge transfer proceeds and 4) if we will confirmed the transfer between Trp and Asp explain why the transfer dos not take place between Trp and Glu that is nearer to the donor than Asp.

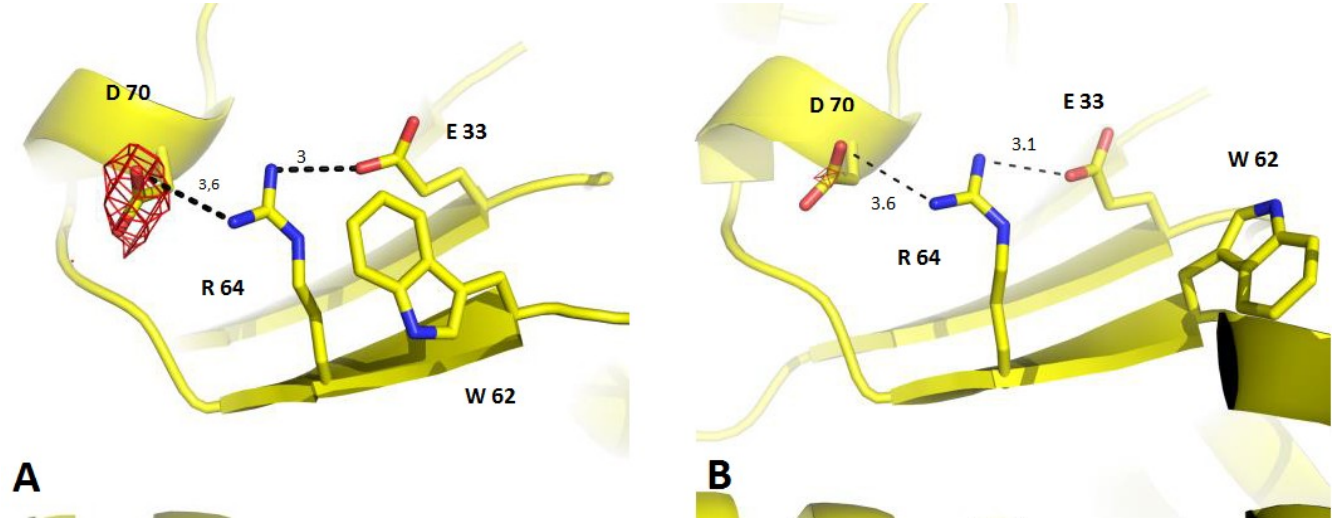


Figure 1 Damage to Asp70 in the apo (A) and the holo form (B) of TtLDH. Dashed lines represent H-bonds, whose lengths are indicated. Negative electron density in $F_o^4 - F_o^1$ difference Fourier maps is shown in red. The negative density on the carboxyl group of Asp70 in the apo form (A) indicates this residue is decarboxylated under the influence of X-irradiation. In the holo form (B), the negative density on Asp70 is much smaller, indicating a lower radiation sensitivity than in the apo.

Methods

System setup

We prepared the system from the PDB files, for both Apo and Holo forms, using pDynamo[17] software following these steps: (1) We generated a MM model for the system adding hydrogens and using the standard protonation states for the acidic or basic residues. We did not take into account for our calculations the presence of the ligand at the active center, because is too distant to have a significant role in the process. (2) We minimized the structure of the vacuum system (3) We solvated the system, with an equilibrated water box and do a short dynamic of the water molecules (0.2ps) (4) We then pruned the system to a sphere of 25Å of Arg64. This was done to minimize the number of point-charge interactions to calculate with QM/MM as the Orca[8] interface calculates all interactions, without cutoffs. The resulting system had 7251 atoms in the apo form and 7528 atoms in the holo. An outer sphere of 2 Å was kept fixed. We performed 2 Langevin dynamics of 0.1ns each with tethered heavy atoms, using force constant of 1000 and 500 kJ/Å^2. We did not use a zero force constant at the end, because the salt

bridge between Asp70 and Arg64, present in the crystal, was lost when removing all tethers. (5) We performed a production Langevin dynamics for each of the crystal forms. The first 500ps were kept as equilibration. Then we took around 10 snapshots of the trajectory separated 500ps each. These snapshots were used to calculate the decarboxylation and the couplings with QM/MM methods.

Calculation of IP

To calculate the stability of the radical, we implemented calculations of the Ionization Potential (IP) for some residues around Trp70. This were done by individual QM/MM calculations, with the relevant residue comprising the QM region, and the rest in the MM region. Four residues were considered: Trp62, Arg64, Glu33 and Asp70.

Based on Koopman's theorem, we estimated the IP of each residue from the energy of the HOMO. These calculations gave us a measure of the energy cost to generate a hole in each of the residues. The higher the IP is, the less stable a hole is. This calculation were done for 50 frames of both molecular dynamics trajectories, where each frame separation is 100ps.

Calculation of decarboxylation profiles

To calculate the energy profile for the decarboxylation process, we performed reaction coordinate scans for structures where the radical is located on Trp62 and Asp70. We used as starting structures the snapshots of the molecular dynamics, as described above.

The QM region for the optimization consisted of the atoms of Trp62 and Asp70 (33 atoms). The QM region was described with the BHLYP functional and the SVP[9] basis set. We chosen this functional because it has a high content of Hartree-Fock exchange, and avoids the excessive delocalization of radicals that pure DFT gives. Previous studies showed that it is adequate to treat radical species[20, 21]. Single point calculations were performed on top of the optimized geometries. These calculations included Trp62, Asp70, Glu33 and Arg64 in the QM region (66 atoms), using the BHLYP[22, 23] functional , and the SVP basis set. To improve the energetics of the process, we recalculated the exothermicity with MP2[24, 25] and a TZVP[19] basis set and the large QM region: the difference in energy between the minimum with the hole on the Trp and the final point for the decarboxylation curve. We only report these pairs of points because the MP2 was unstable in the regions where both curves mix. Table 1 compares the total exothermicities with DFT and MP2.

In order to locate the radical on Trp62 or Asp70, only these residues were set in the QM region in an initial optimization. When enlarging the QM region, both for the optimization or the single-point calculations we checked that the orbitals still described the same radical. Along the scans, the orbitals of the previous point were used as starting guesses to help maintain the radical at a given residue.

We calculated about 10 different decarboxylation processes for each crystal form. In some of them the geometry scan was discontinuous. We could correct some with several passes of forward and backward scans. In some cases the discontinuities remained. We finally kept 8 profiles for each form to perform all further analysis.

Calculation of electron couplings

We have used the Fragment Charge Differentiation (FCD) method to calculate the coupling term[26]. This method has been developed by Voityuk and co-workers and has been applied to several biological systems[27, 28, 29]. It gives results very close to the Generalized Mülliken-Hush, at a lower computational cost. The coupling between the donor and the acceptor is defined by:

$$V_{DA}=\frac{\left|\left(E_2-E_1\right)\right|\Delta q_{12}}{\sqrt{\left(\mu_1-\mu_2\right)^2+4q_{12}^2}} \qquad (1)$$

All the terms in the previous equation, based on the one-electron approximation can be calculated from the molecular orbital coefficients and overlap matrix of the neutral species. Ideally, one should thus optimize the geometry of the closed-shell species and estimate the energetics of the process based on the orbital energies. We tried this approach, but the decarboxylation process was poorly described when optimizing the closed-shell species. Thus, we needed to optimize the geometries for both the acceptor and the donor, i.e. for the hole on Trp62 and Asp70. When doing this we can calculate the coupling for two different geometries. As the geometry for the crossing between donor and acceptor states should be intermediate between both geometries, we expect the coupling to be also intermediate between the two calculated couplings, or at least of the same order. As discussed in the results, the difference between apo and holo form are very large, so that the choice of the donor or acceptor geometry does not affect our conclusions.

Average energy barrier

The averaging energy was calculated as the apparent energy barrier that would give the same rate as the average rates arising from the different barriers[30].

$$-\Delta E_{ave}^{\ddagger}=-RTln\left\{\frac{1}{n}\sum_{i=1}^{n}\exp\left(\frac{-\Delta E_i^{\ddagger}}{RT}\right)\right\} \qquad (2)$$

where $-\Delta E_{ave}^{\ddagger}$ is the average barrier height, R is the gas constant, n is the number of energy profiles considered , $\Delta E_i^{\ddagger}$ is the barrier height of each snapshot and T is te temperature.

The kinetic average depends on the temperature. During the collection data, the crystal is cooled with liquid nitrogen but is heated by the radiation, so that the temperature is not well-defined. In the main text, we give the average assuming a temperature of 300K. However, the result is not very sensitive to the temperature. At 300K the average barrier from the initial state is 0.31eV and 0.30eV if calculated at

193K, and changes from 0.39eV to 0.38eV in the holo form. The average barrier from the initial state is 0.29eV if calculated at 100K and changes from 0.36eV in the holo form.

Software used

All calculations have been performed with the pDynamo library[7], coupled to the Orca program[18], which performed the QM calculation. The cclib[31] library has been used to extract orbital coefficients to calculate the electronic coupling. The figures have been created with VMD[32] and the python libraries Matplotlib[33] and Seaborn[34].

Results

Location of the Hole

The fact that tryptophan residues act as a hole sink, where charge holes are primary localized is confirmed by our calculations. One would expect that the hole goes to the residue with a lower ionization potential (IP), because that will create the most stable radical-ion. In the Holo form, Asp70 has a lower ionization potential in 41 out of 51 analyzed structures. In the apo form this ratio increases up to 49 out of 51 structures. On average, the IP of Trp62 is 0.41(0.92) eV lower than that of Asp70 in the holo (apo) form. The fact that decarboxylation is observed in the experimental studies suggests that there must be a mechanisms by which, if the hole is initially located on the Tryptophan, it can be transferred to Asp70.

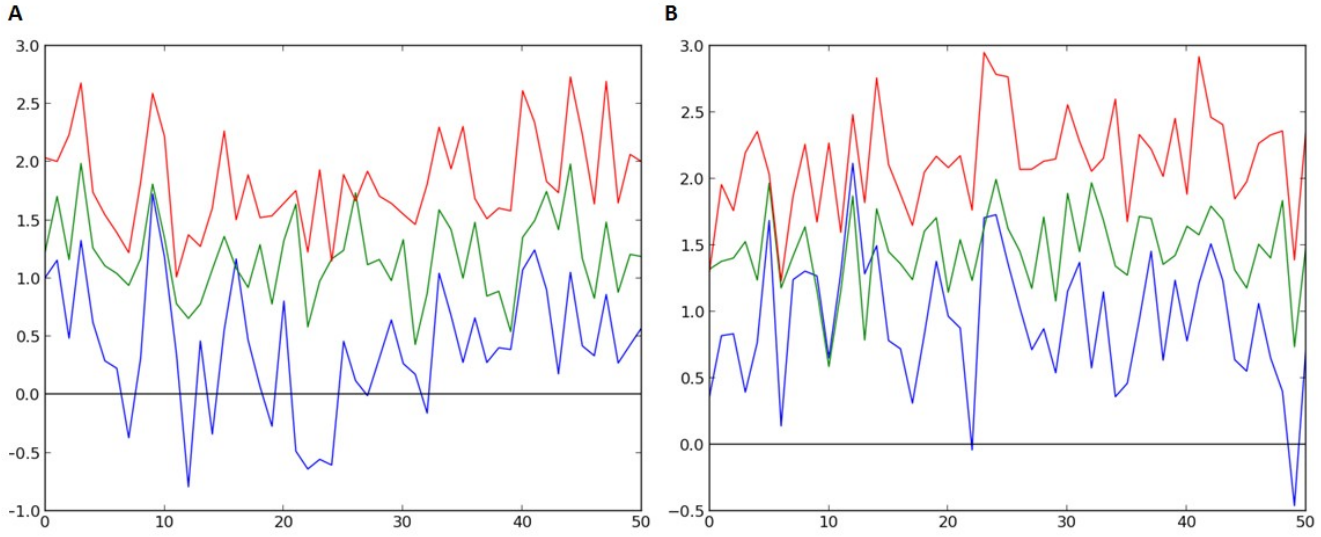


Figure 2. Energy difference, in eV, between the ionization potential of Asp70 (blue), Glu33 (green), and Arg64 (red) and Trp62 as a reference These structures were obtained from molecular dynamics simulations. ***A*** *Holo crystal.* ***B*** *Apo crystal.*

Figure 2 shows that the fluctuations of the ionization potential (IP) difference between Asp70 and

Trp62 is large, and, in both forms, positive in most of the snapshots. Thus, both apo and holo forms seem not to favour a hole transfer from Trp62 to Asp70. Before dealing with the decarboxylation event, let us first consider why the hole should be transferred to Asp70 and not to other closer residues. In particular, if decarboxylation is the expected result from the hole transfer, one may wonder why Glu33 is not decarboxylated, being also a carboxylic acid, and closer to Trp62 (see Figure 1).

As Figures Figure 2 shows, Glu33 IP is higher than Asp70 both in the holo and the apo form. In the Holo form their mean IP difference is 0.77eV and it is 0.53eV in the apo form. Thus, if an electron can be transferred from Asp70 and Glu33, the first one produces a radical that is more stable. As would be expected from a positive residue, Arg64 has the highest IP in both crystal forms. This residue has also been plotted because it plays a role in the charge transfer process, even if it does not host the hole.

Decarboxylation

Let us assume, by now, that a charge transfer is possible from Trp62 and Asp70. How can decarboxylation favour the hole transfer? We have shown that a hole is less stable on Asp70 than on Trp62. That is true for the geometry corresponding to neutral residues and also to the optimized geometries we computed locating the hole on the aspartate and on the tryptophan. However, when the hole resides on the aspartate, the energy barrier for decarboxylation is very low (see Figure 3). On average it is 0.39 for the holo and 0.32eV for the apo form. The lowest barrier is of 0.34 (0.27) in the holo (apo) crystal. This barrier is low and corresponds to a fast process. When it is considered from the starting state, with the hole on the Trp, this average barrier is 0.37 for the apo form and 0.42 for the holo, which is still a low barrier.

The decarboxylation is exothermic, and in several of the snapshots, it leads to structures where the hole on Asp70 is lower in energy than the hole in Trp62 (Figure 3). As we are doing only potential energy scans, we are probably underestimating the exothermicity of this process, as the water molecules will reorganize to better solvating the leaving CO_2 moiety (even at low temperature). Higher level MP2 calculations also show that the BHLYP functional employed underestimates this exothermicity. The average energy difference between the state with the hole on the Trp and the final state with the hole on the decarboxylated Asp is -0.91eV in the apo form and -1.24 in the holo form at the MP2 level. Therefore, if the hole can fluctuate between Asp70 and Trp62, even if most of the time is on Trp62, it will eventually reach Asp70 with a stretched C-C bonds, and that will lead to an irreversible decarboxylation.

Structure	**BHLYP/SVP**	**MP2/TZVP**
Apo-1	-0.614	-1.743
Apo-2	0.332	-0.857
Apo-3	1.139	-0.376
Apo-4	0.217	-0.954

Apo-5	0.906	-0.327
Apo-6	0.566	-0.671
Apo-7	-0.239	-1.602
Apo-8	0.473	-0.720
Holo-1	0.0409	-1.109
Holo-2	-0.187	-1.467
Holo-3	0.219	-1.026
Holo-4	-0.515	-1.693
Holo-5	-0.011	-1.265
Holo-6	-0.020	-1.332
Holo-7	0.427	-0.823
Holo-8	0.041	-1.238

Table 1 Exothermicity of each decarboxylation process, calculated as the energy difference (in eV) between the minimum of the curve with the electron on the Trp62 and the minimum of the curve with the electron on Asp (which corresponds to the decarboxylated species)

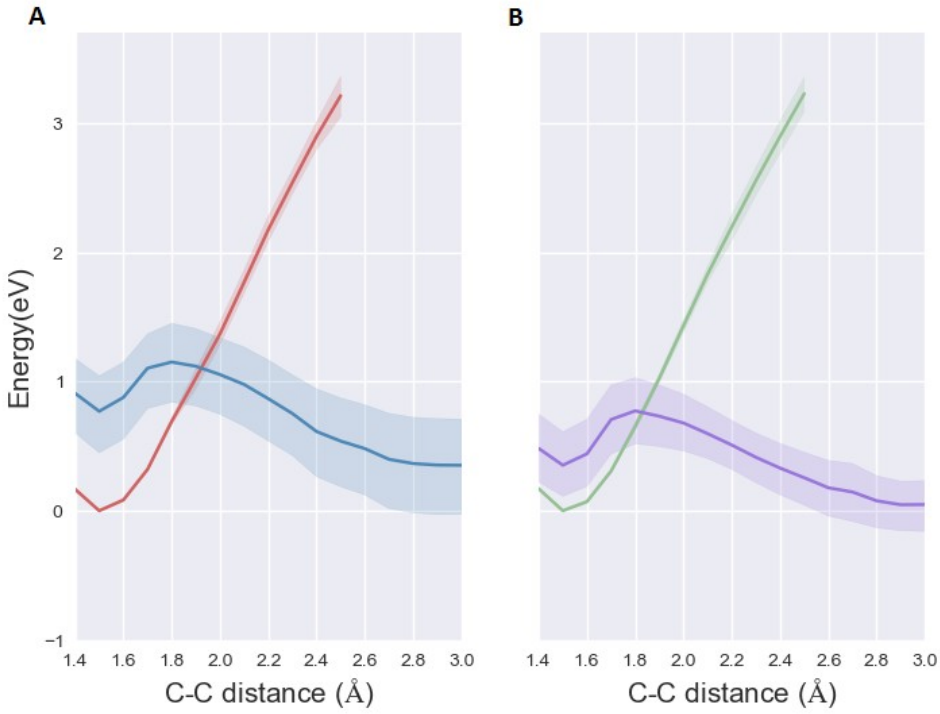


Figure 3 ***A*** *Energy profile for the C-C bond elongation of Asp70 in the apo from. Blue lines correspond to the state where the hole is on the Trp62, and red lines correspond to the state where the hole is on Asp70.* ***B*** *Energy profile for the C-C bond elongation of Asp70 in the holo from. Purple lines correspond to the state where the hole is on the Trp62, and green lines correspond to the state where the hole is on Asp70. We depict 8 decarboxylation processes starting from 8 different snapshots of the*

Molecular Dynamics (see methods) at both the apo and the holo forms. In all cases the zero of energy has been set as the minimum for the hole on the Trp62.

Hole transfer process

Given that a hole on Asp would lead to decarboxylation, the question is whether the hole transfer can take place between the Trp62 and the Asp70.

The rate of a charge transfer process is proportional to the crossing of the donor and acceptor surfaces, the vibronic coupling of these surfaces and the electronic coupling term (V_{DA}). The first two terms usually take similar values in similar systems. The resemblance of the profiles between the two crystal forms suggests that the energy barrier at the crossing point will be similar. The exact nature of the chemical process also indicates that the vibronic coupling will not create significant differences[26, 35, 36]. Thus the electronic coupling arises as the main source of distinction between the apo and the holo form. The rate of a charge transfer process is proportional to the square of the electronic coupling term.

Different mechanisms for the charge transfer determine different coupling terms. The charge transfer between two species can take place directly, when their orbitals are in contact. If a large distance separates donor and acceptor, bridge-assisted (sequential and superexchange) mechanisms dominate[37].

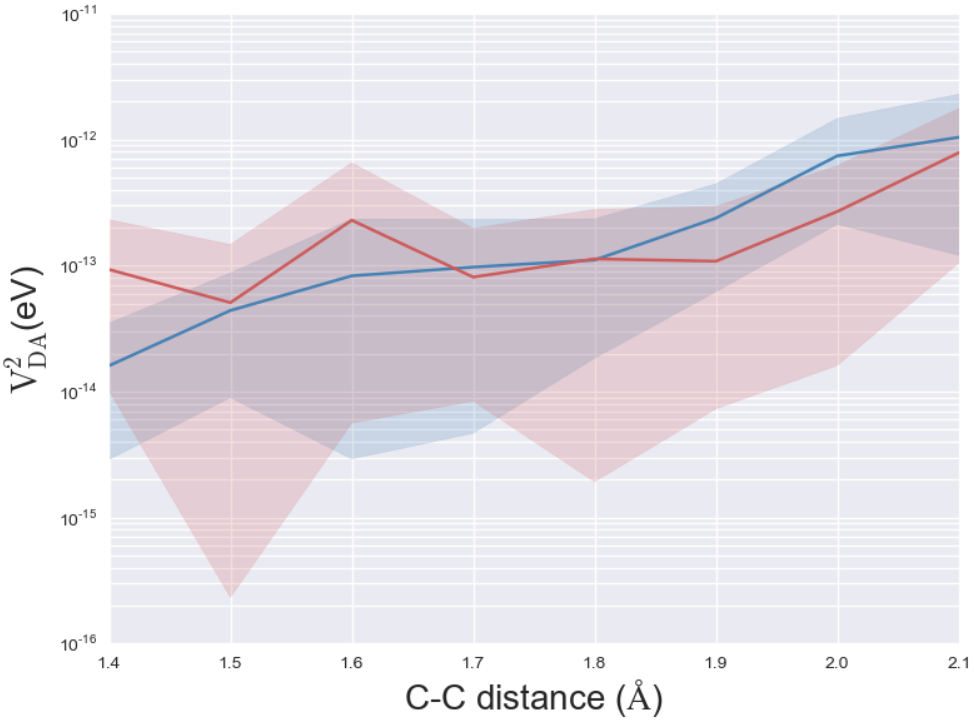


Figure 4 Coupling term for the direct mechanism. The results for the Holo form are zero, and therefore are not shown in the current logarithmic scale.

We have calculated V_{DA} for the direct mechanism. The shortest distance between the Asp and the Trp

in the holo form is 12.40 Å and in the apo form is 7.50 Å (taken from the crystal structure). Both are large distances to expect an overlap between the donor and acceptor orbitals, which results in negligible couplings as shown in Figure 4.

The second possible mechanisms is called sequential, because the hole travels from one donor to the acceptor by hopping to intermediate sites that temporarily host the hole. It corresponds to a chain of charge transfer events and is common in biological systems such as photosynthetic complex[38]. In the system under study, Glu33 and Arg64 are the only intermediate residues that could host the charge. We have seen that the ionization potential of these residues is higher than for the Asp62, which renders this process thermodynamically unfavorable.

The third mechanism is superexchange. In superexchange, the charge travels directly from donor to acceptor, but it travels through the orbitals of intermediate sites, without interact with them.

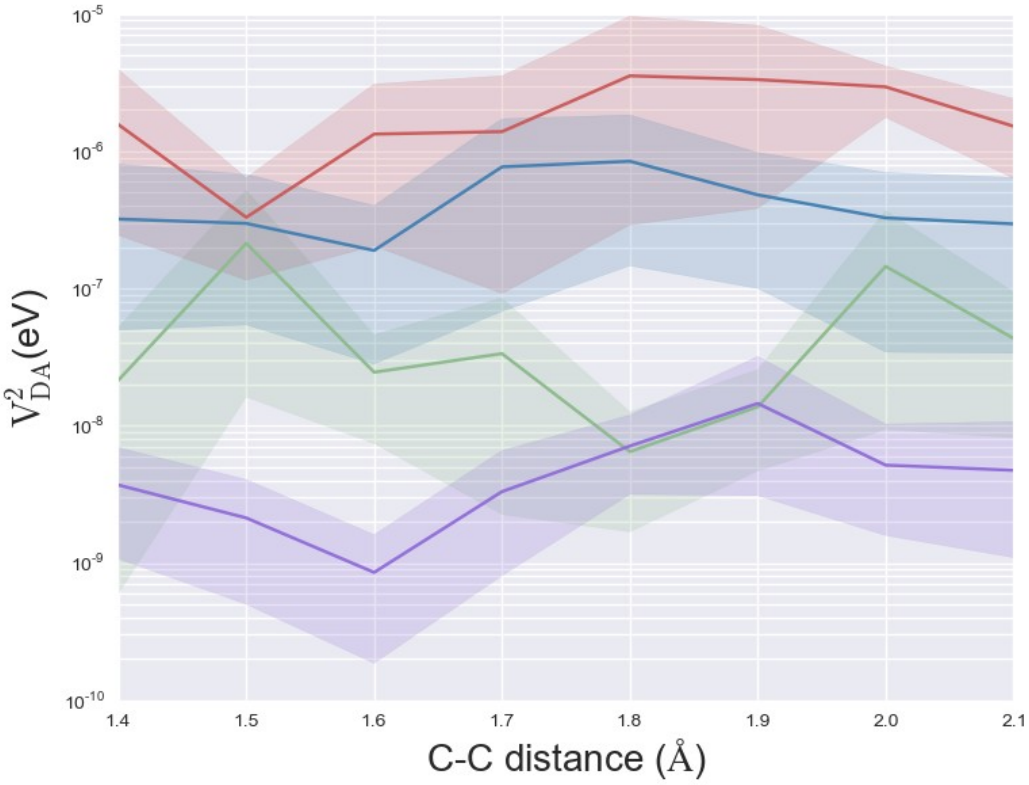


Figure 5 Average coupling term for the superexchange mechanism obtained from the 16 profiles calculated. The coupling was calculated at the optimized geometries with the hole on the Trp (blue: apo form, purple: holo form) and on the Asp (red: apo form, green: holo form).

Figure 5 shows the couplings for the apo and the holo form. It is easy to see that the couplings in the

apo form are up to three orders of magnitude larger than in the holo form. As described in the methods section, calculating a coupling for two different geometries (hole on Trp and hole on Asp) is an approximation, and we expect the actual coupling to lie in between these two values.

The overall rate will be determined by a Boltzmann average of the rates at different C-C distances, and therefore, the most relevant region is the region before the crossing of the two profiles: at C-C distances lower than 2Å. In this region, the apo form has a coupling around 10^{-6}eV. This value is small compared to other biological processes, but we have to consider that the process we observe takes place at the scale of hours. The value of the coupling in the holo form is at least two orders of magnitude lower. This explains why, even if the hole transfer is thermodynamically possible, it is kinetically much slower than in the apo form. So slow that it is not seen in the time-scale of data acquisition in X-ray diffraction.

Conclusions

The biological samples with acidic residues are susceptible to suffer radiation damage events ending in a decarboxylation reaction during its crystallization. In this work we have used TtLDH to study this process. We have checked whether the experimental observations stating that a HT is produced in the Trp62 could be reproducing and explored by QM/MM hybrid simulations.

We have showed that understanding of side-specific radiation damage needs to take into account neighboring residues even if they don't have a direct role into the process. We have proved that the observed HT corresponds to a superexchange mechanism between Asp and Trp (with Arg and Glu as necessary intermediates). Also we showed by electronic couplings calculations why the decarboxylation is produced in both crystal forms but the HT is only observed in the apo form.

Furthermore we have proved that the FCD method is a useful way to characterize charge transfer processes produced by radiation damage because with few experimental parameters the transfer process could be estimate.

Bibliography

1 W. P. Burmeister, Acta Crystallogr D Biol Crystallogr **56**, 328 (2000).
2 R. B. Ravelli and S. M. McSweeney, Structure Fold Des **8**, 315 (2000).
3 M. Weik, et al., Proc Natl Acad Sci U S A **97**, 623 (2000).
4 E. Garman, Acta Crystallographica Section D **66**, 339 (2010).
5 K. A. Sutton, P. J. Black, K. R. Mercer, E. F. Garman, R. L. Owen, E. H. Snell, and W. A. Bernhard, Acta Crystallogr D Biol Crystallogr **69**, 2381 (2013).
6 M. D. Sevilla, J. B. D'Arcy, and K. M. Morehouse, J. Phys. Chem. **83**, 2893 (1979).
7 E. Fioravanti, F. M. Vellieux, P. Amara, D. Madern, and M. Weik, J Synchrotron Radiat. **14**, 84 (2007).
8 D. H. Juers and M. Weik, J Synchrotron Radiat. **18**, 329 (2011).
9 J.J. Holbrook, A. Liljas, S.J. Steindel, M.G. Rossmann, P.D. Boyer (Ed.), The Enzymes, Academic Press, New York (1975), pp. 191–292

10 N. Coquelle, E. Fioravanti, M. Weik, F. Vellieux, and D. Madern, J. Mol. Biol. **374**, 547 (2007).
11 Y. Matsui, K. Sakai, M. Murakami, Y. Shiro, S. Adachi, H. Okumura, and T. Kouyama, J. Mol. Biol. **324**, 469 (2002).
12 M. D. Sevilla and J. B. Darcy, Journal of Physical Chemistry **82**, 338 (1978).
13 Ranguelova, K.; Suarez, J.; Magliozzo, R. S.; Mason, R. P. Biochemistry 2008, 47, 11377.
14 Shafaat, H. S.; Leigh, B. S.; Tauber, M. J.; Kim, J. E. J Phys Chem B 2009, 113, 382.
15 Colin, J.; Wiseman, B.; Switala, J.; Loewen, P. C.; Ivancich, A. J Am Chem Soc 2009, 131, 8557.
16 Smith, A. T.; Doyle, W. A.; Dorlet, P.; Ivancich, A. Proc Natl Acad Sci U S A 2009, 106, 16084.
17 Field, M. J. Journal of Chemical Theory and Computation 2008, 4, 1151.
18 Neese, F. Wiley Interdisciplinary Reviews: Computational Molecular Science, 2, 73.
19 Schafer, A.; Horn, H.; Ahlrichs, R. The Journal of Chemical Physics 1992, 97, 2571.
20 Solans-Monfort, X.; Branchadell, V.; Sodupe, M.; Sierka, M.; Sauer, J. The Journal of chemical physics 2004, 121, 6034.
21 Félix, M.; Voityuk, A. a. The journal of physical chemistry. A 2008, 112, 9043.
22 Becke, A. D. The Journal of Chemical Physics 1993, 98, 1372.
23 Lee, C.; Yang, W.; Parr, R. G. Physical Review B 1988, 37, 785.
24 Neese, F. Journal of Computational Chemistry 2003, 24, 1740.
25 Kossmann, S.; Neese, F. Journal of Chemical Theory and Computation, 6, 2325
26 Voityuk, A. a.; Rösch, N. The Journal of Chemical Physics 2002, 117, 5607.
27 Wallrapp, F. H.; Voityuk, A. A.; Guallar, V. Journal of Chemical Theory and Computation 2010, 3241.
28 Wallrapp, F.; Voityuk, A.; Guallar, V. Journal of Chemical Theory and Computation 2009, 3312.
29 Voityuk, A. A. Chemical Physics Letters 2010, 495, 131.
30 Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. Chemical Society reviews 2012, 3025.
31 O'Boyle, N. M.; Tenderholt, A. L.; Langner, K. M. Journal of Computational Chemistry 2008, 29, 839.
32 Humphrey, W.; Dalke, A.; Schulten, K. Journal of Molecular Graphics 1996, 14, 33.
33 Hunter, J. D. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 2007, 9, 90-95
34 https://github.com/mwaskom/seaborn
35 Newton, M. D. Chemical reviews 1991, 91.
36 Albinsson, B.; Eng, M. P.; Pettersson, K.; Winters, M. U. Physical chemistry chemical physics : PCCP 2007, 9, 5847.
37 Charge and Energy Transfer Dynamics in Molecular Systems; May, V.; Kühn, O., Eds.; Wiley-VCH, 2004; Vol. 6.
38 Cordes, M. and Gies, B., Chem. Soc. Rev., 2009, 38, 892-901

4.2 Global Motions

4.2.1 Cooperativity of secondary structure elements in protein ensembles

Secondary structure is an important element of IDPs. As we have commented (section *1.4.1*) its structure and dynamics are very related to their function. Thus as they are very implicated in disease, understanding their structural characteristics is very relevant.

Some regions (called MoRFs) of IDPs can adopt transient secondary structure configurations. When we generate IDPs conformational ensembles, generally it is difficult to determine its composition. Sometimes the conformational propensities of the individual residues hide the cooperative nature of these ensembles. Thus, as commented previously it is necessary to differentiate when a fragment has regions that adopt a conformation in a secondary structure and when that fragment contains a true secondary structure, with all the residues adopting that conformation (at the same time). Both scenarios can exist and will lead to different experimental results, such as different RDCs, and SS-map will easily visualize the structural differences between them.

SS-map is a python algorithm freely available for download in 'http: code.google. com/p/ss-map/' designed to represent the cooperativity or the correlations in secondary structure conformations for IDPs, where the use of contact orders or native contacts is impossible. Although this was the initial purpose it is also applicable to globular proteins, being a useful tool to analyze the folding process of small proteins and peptides. It can shed light into the actual conception of the protein secondary structure.

To visualize the proteins ensembles secondary structure it use the ϕ and ψ angles, whose values are characteristics of different secondary structure elements. The method incorporates four different definitions of the secondary structure elements based on the Ramachandran (ϕ and ψ) diagram. These definitions corresponds to Profasi, Flexible Meccano, Campari and DSSP programs. Additionally one can use its own definition of the ϕ and ψ angles. The program takes either a folder with multiple PDB files (one for each protein in the ensemble) with which it calculates the angles ϕ and ψ, or an array with the angles ϕ and ψ for each structure in the ensemble. It returns either an image, a numpy array and/or a .txt file containing a matrix (or a graphical representation of this matrix) which shows in how many structures of the ensemble (in %) the residue y is forming a structured region of lenght x.

We checked the algorithm against different type of proteins. First we studied

two folded proteins (HPLC-6 and GB1m2) near its melting temperature (323K), using Profasi, to test our SS-map against other traditional visualization techniques. Then, we visualized an ensemble of a MoRF from a Measles and a Sendai virus nucleoprotein whose ensembles were calculated using the Flexible Meccano (to compare with the results of Blackedge and coworkers), and finally we study the existence of the polyproline II (PPII) helix in IDPs using the data provided by Prof.Rohit Pappu (to test our SS-map against his results).

Using SS-map we realized that for HPLC-6 long α-helix segments are not less frequent than shorter ones, contrary to the observed with usual visualization methods. With them was observed that the percentage of α-helix conformation for each peptide gradually decreases with temperature because the α-helices, that used to emerge from a central residue, get shorter with temperature being them most frequent than longer ones. What is observed is that the long helix (spanning 34 or 35 residues) is lost between 310 and 315K, and then the ensemble is composed of helices of several different lengths. There are a non-negligible percentage of α-helix even at 343K. Then, we studied the GB1m2, a structure that forms a β-hairpin, and has a very similar melting temperature, 324 K, respect to the HPLC-6. The SS-map reproduces its structure, two β-strands linked by a central empty region corresponding to a beta turn and additionally we observed that the β-sheets behaves different than the α-helix regarding the adoption of secondary structure with the temperature. Even above the transition temperature, the strands of the hairpin remain the most populated structures, in contrast to the α-helix. Thereafter, analyzing the PPII helices we also realized that these helices opposite to α-helices do not grow from a central residue. It can be concluded that each type of secondary structure present its own pattern and structural characteristics.

Moreover we studied two IDPs corresponding to the Sendai and Measles virus showing that the picture is more complex of that was observed in previous studies of Blackledge and coworkers(*J Am Chem Soc 2008; 130:8055-61*, *Proc Natl Acad Sci U S A 2011; 108:9839-44*). The helices of a couple of MORFs (H1 and H2 in sendai, and H2 and H3 in measles) mix together and form a higher helix.

Summarizing, we created an algorithm that allows the visualization of the protein secondary structure elements showing the cooperative effect of each residue to the whole secondary structure. We analyzed various globular and disordered proteins showing that SS-map can capture information regarding the size, the presence and the behaviour of secondary structure elements (α-helices, β-strands and PPII helices) that traditional visualization methods can not capture. Furthermore we showed that the β strands behave in a opposite way with respect to the α-helices as also happens with PPII. Differences between α-helices, β-strands and PPII regions become more evident using SS-map.

SHORT COMMUNICATION

Intrinsically Disordered Proteins 1:1, e25323; April/May/June 2013; © 2013 Landes Bioscience

SS-map

Visualizing cooperative secondary structure elements in protein ensembles

Jelisa Iglesias, Melchor Sanchez-Martínez, and Ramon Crehuet*

Institute of Advanced Chemistry of Catalunya; CSIC; Barcelona, Spain

Keywords: intrinsically disordered proteins, IUP, ensembles, visualization, secondary structure, NMR, polyproline II

Abbreviations: IDP, intrinsically disordered protein; SS, secondary structure; RDC, residual dipolar couplling; PPII, Polyproline II helix

We present SS-map, a tool to visualize the secondary structure content of ensembles of proteins. When generating ensembles of Intrinsically Disordered Proteins, we lose the understanding a single native structure gives for folded proteins. It then becomes difficult to visualize the composition of the ensembles or to detect transient helices such as MoRFs. Conformational propensities for single residues also hide the nature of cooperative structures. Here we show how SS-map describes folded and unfolded ensembles of some peptides and gives a new view of the ensembles used to describe Intrinsically Disordered Proteins with residual structure in computational and NMR experiments. This tool is implemented in an open-source python code located at code.google.com/p/ss-map

Intrinsically Disordered Proteins (IDPs) exist in solution as ensembles of structures. This raises a challenge to us, humans, as we tend to understand structures by visualizing them,[1] and we lack ways to represent ensembles. Ensembles contain structural information, even when IDPs satisfy random-coil statistics.[2,3] Some regions of IDPs can adopt secondary structures, at least for a transient time.[4] This can be probed with experimental techniques such as NMR, in particular with Residual Dipolar Couplings (RDCs).[5-8] Structured regions, termed MoRFs, are key to recognition processes mediated by coupled folding-binding events.[9] The interpretation of data derived from NMR is usually done by stating that a certain segment of the protein chain adopts a certain secondary structure in a percentage of the total ensemble, but this conveys information in a difficult way for scientists not familiar with these interpretations. How can the ensembles be represented to better unveil their structure?

When studying protein folding, ensembles coming from computations are represented along the reaction coordinate of native contacts. This shows that for many (small) proteins, folding is a two-state process. Thus it is a cooperative event where most of the ensemble at a given temperature is either folded or unfolded. Victor Muñoz has pioneered the study of downhill folders, which fold in a progressive manner.[10] How do MoRFs of IDPs behave? Contact order discriminates between two-state and downhill folders, but it cannot be used in IDPs because it is based on the concept of a well-defined native structure. MoRFs are usually described as the ratio of residues that adopt a certain secondary structure. It is important to differentiate when residues in a fragment *independently* adopt a conformation in a secondary structure region, from when that fragment contains a true secondary structure, with all the residues adopting that conformation *at the same time*, even if that structure is only adopted rarely. Indeed, if n residues are in an α-helical region 20% of the time, that does not mean an helix of n residues is present 20% of the time. Whether this happens or not will lead to different experimental results, such as different RDCs, and SS-map will easily visualize the structural differences of these ensembles.

In this communication we present a way to represent the cooperativity or the correlations in secondary structure formation for IDPs, where the use of contact orders or native contacts is impossible. We named our approach SS-map, from Secondary Structure map. We first study 2-folded proteins near its melting temperature to link our SS-map with other visualization techniques used in the protein folding community. Then, we visualize an ensemble of a MoRF from a measles[11] and a Sendai[5,12] virus nucleoprotein. Finally we reconsider the existence of the polyproline II helix in IDPs.

The SS-map tool is available for download in htcode.google.com/p/ss-map/, under the GNU GPL v3 license. Graphical output from the SS-map is produced with the matplotlib

*Correspondence to: Ramon Crehuet; Email: ramon.crehuet@iqac.csic.es
Submitted: 05/07/13; Revised: 06/03/13; Accepted: 06/08/13
http://dx.doi.org/10.4161/idp.25323
Citation: Iglesias J, Sanchez-Martinez M, Crehuet R. SS-map: Visualizing cooperative secondary structure elements in protein ensembles. Intrinsically Disordered Proteins 2013; 1:e25323; http://dx.doi.org/10.4161/idp.25323

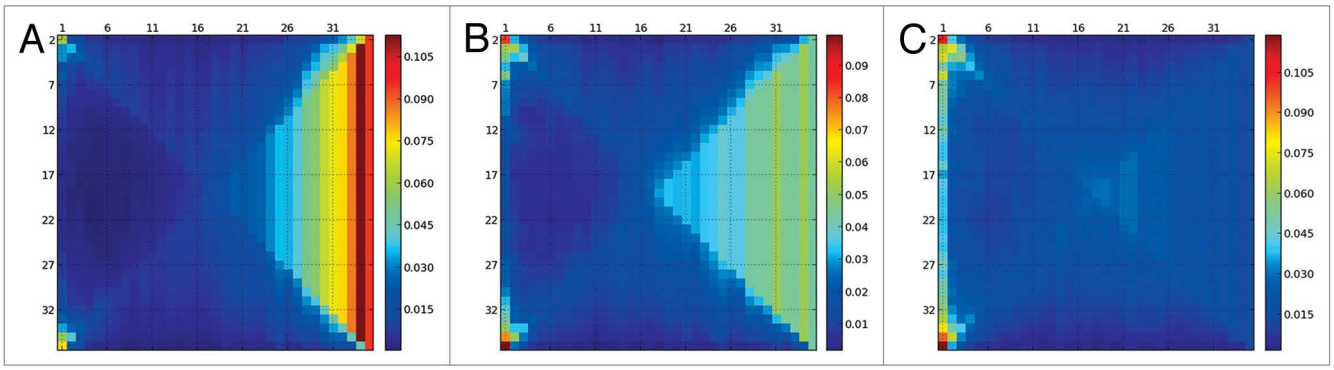


Figure 1. SS-map representing α-helices for the HPLC-6 peptide at different temperatures [(**A**): 313K, (**B**): 320K, (**C**): 327K)]. Large helices are lost below the melting temperature of 323K and all fragments grow from a central residue. At 320K an ensemble of helices with a wide range of lengths is present but shorter helices are not more abundant than longer ones.

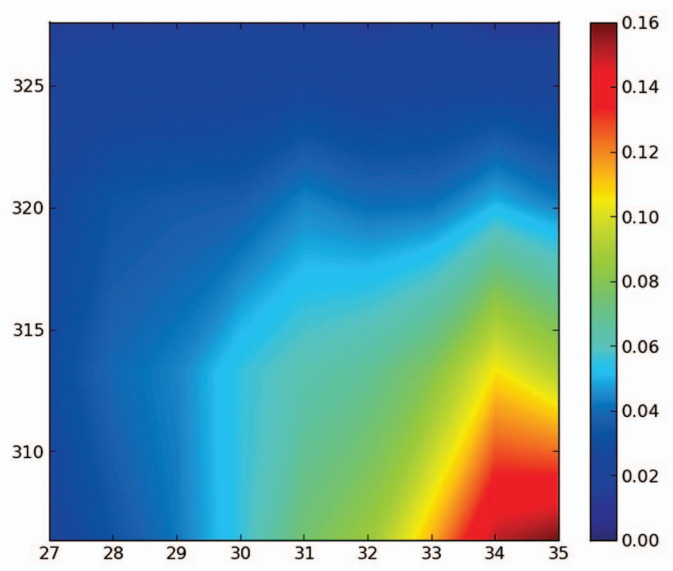


Figure 2. Temperature dependence of the presence of secondary structure elements at different temperatures. The x-axis represents the length of a helix element, and the y-axis the temperature. It shows how long helices are present only at low temperatures, and that helices do not get shorter, they just become much scarcer at higher temperatures (only the region of long helices is plotted, as the remaining region is essentially zero)

library.[13] Details of the simulated ensembles are reported in the Supplemental Material.

The visualization tool presented in this work extends the calculation of secondary-structure percentage per residue one more dimension: we calculate and show the frequency of having *n* exactly contiguous residues in a certain secondary structure. For a protein with N+2 residues, this generates a matrix of *N*x*N*, where an element (*m*,*n*) corresponds to the frequency of having residue *m* forming a secondary structure element of length exactly *n*. Frequencies are normalized, so that if one wants the probability of residue *m* forming an helix of at least 4 residues, one can get it by summing row *m*, elements 4 to N.

There are different definitions of secondary structure elements. Currently our code can use the definition reported in ref. 14, where all the Ramachandran space is assigned to an element; a more restrictive definition as in ref. 15; or a user defined rectangular region of the Ramachandran plot. When the ensemble is input as a set of PDB files, SS-map uses the Bio.PDB[16] module of Biopython[17] to generate dihedral angles. Alternatively, we can use the external code Stride[18] to read the secondary structure. Differences in applying these definitions will be discussed below. A schematic workflow with the different possible input and outputs of SS-map is depicted in **Figure S1**.

The information that SS-map presents requires an image for each of the ensembles. This information can be compressed in two ways to represent several ensembles in one image. The raw-average gives the widely used probability of a certain residue being in the selected conformation, as **Figure S2** shows. The column-average gives new and complementary information: the percentage of fragments of a given length. This information can then be combined for different ensembles, for example, at different temperatures, such as in **Figures 1 and 2**.

We first present a study of the unfolding of the peptide HPLC-6, which forms an α-helix and has a melting temperature of 323K when simulated with the Profasi force field.[19] The percentage of α-helix conformation for each peptide gradually decreases with temperature. This is more prominent at the N- and the C-terminus (**Fig. S2; Fig. S3**). The SS-map shows that at 313K a long helix spanning most of the residues is the most abundant structure (see **Fig. 3**). At 320K, this long helix is lost and fragments of different sizes are almost equally present, but in all cases, these fragments grow from the central residue 19. A representation of secondary structure per residue (**Fig. S2**) suggests that helices get shorter with temperature. This is not true: Long α-helix segments are not less frequent than shorter ones. At 320K, all fragments are rare, and the cumulative percentage of helices larger than 20 residues represents only a 21%. This number, at 313K is of 71%. At 327K, although the overall percentage of α-helix is still 45% (**Fig. S2**), there is no helix as such, only residues that adopt this conformation independently, without any cooperativity. This information cannot be reflected with the visualizations traditionally used, such as **Figure S2**, but it is

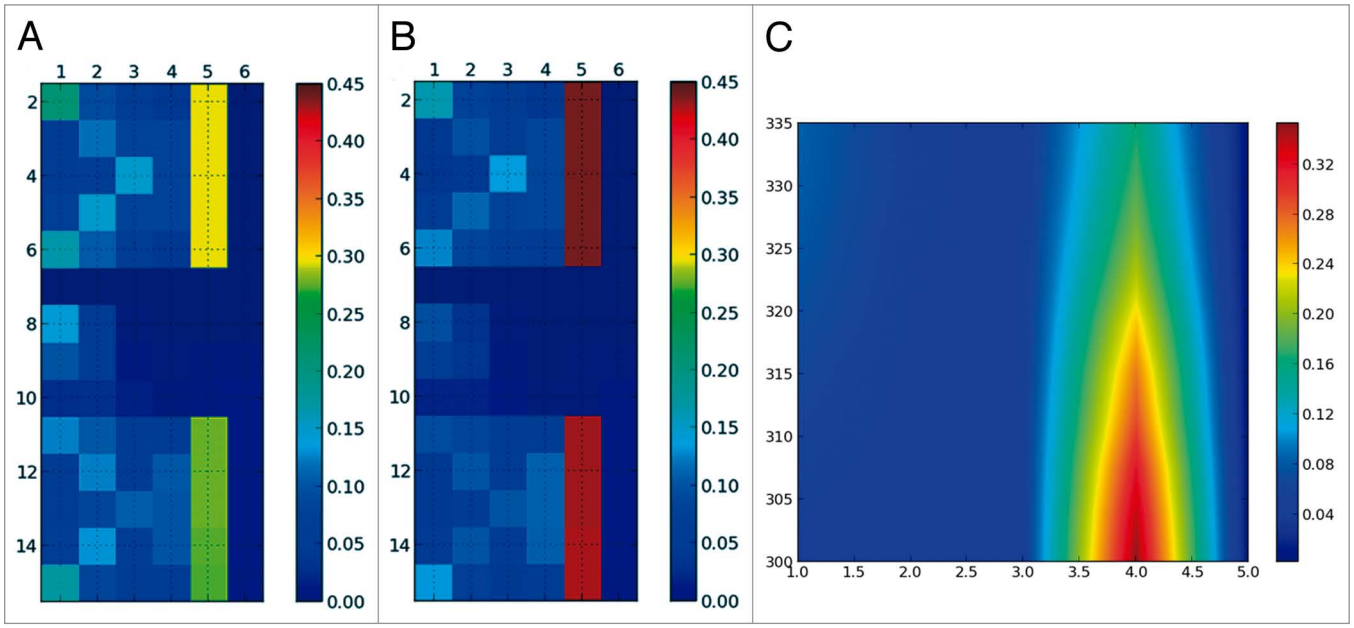


Figure 3. SS-map showing β-strands for the GB1p β-hairpin below the folding temperature [(**A**): 319K and above (**B**): 327K]. The temperature dependence of the SS-map shows that at all temperatures the most frequent strand has 4 residues (**C**).

relevant to interpret the results of circular dicroism that revealed a non-negligible percentage of α-helix even at 343K:[20] our interpretation is that it was only due to isolated residues in α-helix, and not to true helical segments.

The information of a range of ensembles at different temperatures can be compressed as previously explained. **Figure 1** shows that the long helix spanning 34 or 35 residues is lost between 310 and 315K, and then the ensemble is composed of helices of several different lengths. An essentially unfolded ensemble at the melting temperature agrees with recent similar findings for the more complex Protein A.[21]

We now focus on a structure that forms a β-hairpin, i.e., two β-sheets connected by a turn. We have taken a mutated from of the GB1p peptide (GB1m2)[22] also studied with the Profasi force field.[19] The simulated melting temperature for this peptide is very similar to the previous α-helix, 324K. The SS-map shows two β-strands and an empty 4-residue central region, which corresponds to the β-turn (**Fig. 2**). Even above the transition temperature, the strands of the hairpin remain the most populated structures, in contrast to the α-helix. The SS-map shows that the unfolded state of this β-hairpin—ensembles above the folding temperature—has different structural characteristics than the unfolded state of the α-helix (**Fig. 3**; **Fig. 2**). The temperature profile of the SS-map in **Figure 2** also contrasts with the one for the α-helix.

We now focus on a true IDP that contains fragments of partial secondary structure. These fragments are called MoRFs and correspond to binding regions of the IDPs.[9] Partially ordered regions are a challenge for many biophysical techniques,[4] but a successful approach is the use of NMR Residual Dipolar Couplings.[6-8] Here we will consider two proteins: a Measles virus nucleocapsid protein[11] and a Sendai virus nucleoprotein,[5] both studied by Blackledge and coworkers. In both proteins, the authors used a random-coil model named Flexible Meccano[12,23] to generate an ensemble of structures (**Fig. 4**). Then they added helical fragments—in a statistically robust way—until they achieved a satisfactory fit of the RDCs. A special conformational treatment was given to the N-capping residues of the helices. The N-capping modifications are not implemented in the public version of Flexible Meccano, and therefore our ensembles differ from the ones used by Blackledge and collegues (see the SI for a further discussion of this point). **Table 1** describes the composition of both ensembles.

The analysis of the ensemble using SS-map shows that the picture is more complex than it might seem. For example, helix H1 and H2 in the measles virus protein mix together to give an ensemble of helices that have lengths from 5 to 8 residues. Similarly helices H2 and H3 in the Sendai virus protein cannot really be differentiated and extend from the limits stated in **Table 1**. In our ensembles helices extend both toward the N-terminal and the C-terminal sense symmetrically, due to the lack of the N-capping treatment.

SS-map helps to bring light to these features, but as a visualization tool it does not substitute the work to determine what constitutes a correct ensemble. Here we have exploited the statistically sound analysis of Blackledge and coworkers to optimize the ensemble to fit the experimental data and we have only considered their best results.

The presence of polyproline II (PPII) helices in IDPs has been studied in several works. It has been related to the unexpected temperature behavior of IDPs[24] and its content correlates with the net charge of the IDPs[15] because PPII helices are the most stable conformations for charged residues.[25] We have analyzed the simulated ensembles of four IDPs studied by Pappu and colleagues, but here we only report the results for a poly-glutamine of 34 residues (id. 21 in their work[15]) because the results are similar for the other IDPs.

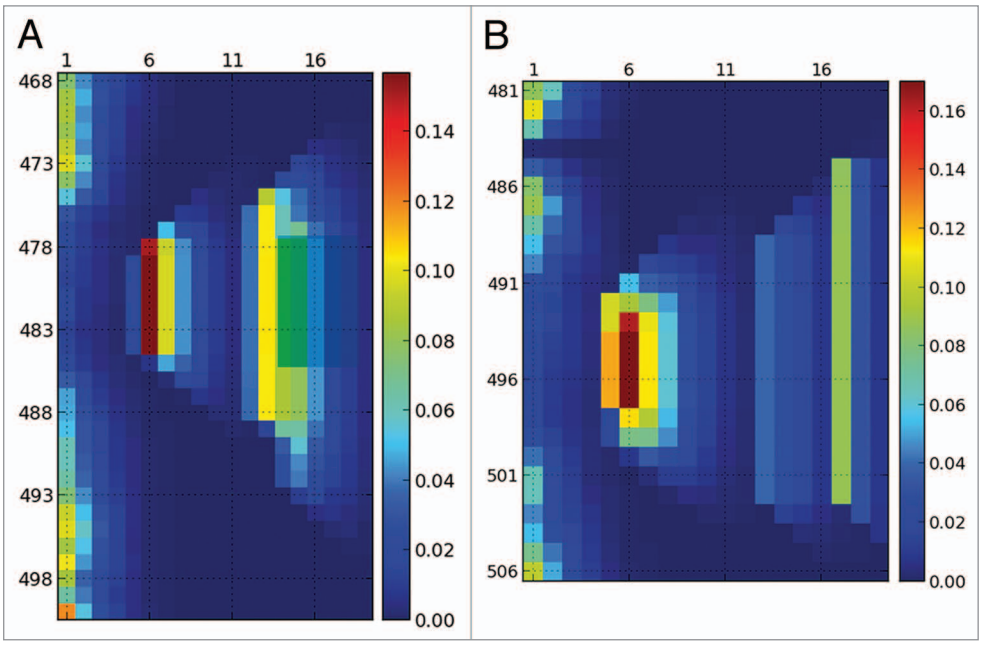


Figure 4. Helical content for the ensembles that reproduces the experimental RDCs of the Sendai virus nucleoprotein[5] (**A**) and the measles virus nucleoprotein[11] (**B**). Both ensembles were generated with Flexible Meccano by mixing ensembles with pre-defined helical content as detained in **Table 1**. Although 3 helices were used for the Sendai protein and 4 for the measles protein, the resulting ensemble is more continuous and mixed than **Table 1** might suggest.

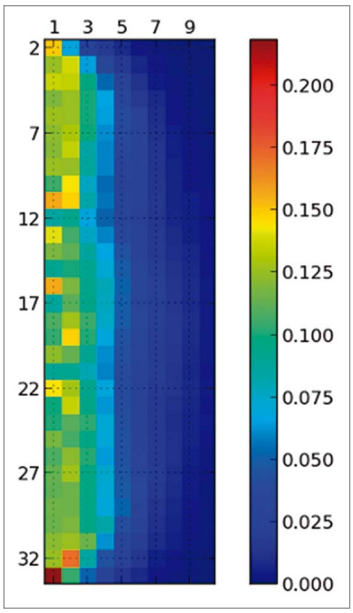


Figure 5. Content of Polyproline II for the 34-residue poly-glutamine, the region defining the polyproline II is the same as in the original study.[15] Although the natural propensity of all the residues is to be in PPII with a relevant frequency, the formation of a helix is not a cooperative process and long helices are absent, in contrast to the α-helices of **Figure 3**.

Table 1. Composition of the ensembles generated with Flexible Meccano[12,23] to reproduce the RDCs for the Sendai virus nucleoprotein[5] and measles virus nucleoprotein,[11] based on the data provided therein

	Residues	Population (%)
	Sendai	
H1	479–484	36
H2	476–488	28
H3	478–492	11
Random coil	468–500	25
	Measles	
H1	494–499	22
H2	492–497	30
H3	489–502	10
H4	485–502	13
Random coil	481–506	25

Remark that the N-capping aminoacids had a special conformational behavior not implemented in the public version of the Flexible Meccano code, and therefore the ensembles reported here differ from those described in the original references.[5,11]

Among all their reported IDPs, this one has the highest PPII content, as expected from its highest charge. Although the total PPII content is 51%, **Figure 5** shows that the longest helices present in the ensembles contain only 5 consecutive residues. To avoid being deceived by single-residue propensities, Pappu and coworkers counted only fragments of 3 or more consecutive residues in PPII conformation. SS-map removes the arbitrariness of that number "3" and conveys more information. As opposed to the α-helix in **Figure 1**, there is no growing helix from any central residue. Thus, long helices of PPII do not cooperatively form in solution, at least

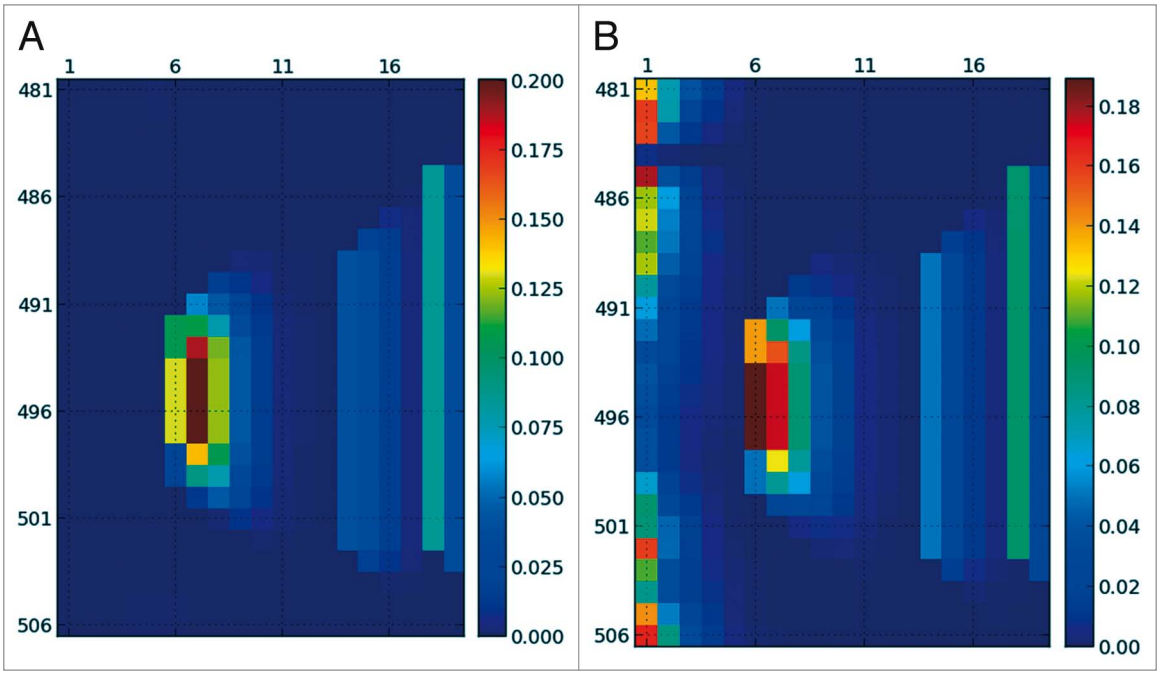


Figure 6. SS-map for the measles virus nucleoprotein showed if **Figure 4** using two different criteria to define the α-helix. The external program Stride (**A**), which only considers a secondary structure element when it is larger than 4 residues, and the definition used in ref. 15 (**B**), which is approximately circular and much smaller than the region used by Blackledge and coworkers.[14]

in the models used by Pappu and coworkers.[15] Considering that electrostatic interactions in water increase with temperature,[26] it would be interesting to study how these ensembles change when heated. We leave that for future work.

Although everybody agrees on the qualitative description of α-helices and β-sheets, different groups partition the Ramachandran plot in different regions. For example, Blackledge and coworkers use big rectangular regions so that any point belongs to a given secondary structure.[14] Although these regions are larger than what is usually accepted, they allow the classification of all points in the Ramachandran plot. Pappu and colleagues use much more restrictive secondary structure elements,[15] closer to more widespread definitions such as the one in the Wikipedia.[27] In SS-map users can also measure with their own definitions. The effect of these arbitrariness could be more important in IDPs than in folded proteins, due precisely to their higher disorder. **Figure 6B** shows the ensembles plotted using different criteria. It is interesting that the Stride program never considers a fragments of less than 4 residues to have a secondary structure, to model as closely as possible how crystallographers represent α-helices and β-strands.[18] Therefore the results differ in those 1 to 3 residue fragments, but agree almost quantitatively in the rest. The more restrictive definitions used by Pappu and coworkers[15] lead to overall lower percentages of secondary structure fragments as expected, but the general picture remains the same (compare **Fig. 6** with **Fig. 5**). Whether a consensus is necessary or not is something the scientific community has to decide, but our present findings suggest that the structural interpretations do not change significantly with varying definitions.

Understanding IDPs with partially folded regions is a challenge to both computation and experiment.[4] Conformations cannot be referenced or compared with a native structure and we need new tools to visualize these heterogeneous ensembles. In this work we presented a tool, SS-map, which literally adds a new dimension to the representation of IDPs ensembles. By including the correlation between secondary structure elements in fragments, a more detailed picture emerges. Differences between α-helices, β-strands and PPII regions become more evident. The ensembles used to reproduces RDCs data can also be visualized and compared. SS-map does not optimize or change the ensembles whatsoever, it only extracts information from them and displays it. The results are as realistic as the underlying ensemble is; finding these ensembles remains a challenge.[28] Finally, this tool can also be useful to analyze the folding process of small proteins and peptides.[29]

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We would like to thank Rohit Pappu for kindly sharing his data on the study of polyglutamines[19] and Martin Blackledge for helpful comments. We acknowledge financial support from the Ministerio de Innovación y Competitividad (CTQ2012-33324) and the Generalitat de Catalunya (2009SGR01472). M.S-M. thanks the Ministerio de Economia y Competitividad for a predoctoral fellowship.

Supplemental Material

Supplemental material may be found here:
http://www.landesbioscience.com/journals/idp/article/25323/

References

1. Gan J, Norman C. 2012 Visualization Challenge. Science 2013; 339:509; http://dx.doi.org/10.1126/science.339.6119.509
2. Fitzkee NC, Rose GD. Reassessing random-coil statistics in unfolded proteins. Proc Natl Acad Sci U S A 2004; 101:12497-502; PMID:15314216; http://dx.doi.org/10.1073/pnas.0404236101
3. Jha AK, Colubri A, Freed KF, Sosnick TR. Statistical coil model of the unfolded state: resolving the reconciliation problem. Proc Natl Acad Sci U S A 2005; 102:13099-104; PMID:16131545; http://dx.doi.org/10.1073/pnas.0506078102
4. Dyson HJ. Expanding the proteome: disordered and alternatively folded proteins. Q Rev Biophys 2011; 44:467-518; PMID:21729349; http://dx.doi.org/10.1017/S0033583511000060
5. Jensen MR, Houben K, Lescop E, Blanchard L, Ruigrok RWH, Blackledge M. Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. J Am Chem Soc 2008; 130:8055-61; PMID:18507376; http://dx.doi.org/10.1021/ja801332d
6. Schneider R, Huang JR, Yao M, Communie G, Ozenne V, Mollica L, et al. Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. Mol Biosyst 2012; 8:58-68; PMID:21874206; http://dx.doi.org/10.1039/c1mb05291h
7. Jensen MR, Markwick PRL, Meier S, Griesinger C, Zweckstetter M, Grzesiek S, et al. Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. Structure 2009; 17:1169-85; PMID:19748338; http://dx.doi.org/10.1016/j.str.2009.08.001
8. Marsh JA, Neale C, Jack FE, Choy WY, Lee AY, Crowhurst KA, et al. Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure. J Mol Biol 2007; 367:1494-510; PMID:17320108; http://dx.doi.org/10.1016/j.jmb.2007.01.038
9. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, et al. Analysis of molecular recognition features (MoRFs). J Mol Biol 2006; 362:1043-59; PMID:16935303; http://dx.doi.org/10.1016/j.jmb.2006.07.087
10. Garcia-Mira MM, Sadqi M, Fischer N, Sanchez-Ruiz JM, Muñoz V. Experimental identification of downhill protein folding. Science 2002; 298:2191-5; PMID:12481137; http://dx.doi.org/10.1126/science.1077809
11. Jensen MR, Communie G, Ribeiro EA Jr., Martinez N, Desfosses A, Salmon L, et al. Intrinsic disorder in measles virus nucleocapsids. Proc Natl Acad Sci U S A 2011; 108:9839-44; PMID:21613569; http://dx.doi.org/10.1073/pnas.1103270108
12. Bernadó P, Blanchard L, Timmins P, Marion D, Ruigrok RWH, Blackledge M. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. Proc Natl Acad Sci U S A 2005; 102:17002-7; PMID:16284250; http://dx.doi.org/10.1073/pnas.0506202102
13. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng 2007; 9:90-5; http://dx.doi.org/10.1109/MCSE.2007.55
14. Nodet G, Salmon L, Ozenne V, Meier S, Jensen MR, Blackledge M. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. J Am Chem Soc 2009; 131:17908-18; PMID:19908838; http://dx.doi.org/10.1021/ja9069024
15. Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc Natl Acad Sci U S A 2010; 107:8183-8; PMID:20404210; http://dx.doi.org/10.1073/pnas.0911107107
16. Hamelryck T, Manderick B. PDB file parser and structure class implemented in Python. Bioinformatics 2003; 19:2308-10; PMID:14630660; http://dx.doi.org/10.1093/bioinformatics/btg299
17. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009; 25:1422-3; PMID:19304878; http://dx.doi.org/10.1093/bioinformatics/btp163
18. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995; 23:566-79; PMID:8749853; http://dx.doi.org/10.1002/prot.340230412
19. Irbäck A, Mitternacht S, Mohanty S. An effective all-atom potential for proteins. PMC Biophys 2009; 2:2; PMID:19356242; http://dx.doi.org/10.1186/1757-5036-2-2
20. Chakrabartty A, Ananthanarayanan VS, Hew CL. Structure-function relationships in a winter flounder antifreeze polypeptide. I. Stabilization of an α-helical antifreeze polypeptide by charged-group and hydrophobic interactions. J Biol Chem 1989; 264:11307-12; PMID:2738067
21. Maisuradze GG, Liwo A, Ołdziej S, Scheraga HA. Evidence, from simulations, of a single state with residual native structure at the thermal denaturation midpoint of a small globular protein. J Am Chem Soc 2010; 132:9444-52; PMID:20568747; http://dx.doi.org/10.1021/ja1031503
22. Fesinmeyer RM, Hudson FM, Andersen NH. Enhanced hairpin stability through loop design: the case of the protein G B1 domain hairpin. J Am Chem Soc 2004; 126:7238-43; PMID:15186161; http://dx.doi.org/10.1021/ja0379520
23. Ozenne V, Bauer F, Salmon L, Huang JR, Jensen MR, Segard S, et al. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics 2012; 28:1463-70; PMID:22613562; http://dx.doi.org/10.1093/bioinformatics/bts172
24. Kjaergaard M, Nørholm AB, Hendus-Altenburger R, Pedersen SF, Poulsen FM, Kragelund BB. Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II? Protein Sci 2010; 19:1555-64; PMID:20556825; http://dx.doi.org/10.1002/pro.435
25. Krimm S, Mark JE. Conformations of polypeptides with ionized side chains of equal length. Proc Natl Acad Sci U S A 1968; 60:1122-9; PMID:16591670; http://dx.doi.org/10.1073/pnas.60.4.1122
26. Thomas AS, Elcock AH. Molecular simulations suggest protein salt bridges are uniquely suited to life at high temperatures. J Am Chem Soc 2004; 126:2208-14; PMID:14971956; http://dx.doi.org/10.1021/ja039159c
27. Wikipedia. Ramachandran Plot, http://en.wikipedia.org/wiki/Ramachandran_plot
28. Fisher CK, Stultz CM. Constructing ensembles for intrinsically disordered proteins. Curr Opin Struct Biol 2011; 21:426-31; PMID:21530234; http://dx.doi.org/10.1016/j.sbi.2011.04.001
29. Irbäck A, Mohanty S. PROFASI: A Monte Carlo simulation package for protein folding and aggregation. J Comput Chem 2006; 27:1548-55; PMID:16847934; http://dx.doi.org/10.1002/jcc.20452

4.2.2 Determination of IDPs ensembles from Residual Dipolar Couplings

A usual problem studying protein dynamics by conformational ensembles is that when we simulate a system and we want to compare against experimental data to validate it, the simulated ensembles does not match accurately the experimental values. The reason why this happens is not unique, can have more than one answer. For instance, it can be a problem of the force-field employed or a problem of sampling due to computing time limitations. As we have explained at the Introduction a lot of efforts are putted into improving the force-fields parameterizing them with experimental data. But there are also other methods that are into development aimed to avoid this mismatch such as reweighting the calculated ensembles to match certain observables.

Trying to shed some light in this direction we created an algorithm (MaxEnt) that incorporates the maximum entropy principle to fit a set of RDCs from a simulated ensemble. The maximum entropy is a statistical method that derives from minimizing the information included in an ensemble to fit certain observables. On the other hand, the usage of RDCs is because it is a very suitable technique to characterize protein secondary structure elements at a residue level and has been widely applied to study IDPs (For instance *J. Am. Chem. Soc., 2010, 132 (24), pp 8407-8418* or *Mol. BioSyst., 2012, 8, 58-68*).

We made MaxEnt, a python algorithm freely available at github ('https:// github.com/ MelchorSanchez/MaxEnt'). It re-weights a set of RDCs values (back-calculated from a set of structures) to fit a given set of Residual Dipolar Couplings values. MaxEnt needs as inputs a matrix of MxN and a vector of N values. The MxN matrix should contain the N RDCs of the M structures in the ensembles, and the second vector should contain the reference RDCs. The RDCs to optimize will be scaled to fit the reference RDCs.

Apart of the Maximum Entropy methodology described at section 3.4.5.4, an important point of our MaxEnt is the calculation of λ, that represents the gradient fit according to that the ensembles are reweighted. The RDCs have to be defined up to a proportionality constant α. Thus the weights in $\langle q_i \rangle = \sum_j^N w^j q_i^j$ need not be normalized. If we know a set of measured RDCs $\mathbf{Q} = \mathbf{Q}_i$, we define the function:

$$f_1(\lambda) = max\big(\frac{1}{M}\|\alpha\langle \mathbf{q}\rangle - \mathbf{Q}\|^2, t^2\big) \tag{4.1}$$

to be minimized. t is a threshold value that is determined by the experimental or the synthetic precision. In that region f_1 is constant. The value of λ can be obtained analytically minimizing $f_1(\lambda)$ which gives:

$$\lambda = \frac{|\langle \mathbf{q} \rangle \cdot \mathbf{Q}|}{\langle \mathbf{q} \rangle \langle \cdot \mathbf{q} \rangle} \tag{4.2}$$

Although with some RDCs is not need to normalize the weights, for clarity and consistency among all types of RDCs we scale them so that they add up to the number of structures. $w = 1$ is equivalent to a structure not being reweighted. Because the scaling adds one degree of freedom, the set of $\lambda = \lambda_i$ that minimize f_1 lies on a 1-dimensional curve. Based on the MaxEnt principle, we seek λ that minimally modifies the ensemble. By $w^j = \sum_i^M exp(\lambda_i q_i^j)$ these are the *lambda* as close as possible to 0. Therefore we add a penalty term:

$$f_2(\lambda) = \frac{k}{M} ||\lambda||^2 \tag{4.3}$$

and minimize $f = f_1 + f_2$. The value of the new introduced parameter is only determined by the user-defined threshold t. If k is large, f_2 will dominate and will force low λ that will result in f_1 higher than the threshold. Once k is small enough, f_1 reaches the threshold and further reduction of k results in the same optimal . Therefore the selection of k is done by the algorithm.

To simplify the generation of the MxN RDCs we also created the python script RunPales, available in the same github folder as MaxEnt. It is a python 3 script that can call the PALES (a software to backcalculate RDCs) executable, generating the corresponding RDCs and converting it into an array. In other words, RunPales is nothing but an interface to call the PALES program with the suitable options, and store the generated results.

Our implementation of the maximum entropy principle present some interesting points. First of all we modified its implementation making it scale invariant to work with RDCs. Then, the MaxEnt can be used by different experimental groups using different ensembles, as it can use any given set of structures. Its only relies on the RDCs values not in the way in that are generated. Thus although we provide the RunPales, its use is not necessary. Also due to its independence of the structures we avoid the risk of overfitting problems as the number of parameters is based only on the number of experimental data. Finally another important characteristic is its velocity: it can use thousands of structures and converges in a few seconds.

To check the algorithm we tested it over synthetic and experimental data sets of the Sendai virus nucleoprotein. The calculated data set were obtained from REMC simulations using the Profasi and the Campari (coarse grained) physics-based force fields. Our results showed that despite their limitations, both can generate better ensemble than simple random-coil methods (widely employed to

simulate IDPs in general and the Sendai virus nucleoprotein in particular), showing a more predictive power. However the generated ensembles has to be be slightly modified to fit the reference data to good accuracy. Thus it is necessary the use of reference (usually experimental) data to improve these ensembles. But this data is insufficient to fully determine a good representative ensemble, and thus the pervasive influence of the force field cannot be overlooked, if we wish to have consistent representations of IDPs ensembles. If the force-field does not correctly represent the real structural ensemble, although the ensemble will be reweighted to a stronger extent and correctly fits the 'reference' data, its composition is going to change only slightly.

Summarizing, we devised the suitable implementation of MaxEnt to fit RDCs data sets in a fast way, avoiding overfitting and being scale invariant (that make the model independent and potentially applicable to any kind of observable not only RDCs). Also, we highlighted the relevance of the underlying model, which sometimes is understimated. And finally, we tested whether some coarse grained methods (Profasi and Campari) could produce more accurate ensembles than random-coil-based Force Fields and thus increase the prediction of RDCs.

PCCP

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c4cp03114h

Application of the maximum entropy principle to determine ensembles of intrinsically disordered proteins from residual dipolar couplings†

M. Sanchez-Martinez and R. Crehuet*

Received 15th July 2014,
Accepted 10th October 2014

DOI: 10.1039/c4cp03114h

www.rsc.org/pccp

We present a method based on the maximum entropy principle that can re-weight an ensemble of protein structures based on data from residual dipolar couplings (RDCs). The RDCs of intrinsically disordered proteins (IDPs) provide information on the secondary structure elements present in an ensemble; however even two sets of RDCs are not enough to fully determine the distribution of conformations, and the force field used to generate the structures has a pervasive influence on the refined ensemble. Two physics-based coarse-grained force fields, Profasi and Campari, are able to predict the secondary structure elements present in an IDP, but even after including the RDC data, the re-weighted ensembles differ between both force fields. Thus the spread of IDP ensembles highlights the need for better force fields. We distribute our algorithm in an open-source Python code.

Introduction

Intrinsically disordered proteins (IDPs) are an emerging family of proteins characterized by their ability to adopt a vast number of configurations in solution. Their role in cell signalling, transcription and aggregation turns them into key proteins in cancer and neurodegenerative diseases.[1,2] One would expect many of them to be drug targets; however very few studies have addressed the interaction of IDPs with small molecules.[1,3] One reason for this is the difficulty in both generating and characterizing the ensemble of configurations that turn an IDP functional.[4] A common mechanism of IDPs is a folding transition upon binding partner proteins.[5] The amount of secondary structure elements in the unbound IDPs governs the kinetics of this binding process,[6] thus the need to understand IDP secondary structure elements in solution. These regions are also called MoRFs[7,8] and many studies aim at their identification.

A very suitable technique to characterize the secondary structure at a residue level is the NMR residual dipolar couplings (RDCs),[9] a technique that has been thoroughly developed by Blackledge[10–13] and Forman-Kay[14–17] groups, among others. In an isotropic medium, such as liquid water, dipolar couplings average out to zero. But if the medium has some preferential directions, then there is a partial alignment of the molecules and a residual coupling can be measured.

Contrary to what is the case for folded proteins, in IDPs the alignment tensor is essentially determined by the local (secondary) structure.[16] When the main mechanism of alignment is steric, repulsion between the protein and the alignment medium tends to align secondary structure elements parallel to the medium. For this reason N–H couplings convey important information on the secondary structure. When the alignment medium is parallel to the field they are positive in α-helices – as all N–H are parallel to the helix – negative in β-sheets – as N–H are perpendicular to the sheet – and are very low for regions without any secondary structure, where residue orientations are random. A qualitative interpretation of RDCs can be based on these principles, but a quantitative explanation can be achieved if one is able to generate an ensemble of configurations that reproduce the measured RDCs.[11,12,15,16]

The generation of the ensemble that fit the RDCs is the crux of several approximations used in this field.[18] A common approach is to sample random coil regions of the Ramachandran plot with codes such as Flexible Meccano,[13,19] TraDES,[16,20] or BEGR[21] and then introduce secondary structure regions and weight them with a statistical analysis[11,17] or a genetic algorithm.[13] This is because the physics behind these force fields is very simple and cannot predict secondary or tertiary structure. These methods have proved extremely successful in interpreting several IDP studies, but lack predictive value in terms of secondary structure elements.

The problem of optimizing an ensemble is a case of inferential structure determination,[22] albeit with a much broader probability distribution. If this distribution comes from a simulation, we would like to modify it so that it agrees with

Institute of Advanced Chemistry of Catalunya (IQAC), CSIC, Spain.
E-mail: ramon.crehuet@iqac.csic.es

† Electronic supplementary information (ESI) available: Details of the PALES calculation and Fig. S1 to S9. See DOI: 10.1039/c4cp03114h

the experimental data. Ideally, the inclusion of the experimental data should create ensembles that agree among themselves, even if coming from different simulation methods. Here we explore to which extent this is true.

We present a method based on the maximum entropy principle (MaxEnt) to fit RDC data to simulated ensembles. Maximum entropy is a logically consistent way to fit a distribution to previously known values introducing the minimum possible modifications.[23,24] It has been advocated very recently as a powerful technique to solve structural problems[25] and it has already been applied to SAXS ensemble determination.[26]

We generated our ensembles from two coarse-grained force fields, which have more accurate physical terms than TRaDES or Flexible Meccano while remaining computationally affordable. Coarse-grained methods allow sampling of the large conformational space essential to describe IDPs and converge RDC data. However the simulation force-field does not influence the validity of the presented selection procedure, which can be applied to all types of ensembles.

Our aim of this work is three-fold. First, we develop a fitting algorithm to adjust experimental RDCs to an ensemble of conformations. We implement our method in a publicly available code so that it can be compared to others, and can be used by any research group.[27] Second, we explore the information content of RDC data and the influence of our force field; in other words, how much do the RDCs constrain the initial ensemble. Considerable efforts have been made to determine how much different experimental data determine the properties of the ensembles.[17] Here we want to highlight the relevance of the underlying model, which is often overlooked. And third, we test whether some coarse grained methods can produce more accurate ensembles than random-coil-based force fields and thus increase the prediction of RDCs.

Methods

The maximum entropy (MaxEnt) principle derives from minimizing the information included in an ensemble to fit certain observables. It was first introduced by Jaynes[23] and was recently applied as a way to constrain molecular dynamics on-the-fly.[28,29] Roux and co-workers showed that under certain circumstances, their results were equivalent to the more traditional constrains with harmonic potentials, used also in molecular dynamics,[30] while Vendruscolo and co-workers showed that the restraint strength can be related to the experimental error.[31] Here we present the application of the MaxEnt to the *a posteriori* re-weighting of an ensemble that has already been calculated. We also add some modifications needed to treat RDC data.

We decided to implement an *a posteriori* re-weighting so that our method could be applied to ensembles generated with any software or force field. A second reason is that when applying the constraints on-the-fly, one usually averages by the number of replicas running in parallel[32,33] but the number of replicas needed to converge the RDC values for IDPs is of the order of thousands (see Results section), which means that constraint molecular dynamics could only be run in supercomputers.

In our *a posteriori* re-weighting we assume we have a set of N structures $\{\mathbf{X}_{j=1,N}\}$ that we have previously calculated with a Monte Carlo or molecular dynamics simulation. As such, they have already been generated with a probability proportional to their Boltzmann factor, which depends on each specific force field. For a set of M observables $\mathbf{q} = \{q_{i=1,M}\}$, Pitera and Chodera showed that the application of the MaxEnt principle resulted in a reweighting of the probability of each structure j by a term:[28]

$$w^j = \sum_i^M \exp\left(\lambda_i q_i^j\right) \tag{1}$$

The form of the reweighting is fixed and a single parameter λ_i applied to each observable. As each structure has already been generated with a weight according to a given ensemble (a Boltzmann factor in *NVT*), w^j modifies the weight of the structure to fit the experimental observables. q_i^j represents the value of observable i in the structure $\mathbf{X}_j$. $\mathbf{q}$ is a matrix of dimension $M \times N$. The average value of observable q_i for a given reweighting is

$$\langle q_i \rangle = \sum_j^N w^j q_i^j \tag{2}$$

RDCs have the peculiarity that they can only be defined up to a proportionality constant α, because their absolute value depends on their degree of alignment, which cannot be measured. This has two consequences. First the weights in eqn (2) need not be normalized, and second, one cannot define a simple convex objective function as Pitera and Chodera did.[28] If we know a set of measured RDCs $\boldsymbol{Q} = \{Q_i\}$, we define the function

$$f_1(\boldsymbol{\lambda}) = \max\left(\frac{1}{M}\|\alpha\langle \boldsymbol{q}\rangle - \boldsymbol{Q}\|^2, t^2\right) \tag{3}$$

to be minimized. t is a threshold value that is determined by the experimental precision, and there is no point in optimizing below that threshold, so f_1 is constant in that region. In the case of experimental RDCs, we chose the value of 1 Hz. The value of α can be obtained analytically by minimizing $f_1(\boldsymbol{\lambda})$ which gives

$$\alpha = \frac{|\langle \mathbf{q}\rangle \cdot \mathbf{Q}|}{\langle \mathbf{q}\rangle \cdot \langle \mathbf{q}\rangle} \tag{4}$$

When using N–H and Cα–Hα sets of RDCs a common scaling factor was used.[34] Because of the scaling, the weights need not be normalized, but for the sake of clarity in the figures and in the main text we scale the weights so that they add up to the number of structures, so that a weight equal to 1 is equivalent to a structure not being reweighted.

Because the scaling adds one degree of freedom, the set of $\boldsymbol{\lambda} = \{\lambda_i\}$ that minimize f_1 lies on a 1-dimensional curve. Based on the MaxEnt principle, we seek $\boldsymbol{\lambda}$ that minimally modifies the ensemble. By eqn (1) these are the $\boldsymbol{\lambda}$ as close as possible to 0. Therefore we add a penalty term:

$$f_2(\boldsymbol{\lambda}) = \frac{k}{M}\|\lambda\|^2 \tag{5}$$

and minimize $f = f_1 + f_2$. Although we are introducing a new parameter, its value is only determined by the user-defined

threshold t. If k is large, f_2 will dominate and will force low λ that will result in f_1 higher than the threshold. Once k is small enough, f_1 reaches the threshold and further reduction of k results in the same optimal λ (Fig. S1, ESI†). Therefore the selection of k is done by the algorithm. The lack of sensitivity to k is an important difference with restrained dynamics where its choice is highly non-trivial.[25,30,31] The minimization of f is done with the Newton-GC method implemented in SciPy.[35] For that, the analytic gradient is required. Its expression is deduced in the Appendix.

Our implementation converges in less than 10 seconds for the ensembles used in this work in a 1 processor Xeon machine. This is to be compared with the Bayesian method developed by Stultz,[36,37] which being their most efficient method takes about 30 minutes in an 8 processor Xeon machine with an ensemble of 299 structures. At the time of writing this paper, Das *et al.*[38] published an interesting paper with a full Bayesian approach (called FitEnsemble) based on Monte Carlo sampling and implemented in pyMC.[39] In the results section we compare our method with theirs and we show that the full Bayesian approach does not convey any essentially new information. At present their method cannot deal with scale-invariant quantities such as RDCs, but we do not see any fundamental reason why it could not be extended to treat them and we plan to explore this possibility. That would allow a cleaner way to introduce the uncertainty of RDCs' prediction and the experimental error, which are cumbersome to include in a maximum entropy formalism[40] in an *ad hoc* manner. As the comparison with FitEnsemble[38] will show, both of these terms are small for RDCs and the MaxEnt principle results in a fast algorithm. The extension of generative probabilistic models[40,41] or maximum likelihood approaches[42] to IDPs is also an attractive alternative, but it is beyond the scope of this work to evaluate them. The MaxEnt principle gives results in agreement with the Sparse Ensemble Selection algorithm,[43] but the latter is computationally more expensive and needs some further development to be applicable to IDPs.[43]

Data

As N–H RDCs are the most discussed RDCs for IDPs we focus on these data, but we also explore the additional information carried by Cα–Hα RDCs. We use two kinds of data. First, we test our method with synthetic data, as that allows comparisons with the exact result. Then we apply the method to experimental RDCs to see how it performs. In both cases we use a 53 residue sequence from the nucleocapsid-binding domain of Sendai virus phosphoprotein. This protein has a crucial role in the replication and transcription of the negative strand RNA genome.[11,44] The N-terminal domain of this protein is unstructured but contains some partial secondary structure. The sequence of the simulated fragment is FVTLHGAERLEEETNDEDVSDIERRIAMRLAERRQEDSATHGDEGRNNGVDHE (the charges at the end of the sequence were removed as it is part of a larger protein). This fragment corresponds to the residue numbering 458 to 510 in ref. 11. We have analysed only this region as it contains secondary structure elements[11,44] that cannot be predicted with a simple force field such as Flexible Meccano.

Synthetic data

We run a parallel tempering simulation using the Profasi force field[45,46] in the Profasi code[47] with 16 replicas, from 270 to 330 K.

We take T_1 = 325.6 K as our reference or "experimental" ensemble. We calculated the RDCs for 8000 uncorrelated structures with PALES[48] using steric alignment, because of the NMR setup used (see ESI† for the PALES options used). Then, we have used the ensembles of structures at T_0 = 317.0 K to fit the RDC data at T_1.

Because we have simulated both ensembles, we know that the weight of a given structure j with energy E_j from the T_1-ensemble at temperature T_0 is given by the Boltzmann factor, namely

$$w^j_{\text{Boltzmann}} \propto \exp\left(-\left(\frac{1}{T_1}-\frac{1}{T_0}\right)E_i\right) \tag{6}$$

And this can be compared with the reweighting of our MaxEnt algorithm based on the RDCs.

Experimental data

The experimental data for this study were obtained from the work of Blackledge and co-workers.[11] In their study they measured N–H and Cα–Hα RDCs and made a statistical analysis to evaluate which regions of the α-helix needed to be added to explain the observed results. When comparing with experimental data, our residue number 1 corresponds to residue number 458 in ref. 11. In this region, 31 N–H RDCs and 25 Cα–Hα RDCs were measured. RDCs for the 11 terminal residues are not calculated nor taken into account for the fit side to eliminate the boundary effects in the RDCs.[49,50]

The most interesting part corresponds to residues 18 to 34, because of their tendency to form partial α-helices, also known as MoRFs.[7,8]

These data have been simulated with two different coarse-grained force fields: Profasi[45–47] and Campari.[51] Profasi was chosen for its focus on reproducing the folding behaviour of proteins based on physical terms. We think that using a physics-based force field is important to work with IDPs as knowledge-based force fields are biased towards folded proteins. Profasi has also been applied to IDPs.[52,53] The choice of Campari is justified because it was specifically designed to work with IDPs and has been applied in several studies.[51,54] The Campari system contains 9 sodium ions to neutralize the charge.

The RDCs were calculated from the PDBs with the PALES software.[48] As the alignment media, poly(ethylene glycol), is dominated by steric interactions, we used the steric alignment in PALES (see the ESI† for further details).

Data and code availability

The Profasi and Campari ensembles re-weighted to fit the experimental data have been deposited in the Protein Ensemble Database (pE-DB)[55] with the code 4AAB. Because the pE-DB does not support weighted ensembles, the deposited structures are those structures with weights larger than 0.75 (see below).

Cross-validation

We have performed two types of cross-validation. First, we use experimental N–H RDCs as a training set and leave the experimental Cα–Hα as the test set. Second, we use a set of 10 000 structures as a test set and use a variable number of structures in the training set. We tried the following sizes for the training set: {100, 250, 500, 750, 1000, 2500, 5000, 7500, 10 000}. When using smaller sets, MaxEnt could not converge to the requested accuracy in the training set. Note that the training set is not a subset of the test set, and in the final case, we have a total of 20 000 structures. We compare the error in the fit in the test set with the $\boldsymbol{\lambda} = \{\lambda_i\}$ and the scale factor coming from the training set with respect to the error in that training set. This procedure can tell us the adequate size of the training set and an estimation of the error.

Results and discussion

Size of the ensemble and error estimation

The number of molecules in an NMR experiment is orders of magnitude larger than what can be simulated. How many structures should an ensemble contain? We seek the minimum number of structures needed so that when we add more structures to the ensemble (sampling from the probability distribution given by our force field) the results do not change appreciably.[56] This depends on both the property we measure and the shape of the probability distribution of the ensemble. For example, for several folded proteins, a single structure can reproduce a SAXS curve or a diffraction pattern.

Fig. 1 shows the error in the test set when using different number of structures for the training set to fit N–H RDCs with the Campari ensemble. We can see that for training sets smaller than several thousands, the errors in the test set remain very large, and increase as we improve the fit in the training set. In other words, the optimized $\{\lambda_i\}$ are not transferable. This shows us that we need training sets at least of 7000 structures to determine parameters that do not overfit the experimental results until an RMS error of approximately 1 Hz. Because this number is close to the experimental error, we consider ensemble sizes of 7000–10 000 as adequate.

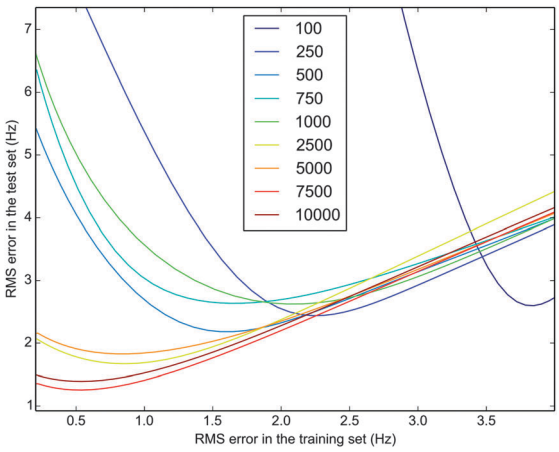


Fig. 1 Plot of the root mean square (RMS) error allowed when fitting the training set with respect to the error in the test set. The test set is always of 10 000 structures whereas the training set increases from 100 to 10 000 structures. Results seem converged above 7500 structures and trying to fit below 1 Hz results in overfitting even for the largest ensembles.

Alternatively, we can estimate the error when calculating the mean value for an RDC: the standard error of the mean. There is certain ambiguity in this value as RDCs can be scaled, but we take here a fixed scale factor obtained from the fit of the 10 000 structures (α = 2.08). Fig. S2 (ESI†) agrees with our conclusion that several thousands of structures are needed to get a mean RDC value of the same order of the experimental error. This result is independent of the residue we are measuring: the convergence of all RDCs is the same. Other studies have also found that the underlying ensembles are more heterogeneous than what the measured mean value may suggest.[56–58]

Several previous studies used a smaller ensemble size[31,32] to successfully simulate IDPs. The size of the ensemble in these MD restrained simulations depends not only on the dispersion of the measured property but also on the other parameters used for the restrain, namely its force constant.[30,31] These studies run simulations in parallel and were limited by computational resources, but formally their results are exact only when the number of replicas tends to infinity. Other computational methods are expensive, thus limiting the size of the ensembles.[4,14,15,17,37] Our method is efficient for thousands of structures so that we prefer to use the full simulated ensemble.

A second important reason to limit the size of the ensembles is to reduce the overfitting. This is an issue when the weights of the structures are the parameters to be optimized, because new structures introduce new parameters, with the obvious risk of overfitting. With the MaxEnt algorithm, the number of parameters is fixed by the number of experimental data and not by the number of structures in the ensemble, which again does not prevent the use of large ensembles.

Synthetic data. What are the RDCs re-weighting?

In this section we analyse to which extent the MaxEnt can recover an unknown ensemble, using some experimental data from that ensemble.

To analyse the secondary structure (SS) content of the ensemble, we use SS-map.[59] SS-map is a software that plots the SS fraction of a given residue on the y axis and the length of the SS element on the x axis, thus providing a picture of the SS distribution of an ensemble with the information of the cooperativity of different SS of individual residues. By plotting both the fraction of SS and its length, it allows to distinguish, for example, a fully formed helix of 10 residues present 50% of the time from 2 fragments of 5 residues spanning the same range.

The ensemble at T_1 represents what in a real situation would be the unknown ensemble, from which we only know the measured RDCs. T_0 is a calculated ensemble that presumably will be similar, but does not have to reproduce the data exactly. MaxEnt should be able to reweight the T_0-ensemble so that it fits the "measured" RDCs. Will the T_0 re-weighted ensemble be more similar to the T_1 ensemble?

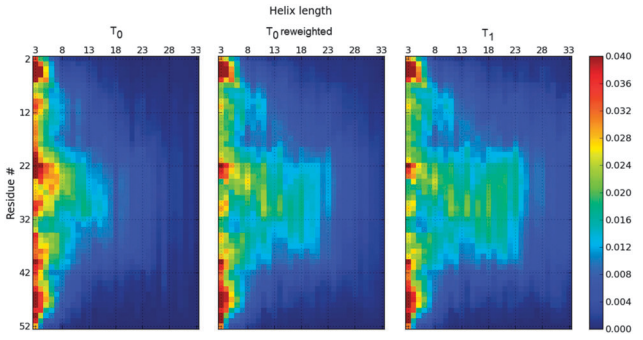


Fig. 2 SS-map of the Profasi ensemble at T_0 = 317.0 K (right) and T_1 = 325.6 K (left) and the T_0 MaxEnt re-weighted ensemble to fit T_1 N–H RDCs (middle). The latter ensemble has fewer long helices than the T_0 ensemble, but it still contains more long helices than the T_1 ensemble despite reproducing the RDCs at T_1.

Fig. 2 shows the SS-map of the synthetic ensembles at temperatures T_0 and T_1 and the re-weighted T_0-ensemble to fit T_1 N–H RDCs. Because T_0 is a lower temperature, this ensemble presents longer helices. Fig. 3 shows that the application of the MaxEnt principle returns a set of weights that can reproduce the final RDCs.

The re-weighting needed to fit the data gives a set of weights that are closer to 1 than the exact Boltzmann reweighting (see Fig. 3). In other words, although the exact Boltzmann weights can reproduce the RDCs of the objective T_1-ensemble (see Fig. S3, ESI†), the MaxEnt principle tells us that, based on the data, we do not need to change the weights that much, and that a lower modification of the ensemble is enough and consistent with the data.

As Fig. 3 suggests, the energy distribution of the reweighted T_0-ensemble is still closer to that of the T_0-ensemble than to that of the objective T_1. On average the energy increases but remains lower than the T_1-energy distribution (see Fig. S4, ESI†). Fig. S5 (ESI†) shows that most of the structures do not get re-weighted, and only a few do. For those that get re-weighted there is a certain correlation between the Boltzmann re-weighting and the re-weighting given by the N–H RDCs. Of course, if more data are used, for example Cα–Hα RDCs, the reweighting will increase, but even when doubling or tripling the number of experimental data, the degrees of freedom of the ensemble are much higher. We explore this in the following section.

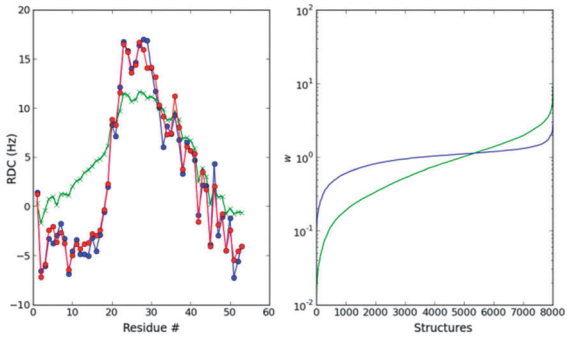


Fig. 3 Left: MaxEnt fit of the Profasi T_0 = 317.0 K ensemble to the Profasi T_1 = 325.6 K average N–H RDCs (blue). The unweighted ensemble (green) has too many long alpha-helices compared to the optimized ensemble (red). Right: distribution of the weights after the ME optimization (blue) compared to the exact Boltzmann weights.

The N–H RDCs do not give information on the energy but on the SS content of the structures; thus we expect the re-weighting to change the SS distribution. Fig. 2 and Fig. S6 (ESI†) reveal that the re-weighting of the data produced goes in the expected directions: the T_0 ensemble gets depleted from the long helices that give too large RDCs. But these figures also show that the SS-map of the resulting ensemble remains different from that of the objective T_1-ensemble. There are still regions of long helices much less populated in the T_1-ensemble. In the following section we will give a reason why the reweighting is not complete and only affects some of the structures.

The results from this section suggest that the RDCs give some information on the SS content of an ensemble, but this information is limited and cannot fully determine the helical propensity nor the helical lengths of an ensemble.

Application to experimental RDCs

We now focus on the reproduction of the experimental RDCs. First we use N–H RDCs and then we include Cα–Hα RDC either as a form of cross-validation or as a source of further structural information. Here, we treat the temperature of the simulation as a parameter, so that we first select the ensemble that best fits the N–H RDCs. For Profasi, this temperature is 325.6 K, and for Campari, the temperature is closer to the experimental one: 300.5 K. As these are the only ensembles we will use from now on, we will refer to them as Profasi and Campari ensembles. Previous studies showed that some force fields need higher-than-experimental temperatures to agree with the data;[57] however this adds a parameter that limits the predictive power of Profasi.

The Profasi ensemble fits the N–H RDCs reasonably well, but shows a region, around residue 35, of too much alpha helices. The MaxEnt algorithm produces a small reweighting of this ensemble, with most of the structures retaining a weight close to one. Therefore the SS-map of the ensemble is visually indistinguishable from the one shown in Fig. 2.

We can use the Cα–Hα RDCs to cross-validate this refined ensemble. The Cα–Hα RDCs are very similar to the original ones, showing that we did not incur overfitting, but differ significantly from the experimental RDCs (Fig. S7, ESI†). This shows that Cα–Hα and N–H RDCs are not correlated, and depend on different structural properties of the ensemble. The lack of agreement with Cα–Hα indicates that the Profasi ensemble does not correctly represent the real structural ensemble.

As fitting one set of RDCs does not affect the other, we can use MaxEnt to also fit Cα–Hα RDCs The resulting ensemble is reweighted to a stronger extent and correctly fits the 56 RDCs (Fig. S8, ESI†). However Fig. 5 shows that despite the use of the additional 25 Cα–Hα RDCs, the fitted Profasi ensemble has only changed its composition slightly (compare with Fig. 2). This change went in the expected direction, increasing the long

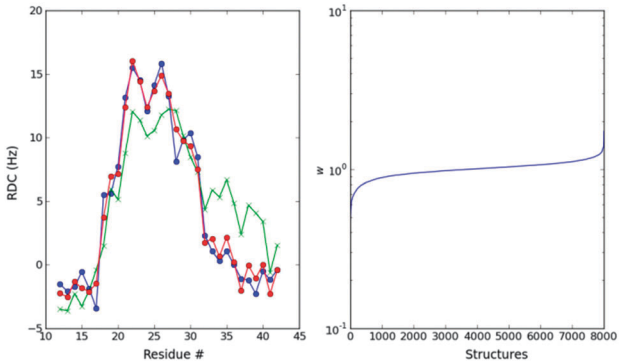


Fig. 4 Left: MaxEnt fit of the Profasi T_1 = 325.6 K ensemble to the experimental N–H RDCs (blue). The unweighted ensemble (green) has a region of too much alpha-helices compared to the optimized ensemble (red) between residues 32 and 40. Right: distribution of the weights after the optimization.

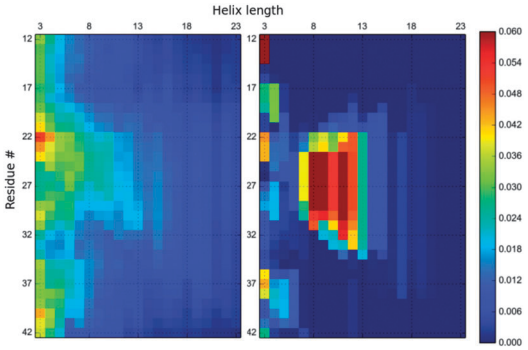


Fig. 5 SS-map of the MaxEnt re-weighted Profasi (left) and Campari (right) ensembles using 31 N–H RDCs and 25 Cα–Hα RDCs. Both ensembles fit the experimental RDCs to the same accuracy.

helices in the region of residues 20–27 and depleting the ensemble from helices in the region 31–39 (Fig. S9, ESI†). However this change was minor compared to the overall composition of the ensemble. Thus, even the use of 56 RDC data does not qualitatively change the Profasi ensemble and hints that it is still far from the real ensemble. We believe that this information can be used by developers to improve the quality of this force field. The spread of IDPs' energy landscape makes them a good target to find the balance between secondary structure populations and lengths *versus* random coils.

The Profasi ensemble differs from the ensemble deduced by Blackledge and co-workers,[11,44] which was mainly composed of random coil regions and three long helices. Their helices add up to 75% of the ensemble, and the longest helix has a population of 11% and ranges from residue 20 to 35. The robustness of their choice was checked by statistically significant improvement compared to other helical combinations. Despite Profasi being able to reproduce the folding of peptides and small proteins *ab initio*,[45,60] it does not predict the long helical elements suggested by Blackledge and co-workers.

The introduction of the experimental data does not reweight all the structures equally, because the weight of a structure depends on its RDC values.

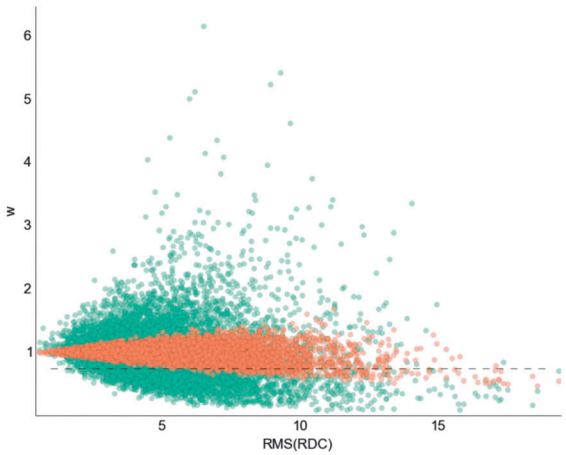


Fig. 6 Optimized weights for the Profasi ensemble to fit the experimental RDCs. The *x*-axis represents the root-mean-square of the RDCs for each of the 8000 structures, showing that the structures that get significantly reweighted are the ones that have large RDCs. When using only N–H RDCs (orange) the reweighting is smaller than when also using Cα–Hα RDCs. The dotted lines are set at w = 0.75, and define a fraction of structures that, if removed, improve significantly the fit. See the text for more details.

The set of RDCs forms a 31-component vector that is difficult to compare to weight of the structures. We can compress the information of this vector in its root-mean-square (RMS) value. If we plot the optimized weights *vs.* the RMS of the RDC vector for each structure, a clear trend appears (Fig. 6): the higher the RMS(RDC) the more reweighted the structure is. This makes sense, as reweighting a structure with small RDCs does not improve the fit. In other words, MaxEnt (or any other fitting procedure) is blind to structures that have low RDCs. Because RDCs can be scaled, "low" or "high" RDC refers to the value with respect to the other structures. As is well known, large RDCs correspond to long helices, and these structures are the ones MaxEnt finally re-weights to a larger extent.

Only 208 structures out of 8000 have a weight lower than 0.75 (see Fig. 7) when fitting N–H RDCs. Just by removing these structures from the ensembles and letting the others unchanged, the fit is almost as good as the optimized one in Fig. 4 (RMSD = 1.96 Hz compared to the optimized 1.00 Hz,

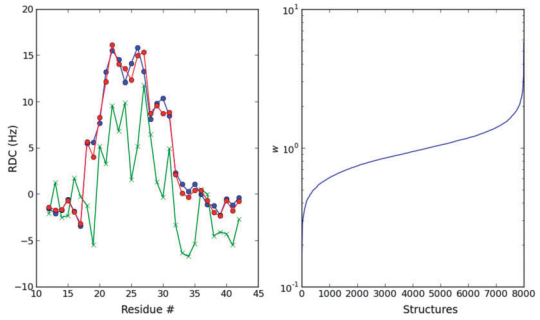


Fig. 7 Left: ME fit (red) of the Campari ensemble to the experimental RDCs (blue). The unweighted ensemble is shown in green. Right: distribution of the weights after the optimization.

Fig. S10, ESI†). The SS-map of these structures (Fig. S11, ESI†) reveals that these 208 structures are mainly long helices in the region of residues 32–40, just where the original Profasi ensemble gives RDCs that are too large. Thus the MaxEnt re-weighting agrees with our biophysical intuition.

We now turn to the comparison with the Campari ensemble. This comparison is illustrative because it allows disentangling the fitting procedure with prior distribution of the ensemble. Indeed, the comparison we did with Blackledge and co-workers was comparing a different ensemble and a different fitting procedure. This is a common practice in this field: different groups have developed sampling force fields and fitting procedures and the results contain information of both. For example, Forman-Kay group results are based on their ENSEMBLE selection procedure[15,17] from a TRaDES force field[16,20] generated structures. The present comparison will shed light on the information RDCs provide giving two different ensembles and *the same* fitting procedure.

The temperature of the Campari force field is better defined than that of Profasi, because the best fitting temperature corresponds to the experimental temperature. However, the initial ensemble has a worse agreement with the experimental N–H RDCs and therefore it needs a larger re-weighting (Fig. 7).

The secondary structure of this ensemble is considerably different from that of Profasi. It lacks the very abundant short helices of the Profasi ensemble and contains mainly helical fragments in the region of residues 22–32. This is, indeed, the region that the RDCs suggest should have helical fragments, and the region where Blackledge and co-workers deduced the helices were. There is a quantitative difference because the amount of helices in the Campari ensemble is lower than that obtained by Blackledge[11] (see also Fig. 4 in ref. 59). However, it is true that both convey a similar ensemble, whereas the Profasi one is qualitatively different. Despite the differences, the Campari and the Profasi ensemble to fit N–H RDCs have similar scaling factors (α = 3.97 and 3.67, respectively).

As before, the initial ensemble is similar to the optimized one, so that because the original Profasi and the Campari ensemble differ, the optimized ensembles still differ, even qualitatively. Even using the same fitting procedure, the starting ensemble has a pervasive influence in the optimized one. This is because the MaxEnt principle minimizes the modifications to the original ensemble, but this is a positive quality because it avoids overfitting or biasing the optimization procedure.

Again, we can introduce the Cα–Hα RDCs to increase the number of experimental data. As with Profasi, the reweighting increases, but the final ensemble is qualitatively very similar to the original. The cross-validation with Cα–Hα RDCs shows that the Campari predicted values are closer to the experimental ones. In spite of being closer, the N–H RDC reweighted ensemble does not improve the Cα–Hα (Fig. S7, ESI†) in agreement with the results of Profasi, and suggesting that the Cα–Hα are independent of the N–H RDCs.

Despite the difference between the Campari and Profasi ensemble, it is worth emphasizing that both are able to reproduce the positive N–H RDCs in the central region, and that the MaxEnt re-weighted ensemble does not differ significantly from the original ones. This may seem disappointing – if we expected them to collapse to the same final ensemble – but it also shows that the initially generated ensembles are physically reasonable. Based on the relation between energy and probability $\Delta E_i = -RT\log(w_i/w_i^0)$, where $w_i^0 = 1/N$, the energy difference for a reweighting of 0.5 is only 0.4 kcal mol^{-1}. Unfortunately, if we want to predict secondary structure elements we need these force fields to do better, and the RDC data can be used to improve them. The weight distribution of IDP structures is not peaked as with folded proteins, and thus can be easily reweighted to fit experimental data. Therefore agreement with experimental data does not guarantee a real structural ensemble. If we expect insights from the simulated ensembles we need force fields to have more predictive power. Campari seems to be more successful in this respect.

The Campari ensemble is ''simpler'' to interpret, but this does not seem to us a valid reason to favour it. In contrast, the Profasi ensemble needs less re-weighting and thus has more predictive power. It is true, however, that the use of an artificially high temperature in the Profasi ensemble is introducing a parameter that the Campari force field predicts to a good accuracy and this can also be the cause for the higher errors of the Cα–Hα in the Profasi ensemble. The Profasi temperature was originally defined as the correct scaling parameter of the energy to reproduce the melting temperature of the Trp cage peptide.[45] For IDPs maybe this parameter can be slightly scaled and it is then transferable to other sequences or maybe rescaling some of the energy terms results in a shifted temperature. Further systems need to be tested but our preliminary results suggest that the higher temperature is transferable among IDPs.

If, as before, we remove the structures that have $w < 0.75$ and leave the remaining unweighted, the fit of the Campari ensemble is very good (Fig. S12, ESI†). In this case, the number of structures removed is larger, 2074 out of 8000 (Fig. S13, ESI†). As with the Profasi ensemble, the structures that get a larger re-weighting are the ones that have larger RDC norm. The consistency of the re-weighting starting from different ensembles with different RDCs strengthens our confidence in the validity of the MaxEnt algorithm that we present.

Ideally, one wishes to start with a large pool of structures and let the data select the ones that agree with the ensemble. Different initial distributions should swamp to the same re-weighted distribution. Unfortunately, this is not the case: not even for folded proteins![41] RDCs do not convey enough information to make the initial distribution irrelevant. Our perspective is that the biophysical community has made heroic efforts in developing experimental techniques to probe IDPs, and then has hoped the data to speak by themselves, overlooking the influence of the prior distribution that the force fields produce.

Profasi and Campari can predict secondary structure elements in IDP ensembles based only on first principles, *i.e.* they can go beyond random coil force fields. But the ensembles they generate are different, and the RDC fitting cannot make them equal, not

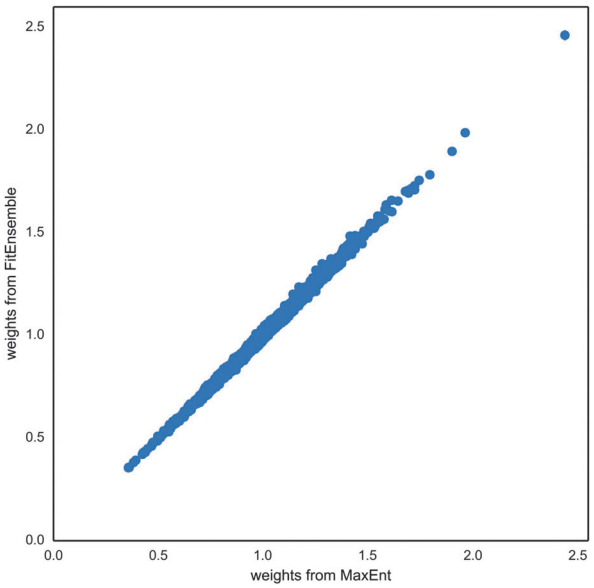


Fig. 8 Comparison of the fitting of MaxEnt and FitEnsemble.[38] FitEnsemble results include the estimated error from the Bayesian procedure, but it is of the order of the point size.

even similar. They do have an influence on the final ensemble that can fit the RDC data. This is not to say that the RDCs are not informative, but that the ensembles that fit the data combine the information from the RDCs with that of the force fields. Efforts should be made to improve both experimental methods and force fields. Indeed, we believe that the efforts in the latter field lag behind the experimental developments attained in the IDP world.

Comparison with FitEnsemble

The recent publication of FitEnsemble,[38] a method to reweight calculated ensembles to experimental data, prompted us to compare this approach with ours. The advantage of FitEnsemble is that it is a fully Bayesian approach. It is one order of magnitude slower than MaxEnt, but that involves times of a bit more than a minute, which is still very competitive. The problem is that it cannot work with scale invariant quantities such as RDCs. Here we take the scaling factor of the optimized ensemble with MaxEnt to compare both methods.

The agreement with both methods is very high (Fig. 8). We also see that the uncertainty in the weights is low compared to its dispersion. That confirms our assumption that this is not a key parameter. We found that the resulting FitEnsemble fit has much lower errors than the introduced experimental uncertainty. In particular, for an uncertainty of 1 Hz, the fit has a root-mean-square error of 0.2. Therefore we optimized our MaxEnt to a threshold of 0.1. For the FitEnsemble, we used a regularization strength of 3, as suggested by the authors but we checked that values of 0.3 and 30 essentially produced the same average results and the same dispersion.

The extension of FitEnsemble to include a scale parameter seems to be an interesting approach. Still, questions about the convergence of MCMC for RDC ensembles need to be addressed, as well as ensuring that it remains a computationally affordable method.

Conclusions

We present an algorithm based on the maximum entropy principle, which minimizes the information introduced in the fitting of experimental data to a given ensemble. We adapted the algorithm to work with scale invariant measures, such as RDCs. The algorithm is implemented in an open source code freely available.[27] The advantage of our method is that it can be used by different experimental groups using different ensembles, as it can use any given set of structures. It can use thousands of structures and converges in a few seconds. It also avoids the risk of overfitting, as the number of parameters depends only on the number of experimental data, and not on the number of structures in the ensemble. Cross-validation shows that more than 7000 structures need to be used to get errors close to the experimental errors of 1 Hz.

It has been claimed that RDCs are one of the best probes of IDPs' residual secondary structure,[12] but other studies have questioned the relevance of RDCs in IDP modelling.[17] Our results, with both a synthetic and an experimental data set, suggest that RDCs can shift the ensembles' secondary structure composition, but only to a limited extent. Different sets of RDCs – N–H and Cα–Hα – give complementary information and improve the reweighting; however the vast conformational space that IDPs can sample makes it a complex case of inferential structure determination,[22] so that even with the large number of RDC experimental data, the amount of data is sparse compared to the size of the ensemble.[40]

Neither all-atom nor coarse-grained force fields have the precision to describe an IDP ensemble,[61] as errors of 1 or 2 kcal mol^{-1} can significantly shift the populations of helices or other secondary or tertiary structure elements. Therefore the need to use experimental data to improve these ensembles is mandatory. But the experimental data is insufficient to fully determine this ensemble, and the pervasive influence of the force field cannot be overlooked, if we wish to have consistent representations of IDP ensembles.

Even though both Campari and Profasi predict certain secondary structure elements, their ensembles are qualitatively different. That determines the composition of the MaxEnt reweighted ensembles. The combination of Cα–Hα and N–H RDCs suggests that Campari is more suitable to describe IDPs than Profasi. We still need further work to test other force fields, improve them, and check other complementary sources of data that help up further select the ensembles. One of our future goals is to include SAXS and chemical shifts in our maximum entropy code.

Appendix

Here we derive the expression of the gradient of f_1 and f_2, needed for their optimization.

For the sake of simplicity we will derive the gradient of f_1 piecewise. We only consider when the argument in eqn (3) is

larger than the threshold; otherwise the gradient is the null vector. The gradient of the average RDC is

$$\mathbf{g}(\langle\mathbf{q}\rangle) := \frac{\partial\langle q_n\rangle}{\partial\lambda_i} = -\sum_j^N q_i^j q_n^j \exp\left(\sum_l^M -q_l^j\lambda_l\right)$$

The gradient of the scaling factor α is

$$\mathbf{g}(\alpha) := \frac{\partial\alpha}{\partial\lambda_i} = \frac{s\mathbf{g}(\langle\mathbf{q}\rangle)\cdot\mathbf{Q}(\langle\mathbf{q}\rangle\cdot\langle\mathbf{q}\rangle) - 2|\langle\mathbf{q}\rangle\cdot\mathbf{Q}|\langle\mathbf{q}\rangle\cdot\mathbf{g}(\langle\mathbf{q}\rangle)}{(\langle\mathbf{q}\rangle\cdot\langle\mathbf{q}\rangle)^2}$$

where s is the sign function of $\langle\mathbf{q}\rangle\cdot\mathbf{Q}$. Finally,

$$\frac{\partial f_1}{\partial\lambda_i} = \frac{2}{M}(\boldsymbol{g}(\alpha)\times\langle\mathbf{q}\rangle + \alpha\mathbf{g}(\langle\mathbf{q}\rangle))\cdot(\alpha\langle\mathbf{q}\rangle - \mathbf{Q})$$

where $\times$ represents the outer product. The gradient for f_2 is trivial:

$$\frac{\partial f_2}{\partial\lambda_i} = 2\frac{k}{M}\lambda_i$$

Acknowledgements

We would like to thank X. Salvatella and P. Bernadó for critically reading the manuscript. We acknowledge financial support from the Ministerio de Economía y Competitividad (CTQ2012-33324) and the Generalitat de Catalunya (2009SGR01472). MS-M thanks the Ministerio de Economia y Competitividad for a predoctoral fellowship. We thank the CCUC and the RES (BCV-2013-3-0015) for computational resources.

Notes and references

1 M. M. Babu, R. van der Lee, N. S. de Groot and J. Gsponer, *Curr. Opin. Struct. Biol.*, 2011, **21**, 432–440.
2 V. N. Uversky, C. J. Oldfield and a. K. Dunker, *Annu. Rev. Biophys.*, 2008, **37**, 215–246.
3 J. Wang, Z. Cao, L. Zhao and S. Li, *Int. J. Mol. Sci.*, 2011, **12**, 3205–3219.
4 C. K. Fisher and C. M. Stultz, *Curr. Opin. Struct. Biol.*, 2011, **21**, 426–431.
5 M. Fuxreiter, *Mol. BioSyst.*, 2012, **8**, 168–177.
6 V. Iešmantavičius, J. Dogan, P. Jemth, K. Teilum and M. Kjaergaard, *Angew. Chem., Int. Ed.*, 2014, **53**, 1548–1551.
7 A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker and V. N. Uversky, *J. Mol. Biol.*, 2006, **362**, 1043–1059.
8 W.-L. Hsu, C. J. Oldfield, B. Xue, J. Meng, F. Huang, P. Romero, V. N. Uversky and a. K. Dunker, *Protein Sci.*, 2013, **22**, 258–273.
9 K. Chen and N. Tjandra, *Top. Curr. Chem.*, 2012, **326**, 47–67.
10 L. Salmon, M. R. Jensen, P. Bernadó and M. Blackledge, *Methods Mol. Biol.*, 2012, **895**, 115–125.
11 M. R. Jensen, K. Houben, E. Lescop, L. Blanchard, R. W. H. Ruigrok and M. Blackledge, *J. Am. Chem. Soc.*, 2008, **130**, 8055–8061.
12 M. R. Jensen, P. R. L. Markwick, S. Meier, C. Griesinger, M. Zweckstetter, S. Grzesiek, P. Bernadó and M. Blackledge, *Structure*, 2009, **17**, 1169–1185.
13 R. Schneider, J. Huang, M. Yao, G. Communie, V. Ozenne, L. Mollica, L. Salmon, M. R. Jensen and M. Blackledge, *Mol. BioSyst.*, 2012, **8**, 58–68.
14 W. Y. Choy and J. D. Forman-Kay, *J. Mol. Biol.*, 2001, **308**, 1011–1032.
15 J. A. Marsh, C. Neale, F. E. Jack, W.-Y. Choy, A. Y. Lee, K. A. Crowhurst and J. D. Forman-Kay, *J. Mol. Biol.*, 2007, **367**, 1494–1510.
16 J. A. Marsh, J. M. R. Baker, M. Tollinger and J. D. Forman-Kay, *J. Am. Chem. Soc.*, 2008, **130**, 7804–7805.
17 J. A. Marsh and J. D. Forman-Kay, *Proteins*, 2012, **80**, 556–572.
18 A. F. Ángyán and Z. Gáspári, *Molecules*, 2013, **18**, 10548–10567.
19 V. Ozenne, F. Bauer, L. Salmon, J.-R. Huang, M. R. Jensen, S. Segard, P. Bernadó, C. Charavay and M. Blackledge, *Bioinformatics*, 2012, **28**, 1463–1470.
20 H. J. Feldman and C. W. V. Hogue, *Proteins*, 2000, **131**, 112–131.
21 G. W. Daughdrill, S. Kashtanov, A. Stancik, S. E. Hill, G. Helms, M. Muschol, V. Receveur-Bréchot and F. M. Ytreberg, *Mol. BioSyst.*, 2012, **8**, 308–319.
22 W. Rieping, M. Habeck and M. Nilges, *Science*, 2005, **309**, 303–306.
23 E. Jaynes, *Phys. Rev.*, 1957, **106**, 620–630.
24 S. Pressé, K. Ghosh, J. Lee and K. A. Dill, *Rev. Mod. Phys.*, 2013, **85**, 1115–1141.
25 W. Boomsma, J. Ferkinghoff-Borg and K. Lindorff-Larsen, *PLoS Comput. Biol.*, 2014, **10**, e1003406.
26 B. Róycki, Y. C. Kim and G. Hummer, *Structure*, 2011, **19**, 109–116.
27 https://github.com/MelchorSanchez/MaxEnt.
28 J. W. Pitera and J. D. Chodera, *J. Chem. Theory Comput.*, 2012, **8**, 3445–3451.
29 A. D. White and G. A. Voth, *J. Chem. Theory Comput.*, 2014, **10**, 3023–3030.
30 B. Roux and J. Weare, *J. Chem. Phys.*, 2013, **138**, 084107.
31 A. Cavalli, C. Camilloni and M. Vendruscolo, *J. Chem. Phys.*, 2013, **138**, 094112.
32 S. Esteban-Martín, R. B. Fenwick and X. Salvatella, *J. Am. Chem. Soc.*, 2010, **132**, 4626–4632.
33 R. B. Fenwick, S. Esteban-Martín and X. Salvatella, *Eur. Biophys. J.*, 2011, **40**, 1339–1355.
34 S. Meier, S. Grzesiek and M. Blackledge, *J. Am. Chem. Soc.*, 2007, **129**, 9799–9807.
35 E. Jones, E. Oliphant, P. Peterson, *et al.*, SciPy: Open Source Scientific Tools for Python, 2001, see http://www.scipy.org/scipylib/citing.html, accessed 28th October 2014.
36 C. K. Fisher, A. Huang and C. M. Stultz, *J. Am. Chem. Soc.*, 2010, **132**, 14919–14927.
37 C. K. Fisher, O. Ullman and C. M. Stultz, *Pac. Symp. Biocomput.*, 2012, 82–93.
38 K. a. Beauchamp, V. S. Pande and R. Das, *Biophys. J.*, 2014, **106**, 1381–1390.

39 A. Patil, D. Huard and C. J. Fonnesbeck, *J. Stat. Softw.*, 2010, **35**, 1–81.
40 S. Olsson, J. Frellsen, W. Boomsma, K. V. Mardia and T. Hamelryck, *PLoS One*, 2013, **8**, e79439.
41 S. Olsson, W. Boomsma, J. Frellsen, S. Bottaro, T. Harder, J. Ferkinghoff-Borg and T. Hamelryck, *J. Magn. Reson.*, 2011, **213**, 182–186.
42 S. Olsson, B. R. Vögeli, A. Cavalli, W. Boomsma, J. Ferkinghoff-Borg, K. Lindorff-Larsen and T. Hamelryck, *J. Chem. Theory Comput.*, 2014, **10**, 3483–3491.
43 K. Berlin, C. A. Castañeda, D. Schneidman-Duhovny, A. Sali, A. Nava-Tudela and D. Fushman, *J. Am. Chem. Soc.*, 2013, **135**, 16595–16609.
44 P. Bernadó, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok and M. Blackledge, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 17002–17007.
45 A. Irbäck and S. Mohanty, *Biophys. J.*, 2005, **88**, 1560–1569.
46 A. Irbäck, S. Mitternacht and S. Mohanty, *PMC Biophys.*, 2009, **2**, 2.
47 A. Irbäck and S. Mohanty, *J. Comput. Chem.*, 2006, **27**, 1548–1555.
48 M. Zweckstetter, *Nat. Protoc.*, 2008, **3**, 679–690.
49 G. Nodet, L. Salmon, V. Ozenne, S. Meier, M. R. Jensen and M. Blackledge, *J. Am. Chem. Soc.*, 2009, **131**, 17908–17918.
50 O. I. Obolensky, K. Schlepckow, H. Schwalbe and A. V. Solov'yov, *J. Biomol. NMR*, 2007, **39**, 1–16.
51 A. Vitalis and R. V. Pappu, *J. Comput. Chem.*, 2009, **30**, 673–699.
52 X. Cong, N. Casiraghi, G. Rossetti, S. Mohanty, G. Giachin, G. Legname and P. Carloni, *J. Chem. Theory Comput.*, 2013, **9**, 5158–5167.
53 S. A. Jónsson, S. Mohanty and A. Irbäck, *Proteins*, 2012, **80**, 2169–2177.
54 A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine and R. V. Pappu, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 8183–8188.
55 M. Varadi, S. Kosol, P. Lebrun, E. Valentini, M. Blackledge, A. K. Dunker, I. C. Felli, J. D. Forman-Kay, R. W. Kriwacki, R. Pierattelli, J. Sussman, D. I. Svergun, V. N. Uversky, M. Vendruscolo, D. Wishart, P. E. Wright and P. Tompa, *Nucleic Acids Res.*, 2014, **42**, D326–D335.
56 R. Bürgi, J. Pitera and W. F. van Gunsteren, *J. Biomol. NMR*, 2001, **19**, 305–320.
57 D. S. Weinstock, C. Narayanan, A. K. Felts, M. Andrec, R. M. Levy, K.-P. Wu and J. Baum, *J. Am. Chem. Soc.*, 2007, **129**, 4858–4859.
58 B. Richter, J. Gsponer, P. Várnai, X. Salvatella and M. Vendruscolo, *J. Biomol. NMR*, 2007, **37**, 117–135.
59 J. Iglesias, M. Sanchez-Martínez and R. Crehuet, *Intrinsically Disord. Proteins*, 2013, **1**, e25323.
60 S. Mohanty, J. H. Meinke and O. Zimmermann, *Proteins*, 2013, 1–11.
61 M. R. Jensen and M. Blackledge, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, E1557–E1558.

Chapter 5

Conclusions

With the aim of providing a picture of different aspects relevant to protein dynamics (regarding both local and global motions), in the present thesis we studied several proteins using a variety of computer simulation methods. Our main observations in these two different kind of motions are the following:

Local Motions

- The catalytic proficiency of EcNAGK lies in open-close (collective) motions accessing properly oriented and highly compressed active site conformations. That supports the 'conformational compression' hypothesis inferred by Rubio and coworkers.
- At least for EcNAGK the protein motions leading to compressed catalytic conformations seems to be the limiting process instead of the chemical step.
- The Swarms of Trajectories (SoT) method can be applied to study enzyme catalysis because the method is independent of the dynamics of the CV. We devised a suitable implementation to obtain the contribution of each collective variable to the free energy profile.
- Within the SoT method the addition of a CV without a role in the chemical reaction has no effect in the free energy profile nor in the computational cost of the SoT simulation. However missing a CV which is redundant for the chemical reaction results in the free energy barrier being underestimated.
- Proteins with acidic residues are susceptible to suffer radiation damage effects when are exposed to high radiation as the emitted by synchrotron techniques. In LDH this damage is translated to a decarboxylation. This process is a HT that proceeds through a superexchange mechanism, determined by the distance between Asp and Trp.

Global Motions

- The traditional visualization methods difficult the visualization of cooperativity effects in secondary structure elements leading to incomplete or incorrect interpretations of their propensities, both in globular proteins and in IDPs.

- The different secondary structure elements present different ϕ and ψ angle values in the Ramachandran diagram as well as different structural signatures between them. α-helices and β-strands behave different with the temperature and PPII-helices do not grow from a central residue oppositely to α-helices.

- The Maximum Entropy principle is a good choice to introduce experimental information in an ensemble. We devised a scale invariant implementation of it to study RDCs. Our implementation also allow to avoiding the overfitting as well as accelerating its performance (it can treat thousand of structures in seconds).

A general conclusion from this thesis is that computational methods are an efficient and useful tool to characterize protein motions. However the current computational approaches present limitations and to solve them the incorporation of experimental data and its correct interpretation is crucial. This necessity of be complemented comes from different sides: 1) from experiments to computations and 2) from computations to experiments, as it is good exemplified in IDPs. The experimental techniques used to study IDPs need computation to provide a global and unified vision of the conformational landscape (as NMR give only snapshots of it) and the computational methods need experimental data to improve their performance.

The convergence of experimental and computational techniques to the same point is key to achieve a deep understanding of protein dynamics.

Chapter 6

Sumario

La presente tesis se centra en el estudio computacional de la dinámica de las proteínas. Las proteínas son entidades flexibles y como tales se mueven. Este movimiento es indispensable y esta directemente relacionado con su función. La dinámica de las proteínas se puede dividir en dos grandes bloques conceptuales según el número de átomos inolucrados, la escala de tiempo en que que tiene lugar y la amplitud y dirección de la misma.

Por un lado se encuentran las dinámicas a nivel local, es decir, aquellas que se producen a nivel de 'centro activo' que implican la reorganización de unos pocos átomos de la cadena lateral de los aminoácidos o del esqueleto de la proteina. Estas dinámicas locales también suelen considerarse como movimientos rápidos ya que la escala de tiempo en la cual tienen lugar se encuentra por debajo del milisengundo (ms). A su vez también se clasifican como dinámicas de pequeña amplitud. Por otro lado los movimientos globales se dan a nivel de estructura y engloban procesos como el alosterimo, la modulación conformacional e incluso el plegamiento de la proteico. Atendiendo a la escala de tiempo se consideran dinámicas lentas porque tienen lugar en escalas de tiempo iguales o superiores al miliegundo. Además son consideradas dinámicas de gran ampitud.

Hay cierta controversia con la terminología clasificatoria porque a veces los movimientos locales también se consideran lentos si tenemos en cuenta la frecuencia con la que ocurren ya que generalmente es neceario un movimiento global para que tenga lugar un movimiento local. Sea como fuere en el momento en el que suceden son dinámicas muy rápidas, y es por ello que en esta tesis las hemos definido como tal ya que creemos que esta terminoogía describe mejor la naturaleza de estos movimientos. Para cartacterizar y estudiar estos movimientos existen una amplia gama de técnicas experimentales y computacionales.

En esta tesis doctoral se ha tratado de dar respuesta a varios fenómenos observados en relación con la dinámica de las proteínas. Concretamente hemos realizado

estudios a nivel local, de 'centro activo', relacionados con la catálisis enzimática y el daño proteico así como, a nivel global, con la determinación y el análisis de conjuntos conformacionales de proteínas. Estos estudios, se han realizado usando métodos propios de la química, la bioquímica y la biofísica computacionales, los cuales se han mostrado como herramientas muy útiles a la hora de estudiar la dinámca de las proteinas.

Efecto de los movimientos conformacionales en la catálisis enzimática

Las enzimas son macromoléculas biológicas de naturaleza proteíca capaces de acelerar la velocidad de las reacciones (bio)químicas celulares en más de 10 órdenes de magnitud, alcanzando escalas de tiempo biológicamente relevantes. La clave de la extraordinaria eficiencia catalítica de las enzimas reside en la gran preorganización de su centro activo, el cual presenta aminoácidos con distinta polaridad en una conformación óptima para unir el sustrato y estabilizar el estado de transición de la reacción.

La función de los movimientos conformacionales en las proteínas es un tema de debate muy actual y que genera una gran controversia. La existencia de los movimientos conformacionales esta ampliamente aceptada por la comunidad científica, así como el hecho de que la dinámica proteica tiene lugar dentro del ciclo catalítico, ya que hay estudios computacionales y experimentales que así lo demuestran. Sin embargo, hay un sector que postula que estos movimientos catalizan el paso químico mientras otro defiende que no. Hay un gran debate en torno al papel de los 'efectos dinámicos' y lo que se entiende por este término. Mientras para unos solamente son desviaciones de la teoría del estado de transición para otros representan cualquier modificación conformacioanl que sufra la proteína dependiente del tiempo. El debate, si se analiza en detalle y de forma objetiva, parece ser de alguna forma mayoritariamente semántico y para terminar con él lo que se necesita es una definición clara y consenuada de lo que se considera como 'efectos dinámicos'.

Sea como fuere, y por medio del estudio de la enzima EcNAGK nosotros aportamos nuestro granito de arena esclareciendo el papel de los movimientos conformacionales en esta enzima. Experimentalmente esta enzima había sido caracterizada y ampliamente estudiada junto a otros miembros de su familía (AAK) por el grupo del Prof. Rubio en Valencia, al igual que computacionalmente por nuestro grupo. Es una enzima que constituye un ejemplo idóneo porque hay depositadas en el PDB seis estructuras caracterizando diferentes estados de la reacción (bio)química. Rubio y sus colabroadores mostraron que aquellos que correspondían a estrucutras análogas al estado de transición presentaban distancias entre los substratos de la reacción más bajas que las estructuras cristalinas que

representaban el estado de reactivos. De aquí infirieron que la 'compresión conformacional'(distancia O-O entre los substratos, ATP y NAG) del centro activo favorecía la catálisis.

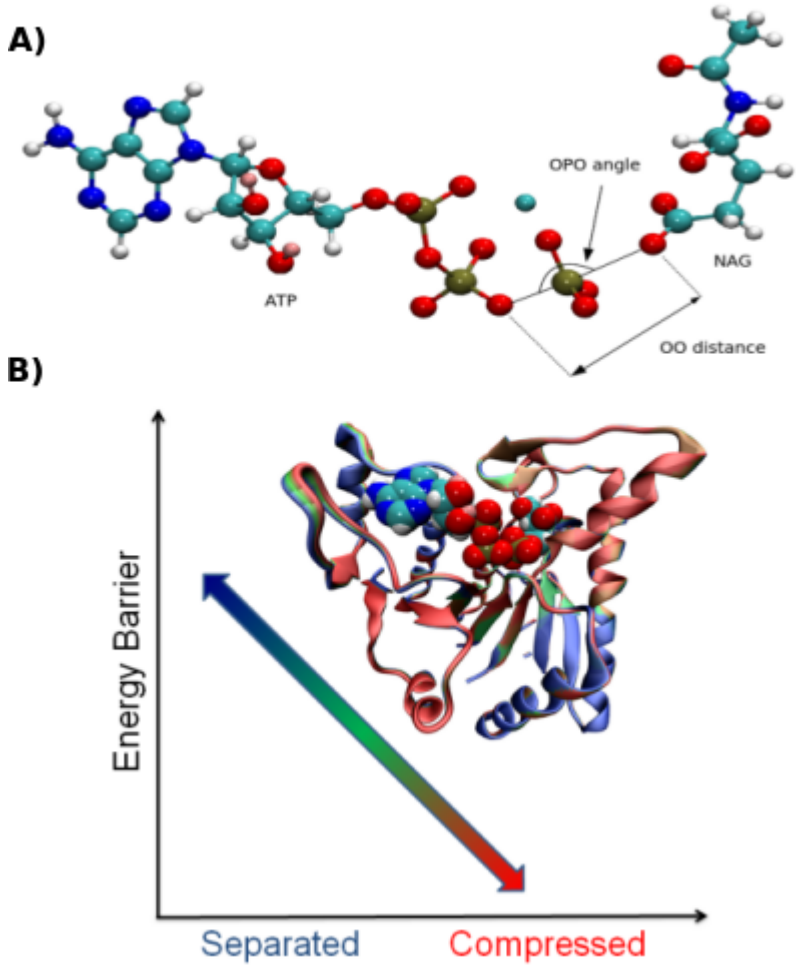


Figure 6.1: A) Representación esquemática de los substratos naturales de la EcNAGK. B) Representación esquemática de la dinámica de EcNAGK. Cuanto más rojo más comprimída es la conformación y cuanto más azul más abierta. Las esferas representan los substratos indicados en el panel A.

Para investigar esta hipótesis complexamos cuatro estrucutras cristalinas representativas (código PDB 1GS5, 1OH9, 1OHA, y 2X2W) con los substratos naturales de la rección, ATP y NAG. Por medio de 1) cálculos de dinámica molecular (MD) seguidos de 2) cálculos de mecánica cuántica / mecánica molecular (QM/MM) (a nivel DFT usando el funcional mPWPW91) sobre algunos snapshots de la trayectorias MD generadas y finalmente 3) análisis estadísticos (PCA+PLSR) sobre los resultados de estos últimos, investigamos la reactividad de la enzima NAGK. Tratamos de esclarecer el papel de los movimientos confor-

macionales, así como la influencia de la distribución espacial de los reactivos y la distancia entre ellos en el perfil de la reacción y por tanto en la catálisis enzimática. Entre todas las estructuras cristalinas empleadas descubrimos que dos de ellas (1OHA and 1OH9) eran las más adecuadas para nuestros estudios.

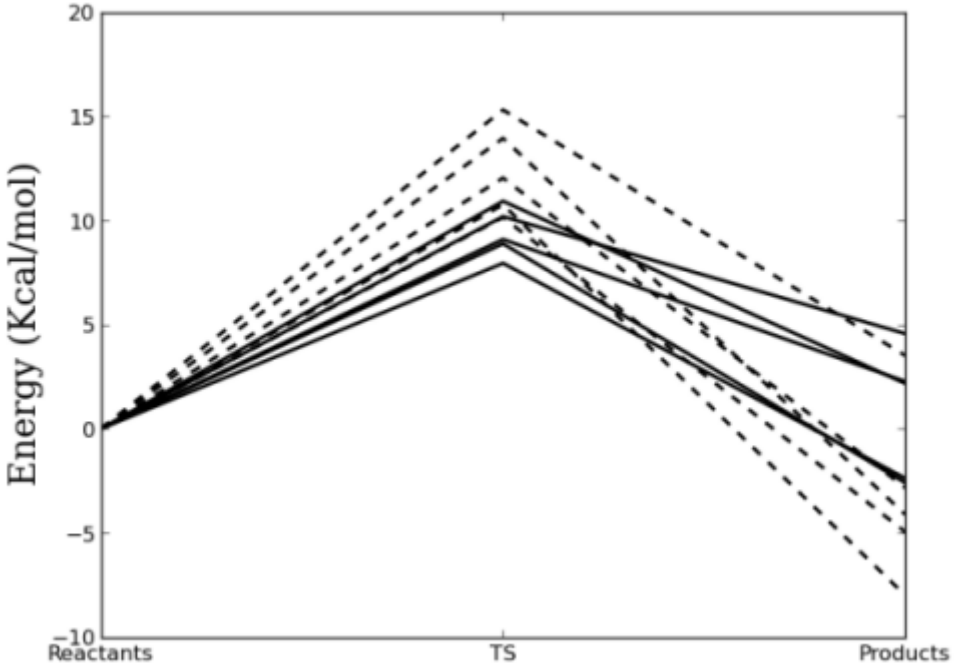


Figure 6.2: Energías de las conformaciones de reactivos, estado de transición y productos de las cinco estructuras 1OH9 (líneas discontinuas) y 1OHA (líneas sólidas). La dispersión de los valores energéticos es larga incluso para las esructuras que vienen del mismo cristal.

La variedad de perfiles de reacción obtenidos incluso sobre la misma estructura cristalina, indican que la barrera de energía no esta determinada por el cambio en la distancia (entre el grupo nucleófilo y el saliente, O-O), al menos no únicamente. Hay una correlación notable entre la distancia de los reactivos y la barrera enérgetica, cuanto menor es la distancia menor es la barrera, tal y como determinaron los análisis estadíticos. Sin embargo hay que tener en cuenta la dependecia de esta distancia y por ende de la barrera respecto a la orientación espacial de los substratos, es decir, el ángulo lineal entre el grupo fosforilo que se transmite, el grupo nucleófilo y el grupo saliente. A mayor linealidad del ángulo, menos barrera. Por lo tanto la hipótesis de Rubio se refórmula: A menor distancia O-O y mayor linealidad del ángulo O-P-O, menor barrera enérgetica y por lo tanto menos inestabilidad del estado de transición.

La barrera energética calculada para el paso químico en todas las estructuras

crstalinas, cuyo valor medio de 9 kcal/mol, es significativamente menor que la barrera energética experimental de 16 kcal/mol. Esta última representa la energía libre de la reacción, e incluye correcciones dinámicas y de efecto tunel, además de que nuestros cálculos dan valores de energía potencial no de energía libre. Esto implica que no podemos comparar directamente ambos valores, pero la diferencia energética es tan grande que no puede ser cancelada por este motivo. Esta diferencia enérgetica tan grande sugiere, al igual que ocurre en otras enzimas, que los movimientos de apertura y cierre pueden ser más lentos que el propio paso químico de la reacción.

Los resultados sugieren que la velocidad del enzima no depende del paso químico de la reacción sino de la unión o no del NAG en su centro activo, siendo los movimientos de apertura y cierre del centro activo, inducidos por la presencia o no del NAG, los que limitan su eficiencia catalítica.

Cálculo de caminos de energa libre en catálisis enzimática

Una de las mayores problemáticas dentro del estudio computacional de la catáisis enzimática reside en el cálculo de la energía asociada al camino de reacción, concretamente del de mínima energia libre. Los enzimas presentan superfices de energía rugosas y multidimensionales que dificultan el cálculo de caminos de energía libre.

Para calcular caminos de reacción los métodos actuales se podrían englobar dentro de dos grupos. EL primero se basa en el uso de variables colectivas (CV) o coordenadas de reacción, para describir la superficie de energía libre (o Potencial de fuerza media). Estas variables idealmente han de ser pocas y su elección muy precisa ya que el añadir o no una variable supone un incremento notable del tiempo de cálculo. Desafortunadamente este no es el caso de las reacciones enzimáticas por lo que estos métodos son muy costosos. El segundo grupo de métodos no utiliza las CV pero necesitan un estado inicial y final de la reacción entre los cuales interpolar el camino de la misma. Estos métodos aunque menos costosos computacionalmente, en sus versiones más simples solo generan caminos de energía potencial, debido a que omiten las contribuciones entrópicas y de muestreo (sampling).

En los últimos años han aparecido los llamados métodos híbridos que incorporan lo mejor de los 'dos mundos'. Dentro de estos se encuentra el llamado enjambre de trayectorias (Swarms of Trajectories (SoT)). Este método desarrollado por Roux y sus colaboradores está basado en el método original de la cadena (String method) con CV, de Vanden Eijnden. Su característica más relevante es que para estimar el desplazamiento (durante la optimización) de cada punto del camino de reacción, se realizan una serie de trayectorias cortas sin ningún tipo de restricción conformacional (un 'enjambre' de trayectorias) sobre las cuales se

calcula el deslazamiento medio entre la inicial y la final. Repecto al valor obtenido se evoluciona el camino de reacción.

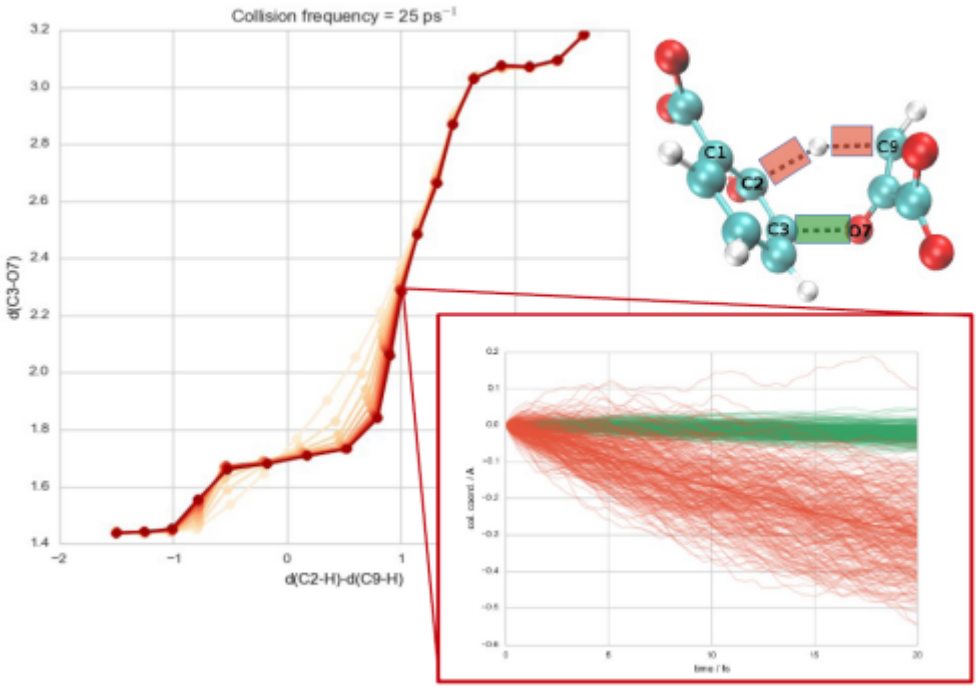


Figure 6.3: Represenatción esquemaática de un camino de reacción optimizado con SoT para la enzima Isocorismato piruvato liasa. La optimización va de claro (camino inicial) a oscuro (camino final optimizado). La figura englobada dentro del cuadrado representa el enjambre de trayectorias calculado sobre un punto del camino.

En un estudio reciente se compararon ambos métodos y se demostró que SoT, algo menos costoso computacionalmente, da resultados equivalentes al método original, siendo ambos matemáticamente equivalentes. También, se sugirió su aplicacbilidad a sistemas biomoleculares debido a que se hipotetizó que podía funcionar a régimenes de tiempo cortos en sistemas inerciales. Hasta el momento solamente se había aplicado a sistemas no inerciales en los cuales las variables colectivas dependían de forma lineal de la dinámica del sistema, lo cual dificultaba su aplicacion a reacciones de catálisis enzimática.

En esta tesis, en colaboración con el Prof. Martin Field en Grenoble implementamos este método en la librería pDynamo. Esta librería está escrita en python y es la que utilizamos en la mayoría de nuestros cálculos relacionados con la catálisis enzimática. Una vez implementada, la testeamos sobre dos enzimas ampliamente estudiadas y que requieren un nivel de cálculo, y por tanto de tiempo de computación, bajo. Esto nos permitió realizar muchas pruebas. Estos enzimas son

la Corismato mutasa (Chorismate Mutase (CM)) y la Isocorismato piruvato liasa (Isochorismate Pyruvate Lyase (IPL)). Para ello realizamos cálculos QM/MM a nivel AM1.

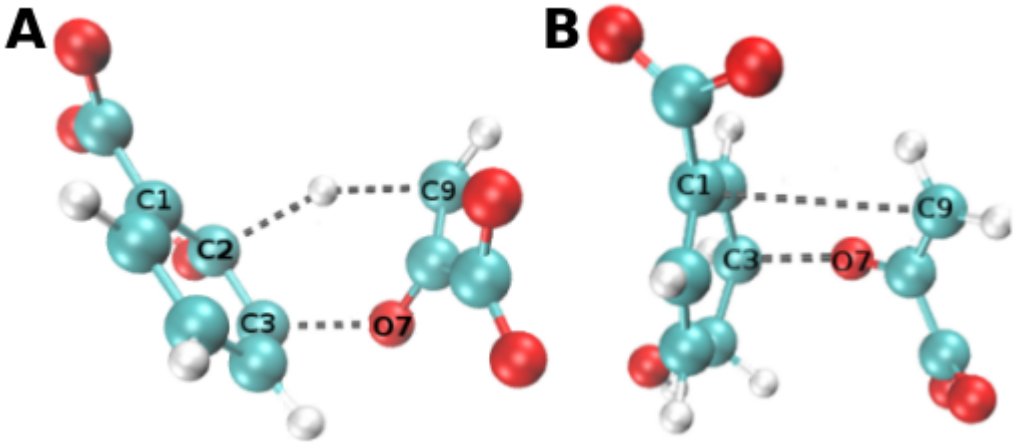


Figure 6.4: Represenatción del las estructuras del estado de transición de los susbtratos de las enzimas utilizadas para testear el método SoT. A) IPL. B) CM.

En este estudio, fuimos capaces de encontrar la forma correcta de introducir SoT en la librería pDynamo, así como de corroborar la hipótesis que postulaba que SoT podía funcionar en regímenes inerciales con pasos de tiempo cortos. Esto además demostró la independencia de la dinámica del sistema respecto del método. Por otro lado, debido a como lo implementamos, se pueden construir perfiles de energía libre a partir del camino de reacción, pudiendo descomponerlos y conocer la contribución de cada variable colectiva al mismo. Por otro lado si añadimos una CV que no tiene ningún efecto sobre el camino de la reacción, su valor permanece invariable a lo largo de la optimizaión y sin un coste computacional adicional. Esto implica que podemos añadir CV si estamos dudando entre si tiene un efecto o no ya que no supone un coste adicional, lo cual es interesante debido a la complejidad de las reacciones enzimáticas.

Daño proteico inducido por radiaciones de alta intensidad

La cristaligrafía macromolecuar por rayos X es una técnica ampliamanete utilizada para caracterizar estructuras enzimáticas. Los diferentes subestados de la reacción bioquímica (análogos de reactivos o del estado de transición por ejemplo) quedan atrapados (caracterizados) en diferentes cristales. Para realizar los experimentos correspondientes a esta técnica es común el uso del sincrotrón. Sin embargo, esto puede generar problemas debido a que la alta radicación utlizada en los métodos de

sincrotrón puede estropear la muestra y provocar daños por radiaición. El estudio de estos fenómenos constituye un campo de investigación activo. Por ejemplo, son capaces de producir reacciones químicas en enzimas que no tienen lugar de forma natural, debido a que producen daños específicos en las cadenas lateral esde los aminoácido.

Las proteínas con residuos ácidos en su interior son especialmente sensibles a estos efectos. El grupo del Prof. Martin Weik en Grenoble es experto en este tipo de fenómenos de daño por radiación. Recientemente estudiando la enzima Lactato Deshidrogenasa (Lactate dehydrogenase (LDH)) observaron que en la forma Apo de le enzima tenía lugar la decarboxilación de un triptófano que no era natural y que sucedía lejos del centro activo. Sin embargo en la forma Holo de la misma no se observaba. Ellos concluyeron que este proceso se debía a un fenomeno de 'Hole Transfer' (HT) entre el Trp62 y el Asp70 mediante la presencia de la Arg64. En colaboración con ellos estudiamos estas dos estructuras para dar una explicación del porque de esta reacción.

Tras revisar el sistema surgieron varias cuestiones de forma natural. Parecía claro que la trasnferencia de carga tenía lugar. Lo que había que corroborar era si el proceso tenía lugar y en caso afirmativo si se dabe entre el Trp62 y el Asp70 o viceversa. También era clave discernir porque se producía en la forma apo y no en la holo. Por otro lado existía la necesiad de conocer la naturaleza del proceso así como desnetrañar porque se producía entre Trp70 y Asp62 y no entre Trp70 y Glu33 que se encuentra más cerca del triptófano y tiene la misma estructura química que el Asp62.

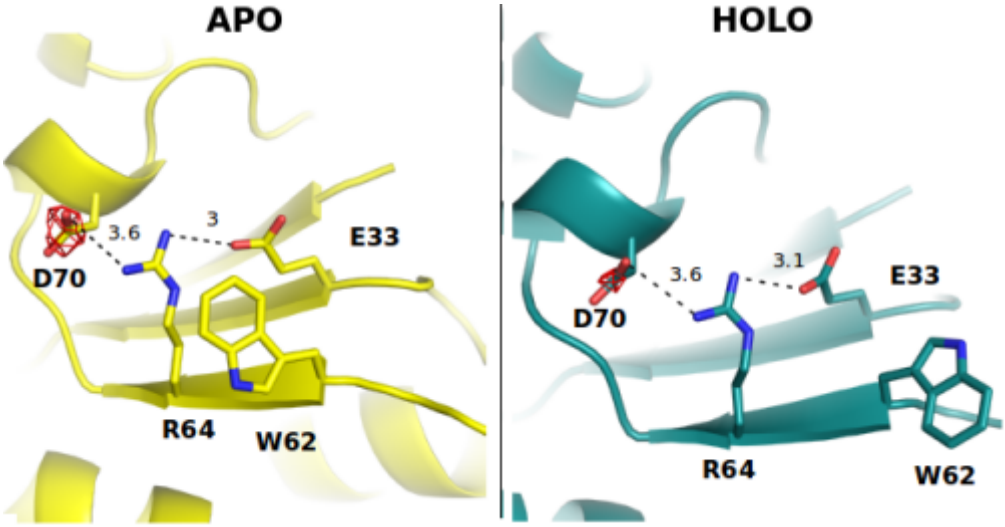


Figure 6.5: Represenatción de los residuos del centro activo de las formas apo y holo de la LDH. En ambos están representados Trp62, Glu33, Arg64 y Asp70.

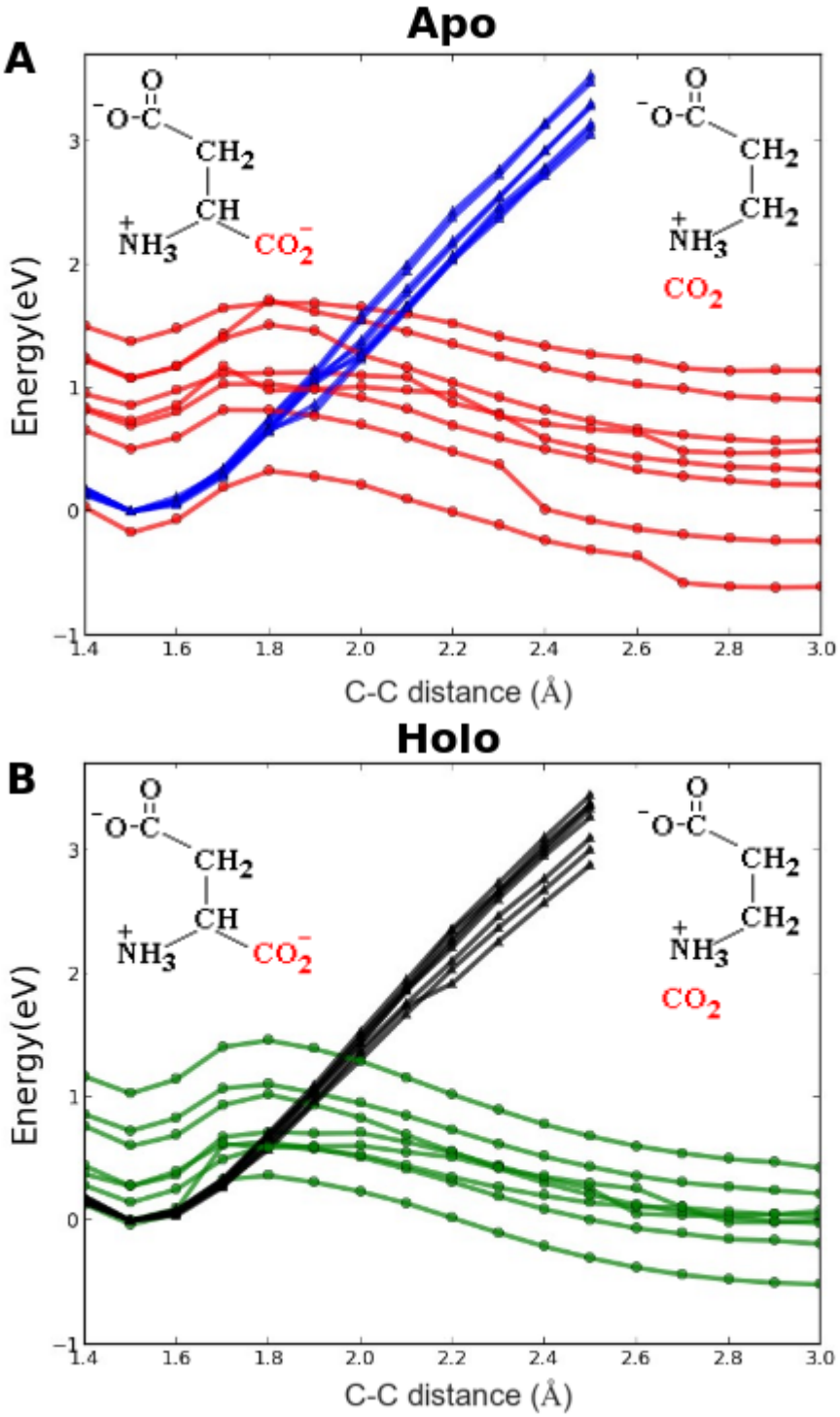


Figure 6.6: A) Perfil de energía de la elongación del enlace C-C del Asp70 en la forma apo. Las líneas azules corresponden al estado en que el hole se localiza en el Trp62 y las rojas cuando lo hace en el Asp70. B) Perfil de energía de la elongación del enlace C-C del Asp70 en la forma holo. Las líneas negras corresponden al estado en que el hole se localiza en el Trp62 y las verdes en el Asp70.

Para responder a todas estas cuestiones realizamos cálculos de MD sobre las estructruas cristalinas de las formas apo y holo, seguidas de cálculos QM/MM (scans de la coordenada de reacción) a nivel DFT usando el funcional BHLYP sobre algunas imágenes de las trayectorias generadas por MD (8 para apo y 8 para holo). Sobre los caminos de reacción realizamos cálculos de acoplamiento electrónico por medio del método de diferenciación de carga fragmentada (FCD).

En primer lugar, confirmamos que el proceso tenía lugar. Por medio del cálculo de potenciales de ionizacion (IP) de acuerdo al teoréma de Koopmans, observamos que el Glu33 tiene un IP mayor que el Asp70. De hecho el IP más bajo es el del Trp62 seguido del Asp70, por ello la transferencia se da entre estos dos aminoácidos, tal como observaron Weik y sus colaboradores. Además confirmamos que termodinámicamente el proceso es plausible en ambas conformaciones, apo y holo.

Tras conocer la naturaleza del proceso, testeamos los diversos mecanismos de HT que se dan de forma habitual en sistemas biológicos: Mecanismo de transferencia directa o mecanismo de puente asistido y dentro de este último los mecanismos de superintercambio y de transferencia secuencial o hopping. El primero que descartamos fué el de transferencia driecta, porque incluso la distancia más baja entre Trp y Asp es demasiado alta en ambas conformaciones: 12.40Åen la forma apo y 7.50Åen la holo, para que se produzca un solapaiento entre los orbitales del dador y el aceptor. Esto da lugar a unos acoplamientos electrónicos negligibles.

EL mecanismo de transferencia secuencial tambien se descartó, porque los cálculos de IP habían demostrado que la transferencia se hacía entre Asp y Trp. Este mecánismo implica que la carga sea transferida entre las especies intermedias, es decir, debería pasar del Trp al Glu, de ahí a la Arg y finalmente al Asp, lo cual no es posible. Así pues el mecanismo por el que se produce es por el de superintercambio. La carga viaja del Trp al Asp a través de los orbitales del Glu y la Arg pero sin interaccionar directamente con ellos. Los acoplamientos electrónicos para el proceso de superintercambio revelaron que en la forma apo el valor de estos es de alrededor de 10^{-6} eV mientras que para la forma holo es más pequeño, alrededor de dos órdenes de magnitud menor. Esto nos indica que es un proceso muy lento, tanto que no es capaz de ser visto en la escala de tiempo que puede abarcar los experimentos de cristalografía de rayos X. Por ello se ve la descarboxilación en la forma apo y no en la holo.

Efecto de la cooperatividad en la estructura secundaria de las proteínas

La estructura secundaria es un elemento importante de las IDPs, como ya hemos comentado. Algunas regiones de las IDPs (llamadas MoRFs) pueden adoptar configuraciones de estructura secundaria transitoria. Cuando generamos conjuntos

estructurales de IDPs, generalente, es difícil visualizar su composición. A veces las propensidades conformacionales de residuos individuales ocultan la naturaleza de las estructuras cooperativas. Por ello, es necesario diferenciar entre cuando un fragmento tiene regiones que adoptan una conformación secundaria y cuando ese fragmento contiene una estructura secundaria completa, todos los residuos del fragmento adoptan dicha conformación (al mismo tiempo). Ambos escenarios son plausibles y por medio de experimentos de RDCs se puede ditinguir entre ellos así como gracias a SS-map visualizar facilmente las diferencias estructurales entre ambos escenarios.

SS-map es un algoritmo escrito en python que se puede obtener de froma gratuita en 'code.google. com/p/ss-map/' y que está diseñado para representar la cooperatividad o las correlaciones de las conformaciones de estructura secundaria. Está especialmente orientado al análisis de IDPs, donde el uso de órdenes de contacto o contactos nativos es imposible. Aunque este fue el propósito inicial también es aplicable a proteínas globulares, siendo una herramienta útil para analizar el plegamiento de proteínas pequeñas y péptidos. Gracias a SS-map, se puede arrojar luz en la percepción real de la estructura secundaria proteica.

Para visualizar los elementos de estructura secundaria de las proteínas se utilizan los valores de los ángulos ϕ y ψ, siguiendo el diagrama de Ramachandran. El algoritmo incorpora cuatro definiciones diferentes de los elementos de estructura secundaria en base a distintos diagramas de Ramachandran (diferentes valores de los ángulos ϕ y ψ). Estas definiciones corresponden a las empleadas por los programas DSSP, Profasi, Flexible Meccano y Campari. Además uno puede usar su propia definición de los ángulos ϕ y ψ. El programa, como entrada necesita varios archivos PDB (uno para cada proteína del conjunto conformacional) de los que extrae los ángulos ϕ y ψ o directamente una matriz con los valores de los ángulos ϕ y ψ para cada estructura del conjunto conformacional. Como salida, devuelve una imagen, una matriz numpy y/o un archivo txt que contiene una matriz (o una representación gráfica de esta matriz) que muestra en cuantas estructuras del conjunto (en %) el residuo y se encuenra formando una región estructurada de longitud x.

Aplicamos el método a diferentes tipo de proteínas. Primero estudiamos 2 proteínas plegadas (HPLC-6 y GB1m2) cerca de su temperatura de fusión, utilizando Profasi, para comparar nuestro programa con otras técnicas tradicionales de visualización. Luego, analizamos MORFs del virus measles y de la nucleoproteína del virus Sendai cuyos conjuntos estructurales se calcularon utilizando el flexible meccano (para poder comparar nuestro resultados con los del grupo de Blackledge). Finalmente, estudiamos la existencia de las hélices de poliprolina II (PPII) en IDPs a partir de datos proporcionados por el Prof. Rohit Pappu (para comprobar los resultados del SS-map con los obtenidos por su grupo).

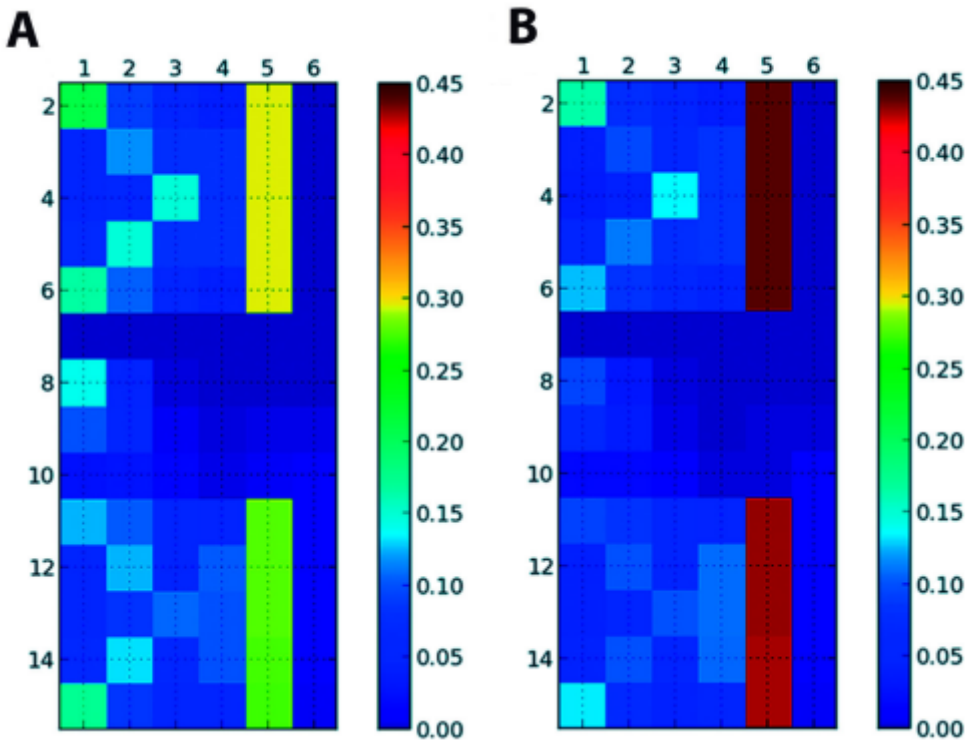


Figure 6.7: Representación de SS-map mostrando las láminas β de la proteína GB1p A) por debajo de la temperatura de plegamiento 319K y B) por encima 327K.

(experimental o sintético)

Nuestros resultado mostraron que para HPLC-6 con los métodos de visualización tradicionales se observa que el porcentaje de $\alpha-$hélice (las cuales emergen desde un residuo central) para cada péptido disminuye gradualmente con la temperatura, y también que la longitud de las hélices se acorta con la temperatura lo cual hace que prevalezca la presencia de fragmentos cortos sobre los largos. Usando SS-map vimos que esto no es cierto y que los segmentos de α hélice tanto largos como cortos son igual de frecuentes.

Después estudiamos la proteína GB1m2 que posee una estructura que presenta un giro β y tiene una temperatura de fusión muy similar, 324 K, a la de HPLC-6. SS-map fue capaz de reproducir su estructura, dos regiones de láminas β unidas por una región central correspondiente a un giro beta. Además, se observó que las láminas β se comportan de forma diferente a las hélices α respecto a la temperatura. Finalmente estudiamos las hélices de PPII. Nos dimos cuenta, al igual que con las láminas β, que estas hélices se comportan diferente a las α. No crecen a partir de un residuo central. Se puede concluir que cada tipo de estructura secundaria presenta sus propias características y patrones estructurales, y que con SS-map podemos diferenciarlos.

Además, estudiamos dos IDPs correspondientes a los virus Sendai y measles. Gracias a SS-map mostramos que la situación es más compleja de lo que Blackledge y sus colaboradores habían mostrado. Las hélices de un par de MORFs (H1 y H2 en Sendai, y H2 y H3 en measles) se mezclan y forman una hélice superior. Gracias a SS-map mostramos la cooperatividad de los elementos de estructura secundaria en estas dos IDPs, efecto que no es visible con lo métodos tradicionales.

Determinación de conjuntos conformacionales de IDPs a partir de RDCs

Una forma habitual de estudiar la dinámica de las proteínas a nivel global, es por medio del uso de conjuntos ('ensembles') conformacionales. Estos 'ensembles' son modelos computacionales, generalmente restringidos por valores experimentales, que describen la estructura de las proteínas. Son herramientas muy potentes para representar el rango de conformaciones que pueden ser sampledas por las proteínas, por lo tanto, permitiendo la representación explicita de la dinámica de las mismas.

Los conjuntos conformacionales se han utilizado para estudiar diferentes aspectos relacionados con propiedades fundamentales de las proteínas como procesos de reconocimieno molecular o de plegamiento proteico. Este tipo de representaciones estructurales no son capaces de describir el ratio de intercambio entre confórmeros o la escala de tiempo de la dinámica, pero como contrapunto informan sobre la amplitud de la dinámica además de sobre diferentes características del comportamiento proteico.

El uso de este tipo de representaciones estructurales es muy útil en el estudio de las proteinas intrínsecamente desordenadas (IDPs), ya que no pueden ser caracteriadas por métodos clásicos de cristalografía de rayos X o por microscopía crioelectrónica. Las IDPs son una familia de proteínas que no cumplen el paradigma tradicional de secuencia-estructura-función, ya que no presentan una estructura plegada. Esta falta de estructura estable, que puede darse en toda la proteína o solo en algunas regiones, les proporciona una plasticidad estructural, imposible de alcanzar por proteínas ordenadas, esencial para llevar a cabo su función celular.

Las IDPs están relacioandas con una amplia gama de enfermedades. Debido a ello son candidatas perfectas para ser dianas terapeúticas, sin embargo no es así. El motivo no es otro que el desconocimiento sobre como estas proteinas realizan su función. Su estructura y su dinámica están ampliamente relacionadas con su unión a ligandos, ya que se trata de proteinas con mútiples dominios de unión. Este fenómeno de unión es importante para la promiscuidad funcional y la regulación de estas proteínas. Por ejemplo bajo la unión a uno o varios ligandos exhiben transiciones orden-desorden (aunque no todas las IDPs lo hacen); adoptan estructura secundaria de forma transitoria. Por ello, la caracterización del amplio

rango de estructuras que pueden adoptar estas proteínas es clave para entender sus propiedades funcionales y conformacionales y las enfermedades en que están implicadas. A esta compensión han contribuido de forma significativa estudios experimentales de RMN así como computacionaes por medio de modelización atomística y de grano grueso ('coarse grained').

Aunque el uso de conjuntos conformacionales es muy común para el estudio de la dinámica de las proteínas, no por ello deja de presentar problemas. Uno muy común es cuando al simular un sistema proteico queremos comparar los valores obtenidos con los de un sistema de referencia para validar las simulaciones. Frecuentemente los valores no coinciden. No hay una sola razón por la cual esto sucede. Por ejemplo, puede ser un problema del campo de fuerza empleado o un problema de muestreo debido a limitaciones de tiempo de computación. Muchos esfuerzos se están centrando en mejorar los campos de fuerza incorporando datos experimentales. También se está avanzando en el desarrollo de métodos que permitan recalcular las estructuras generadas en base a ciertos valores (observables) de un sistema de referencia.

Tratando de arrojar algo de luz en este sentido hemos creado un algoritmo (MaxEnt) que incorpora el principio de máxima entropía para dado un conjunto de RDCs (datos experimentales o simulados) hacer que concuerden con un segundo conjunto (valores de referencia). El principio de máxima entropía es un conocido método estadístico que se basa en la minimización de la información incluida en un conjunto de datos para adaptarse a ciertos observables. El hecho de utilizar RDCs es debido a que estos constituyen una técnica muy adecuada para caracterizar elementos de estructura secundaria en IDPS.

MaxEnt es un algoritmo escrito en python que se puede obtener de forma gratuita desde GitHub ('https: // github.com/ MelchorSanchez / MaxEnt '). A partir de un conjunto primario de datos de RDCs de referencia, dado un segundo conjunto (generalmente calculado a partir de Dinámica Molecular o simulaciones de Monte Carlo), estos se recalculan y minimizan de acuerdo a un ajuste respecto del gradiente. MaxEnt necesita como datos de entrada una matriz M x N y un vector de valores N. La matriz M x N debe contener los N RDCs de las M estructuras a recalcular, y el vector de valores N debe contener los N RDCs de referencia.

Para simplificar la generación de la matriz M x N de RDCs también creamos el script de python RunPales que está disponible en el mismo directorio de github que MaxEnt. RunPales es un script que puede llamar al ejecutable del programa PALES (un software utilizado para calcular RDCs a partir de estructuras simuladas), generar los RDCs correspondientes y guardar sus valores en una matriz. En otras palabras, RunPales no es más que una interfaz que llama al PALES con las opciones adecuadas, y almacena los resultados generados.

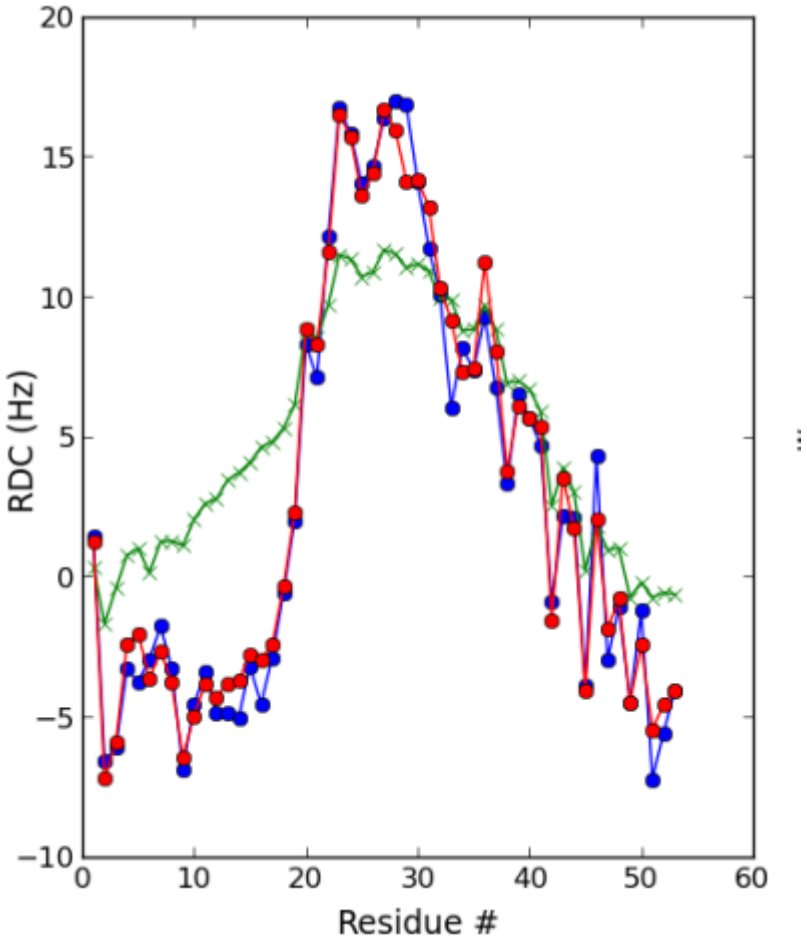


Figure 6.8: Ajuste por medio de MaxEnt de un conjunto de conformaciones generado con Profasi a 325K T1=325.6K a otro conjunto de valores experimenales de RDCs (línea azul). El conjunto estructural sin ajustar (línea verde) presenta una región con demasiadas hélices alfa entre los residuos 32 y 40 en comparación con el ajustado (optimizado) (línea roja).

Nuestra aplicación del principio de máxima entropía presentan algunos puntos interesantes. En primer lugar hemos realizado una implementación que es invariable respecto a la escala del observable para trabajar con RDCs. De esta manera, MaxEnt puede ser utilizado por diferentes grupos experimentales empleando diferentes conjuntos de valores de RDCs, ya que se puede utilizar con cualquier conjunto de estructuras. Se basa en los valores de RDC no en la manera en que se generan. Debido a la independencia estructual evitamos el riesgo de problemas de sobreajuste ya que el número de parámetros se basa únicamente en el número de datos experimentales. Finalmente otra característica importante es su velocidad: puede tratar miles de estructuras y convergir en unos pocos segundos.

Para testear el algoritmo utilizamos varios conjuntos de datos simulados y

experimentales del dominio de unión de la nucleoproteína del virus Sendai. Los datos calculados se obtuvieron a partir de simulaciones de monte carlo utilizando los campos de fuerza Profasi y Campari. Nuestros resultados mostraron que a pesar de sus limitaciones, ambos pueden generar conjuntos de estructuras razonables y mejores que otros métodos ('random coil') muy utilizados para estudiar IDPs, y concretamente esta proteína. Además muestran un mayor poder predictivo que estos. Sin embargo, las estructuras generadas no son perfectas y se hace necesario modificarlas ligeramente para poder ajustarlas a los datos 'objetivo' que se querían reproducir. Por medio del MaxEnt lo logramos. También nos dimos cuenta de que para lograr un buen ajuste y poder reproducir estos valores, es necesario que el campo fuerza con el que se generan las estructuras calculadas sea lo suficientemente preciso. Sino es así incluso usando los datos objetivo no se loogrará generar una estructura representativa. Es por ello que hay que remarcar y tener muy en cuenta el campo de fuerza utilizado, algo que generalmente se subestima.

Conclusiones

Movimientos locales

- La capacidad catalítica de EcNAGK radica en los movimientos (colectivos) de apertura y cierre que permiten acceder a conformaciones cuyo centro activo está orientado adecuadamente y altamente comprimido. Estas evidencias, apoyan la hipótesis de la compresión 'conformacional' formulada por Rubio y sus colaboradores

- Al menos para la enzima EcNAGK los movimiento proteicos que dan lugar a las conformaciones catalíticas son el proceso limitante en lugar del paso químico.

- El método de enjambre de trayectorias (SoT) puede ser aplicado al estudio de la catálisis enzimática porque es independiente de la dinámica de las variables colectivas (CV). Nosotros, encontramos la implementación adecuada para obtener la contribución de cada variables colectiva al perfil de energía libre.

- Dentro del método SoT el uso de una CV sin ninguna función en la reacción química no afecta el perfil de energía libre ni el coste computacional de la simulación. Sin embargo el omitir una CV importante para la descripción de la reacción química es relevante ya que puede provocar que se subestime la barrera de energía libre.

- Las proteínas con residuos ácidos son susceptibles de sufrir daños por radiación cuando están expuestas a fuentes de alta radiación como lo son las técnicas de sincrotrón. En relación con la proteína LDH este daño se traduce en una descarboxilación. Este proceso es un HT que tiene lugar por medio de un fenómeno de superintercambio.

Movimientos globales

- El principio de máxima entropía es una buena opción para introducir información experimental en un conjunto conformacional. Nosotros desarrollamos una implementación independiente de escala del mismo, para estudiar RDCs (aunque potencialemnte cualquier observable puede ser analizado debido a la mentada invariabilidad de escala). Nuestra implementación además permite evitar el sobreajustamiento así como acelerar su ejecución (permite tratar miles de estructuras en segundos).

- Los métodos de visualización tradicionales dificultan la observación de fenómenos de cooperatividad en elementos de estructura secundaria dando lugar a interpretaciones incorrectas o incompletas de sus propensidades tanto en proteínas globulares como IDPs.

- Los diferentes elementos de estructrua secundaria presentan distintos valores de ángulos ϕ y ψ en el diagrama de Ramachandran así como diferentes 'firmas' químicas entre ellos. Las hélices α y las láminas β se comportan de forma diferente respecto a la temperatura y las hélices de PPII no crecen a partir de un residuo central, de forma opuesta a las hélices α.

De todos estos estudios, de forma general, podemos concluir que los métodos computacionales son una herramienta eficaz y útil para caracterizar la dinámica de las proteínas. Sin embargo, los métodos computacionales actuales presentan limitaciones y para resolverlos la incorporación de datos experimentales así como su correcta interpretación es crucial. Pero aunque los métodos computacionales necesitan de los experimentales, esta necesidad también se da de manera opuesta. Por ejemplo, hay métodos experimentales como los utilizados para estudiar IDPs, RMN, que solo nos permiten conocer diferentes conjuntos estructurales, como si fueran fotos de la evolución del sistema a diferentes pasos de tiempo. Para caaracterizar el paisaje conformacional completo es necesario integrar todos estos datos por medio de la computación.

La convergencia de los métodos experimentales y computacionales es clave para poder profundizar en el conocmiento de la dinámica de las proteínas.

Chapter 7

Appendix

Scientific Production

Papers

- Jelisa Iglesias; Melchor Sanchez-Martinez; Ramon Crehuet. *SS-map: Visualizing cooperative secondary structure elements in protein ensembles.* Intrinsically Disordered Proteins 2013; 1:e25323; http://dx.doi.org/10.4161/idp.25323

- Melchor Sanchez-Martinez; Enrique Marcos; Romà Tauler; Martin J. Field; Ramon Crehuet. *Conformational compression and barrier height heterogeneity in the N-acetylglutamate kinase.* J. Phys. Chem. B, 2013, 117 (46), 14261-14272 DOI: 10.1021/jp407016v

- Melchor Sanchez-Martinez; Martin J. Field; Ramon Crehuet. *Enzymatic Minimum Free Energy Path Calculations Using Swarms of Trajectories*, J. Phys. Chem. B **Article ASAP**, DOI: 10.1021/jp506593

- Melchor Sanchez-Martinez; Ramon Crehuet. *Applying the Maximum Entropy Principle to determine ensembles of Intrinsically Disordered Proteins from Residual Dipolar Couplings*, Phys. Chem. Chem. Phys. **Advance Article**, DOI: 10.1039/C4CP03114H

- Melchor Sanchez-Martinez[†]; Nicolas Coquelle[†]; Alexander A. Voityuk; Martin Weik; Ramon Crehuet. *Tryptophan-mediated decarboxylation in proteins: evidence from X-ray crystallography and QM/MM simulations*, **in preparation, [†] co-first authors**

Book chapters

- Enrique Marcos, Melchor Sanchez-Martinez, Ramon Crehuet. *Interplay between enzyme function and protein dynamics. A multi-scale approach to the study of the NAG kinase family and two class II aldolases*, Computational Approaches to Protein Dynamics: From Quantum to Coarse-Grained Methods. Series in Computational Biophysics. Taylor and Francis, **In press**, ISBN 9781466561571

Oral Communications

- *Influence of the conformational dynamics on the activation of the Nacetyl-l-glutamate-kinase.* Melchor Sanchez-Martinez*, Enrique Marcos, Martin J. Field, Ramon Crehuet, XXVIII XRQTC Annual Meeting, 11-12/06/2012, Barcelona

- *Catalytic role of protein motions in NAG kinase.* Ramon Crehuet*, Melchor Sanchez-Martinez, Enrique Marcos, Martin J. Field, XXII Congress SBE/International Congress of SBE. Barcelona, 3-6/07/2012

- *Swarms of Trajectories applied to enzyme catalysis.* Melchor Sanchez-Martinez*, Martin J. Field, Ramon Crehuet, XXX XRQTC Annual Meeting, 26-27/06/2014, Barcelona

Posters

- *Tryptophan mediated decarboxylation caused by X-ray damage in LDH under Crystallographic conditions.* Melchor Sanchez-Martinez*, Nicolas Coquelle, Martin Weik, Ramon Crehuet, Ninth Triennial Congress of the World Association of Theoretical and Computational Chemists (WATOC), 17-22/07/2011, Santiago de Compostela

- *Tryptophan mediated decarboxylation caused by X-ray damage in LDH under Crystallographic conditions.* Melchor Sanchez-Martinez*, Nicolas Coquelle, Martin Weik, Ramon Crehuet, The Catalonian Supercomputing day: Chemistry, Computation and Society, Barcelona, Spain, 26/10/2011

- *Does 'conformational compression' help enzyme catalysis? Influence of the conformational dynamics on Nacetyl- l-glutamate-kinase.* Melchor Sanchez-Martinez*, Enrique Marcos, Martin J. Field, Ramon Crehuet, X Girona Seminar, 2-5/07/2012, Girona

- *Does 'conformational compression' help enzyme catalysis? Influence of the conformational dynamics on Nacetyl- l-glutamate-kinase.* Ramon Crehuet*, Melchor Sanchez-Martinez, Enrique Marcos, Martin J. Field, 8th congress on Electronic Structure: Principles and Applications (ESPA), 26-29/06/2012, Barcelona

- *Intrinsically Disordered Proteins (IDPs). An Emerging family of proteins.* Jelisa Iglesias*, Melchor Sanchez-Martinez, Ramon Crehuet, III New trends in Computational Chemistry for Industry Applications, Barcelona, Spain, 23-24/05/2013

- *Visualizing and re-weighting ensembles of IDPs*: *SS-map and Maximum Entropy principle applied to fit RDC data.* Ramon Crehuet*, Jelisa Iglesias, Melchor Sanchez-Martinez, Intrinsically Disordered Proteins: Connecting Computation, Physics and Biology, 2-5/09/2013, Lugano

  * : Presenting author

Awards and Honours

- Poster prize at the Ninth Triennial Congress of the World Association of Theoretical and Computational Chemists(WATOC), 17-22/07/2011