A Discrete Molecular Dynamics Approach to the Study of Disordered and Aggregating Proteins

Agustí Emperador[1*] and Modesto Orozco[1,2.3*]

We present a refinement of the Coarse Grained PACSAB force-field for Discrete Molecular Dynamics (DMD) simulations of proteins in aqueous conditions. As the original version, the refined method provides good representation of the structure and dynamics of folded proteins, but provides much better representations of a variety of unfolded proteins, including some very large, impossible to analyze by atomistic simulation methods. The PACSAB/DMD method also reproduces accurately aggregation properties, providing good pictures of the structural ensembles of proteins showing a folded core and an intrinsically disordered region. The combination of accuracy and speed makes the method presented here a good alternative for the exploration of unstructured protein systems.

[1] Institute for Research in Biomedicine (IRB) Barcelona. The Barcelona Institute of Science and Technology. Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028

[2] Joint IRB-BSC Program on Computational Biology. Barcelona. Spain

[3] Departament de Bioquímica i Biomedicina, Facultat de Biología, Avgda Diagonal 647, Barcelona 08028, Spain

* Corresponding authors agusti.emperador@irbbarcelona.org, modesto.orozco@irbbarcelona.org

INTRODUCTION

The quality of biomolecular simulation is limited by two main factors: the accuracy in the representation of molecular interactions and the quality of the sampling. Current atomistic force fields implemented in molecular dynamics (MD) algorithms allow the collection of reasonable samplings in the multi-microsecond regime for systems containing in the order of 10^4-10^5 atoms[1], which has made possible the representation of some fast conformational movements in small/medium proteins and even the *ab initio* folding of a few small proteins[2,3]. Unfortunately, the times when MD simulations will be applicable to the study of large systems (above 10^6 atoms) for long (above millisecond) periods of time are still far, which hampers our ability to study complex phenomena, such as protein aggregation, association and dissociation or conformational sampling of large disordered proteins.

The coarse-graining (CG) approach provides a simple strategy to improve sampling, at the expense of a certain loss of accuracy[4]. The main idea behind all CG methods is to reduce the complexity of the system by grouping atoms into beads, whose interactions are presented by a simple energy functional, which typically include solvent in an implicit way[5,6]. The use of CG methods largely accelerate calculations due to the combination of the reduction in the number of particles, the neglect of fast movements and the reduced cost of energy evaluations. The advantages are more evident in systems where the volume fraction of water is very high, like unfolded proteins, because the removal of solvent molecules not only reduces the degrees of freedom, but also the viscosity, facilitating the representation of large conformational changes[7]. The dark side of CG methods is that the energy functional and the sampling strategy require a careful parametrization using structural experimental data, which means that very often protein CG methods are overspecialized to reproduce the structure of folded proteins, those for which more information exist. This specialization generates a transferability problem, as a method very efficient to represent a well-folded protein under diluted aqueous conditions might be unable to reproduce unfolded proteins.

We present here a recalibration of our discrete molecular dynamics DMD/PACSAB force-field for simulations of proteins in aqueous solution[8,9]. The refined method maintains the good ability of the original PACSAB force-field to describe folded proteins, but show much improved representations of unfolded proteins and is able to produce reversible protein-protein binding[10], reproducing correctly dimerization and association-dissociation processes.

METHODS

The basic DMD formalism assumes that macromolecules are a set of particles moving at constant velocity (i.e. in the absence of forces) in a space limited by square wells defined by discontinuous potentials. Within this assumption particles move in a fully predictable way until a collision happens:

$$\vec{r}_i(t+t_c) = \vec{r}_i(t) + \vec{v}_i(t)t_c\,, \qquad (1),$$

where $\vec{r}_i$ and $\vec{v}_i$ stand for positions and velocities and t_c is the minimum amongst the collision times t_{ij} between each pair of particles i and j:

$$t_{ij} = \frac{-b_{ij} \pm \sqrt{{b_{ij}}^2 - {v_{ij}}^2({r_{ij}}^2 - d^2)}}{{v_{ij}}^2}\,, \qquad (2),$$

where r_{ij} is the modulus of $\vec{r}_{ij} = \vec{r}_j - \vec{r}_i$, v_{ij} is the modulus of $\vec{v}_{ij} = \vec{v}_j - \vec{v}_i$, $b_{ij} = \vec{r}_{ij} \cdot \vec{v}_{ij}$, and d is the distance corresponding to the wall of the square well.

When two particles collide in an elastic way, there is a transfer of linear momentum into the direction of the vector $\vec{r}_{ij}$:

$$\begin{aligned} m_i\vec{v}_i &= m_i\vec{v}_i' + \Delta\vec{p} \\ m_j\vec{v}_j + \Delta\vec{p} &= m_j\vec{v}_j' \end{aligned} \qquad (3),$$

where the prime indices denote the velocities after the collision.

In order to calculate the change in velocities upon collision the velocity of each particle is projected in the direction of the vector $\vec{r}_{ij}$ and conservation rules are applied:

$$m_i v_i + m_j v_j = m_i v_i' + m_j v_j' \tag{4}$$

$$\frac{1}{2} m_i v_i^2 + \frac{1}{2} m_j v_j^2 = \frac{1}{2} m_i v_i'^2 + \frac{1}{2} m_j v_j'^2 + \Delta V, \tag{5),}$$

where ΔV stands for the depth of the square well defining the inter-atomic potential.

The transferred momentum can be easily determined from;

$$\Delta p = \frac{m_i m_j}{m_i + m_j} \left\{ \sqrt{(v_j - v_i)^2 - 2\frac{m_i + m_j}{m_i m_j} \Delta V} - (v_j - v_i) \right\}, \tag{6}$$

Note that the two particles overcome the potential step ΔV as long as

$$\Delta V < \frac{m_1 m_2}{2(m_1 + m_2)} (v_j - v_i)^2 \tag{7}$$

Otherwise, if the particles remain in the well Eq. 6 reduces to:

$$\Delta p = \frac{m_i m_j}{m_i + m_j} \left\{ \sqrt{(v_j - v_i)^2} - (v_j - v_i) \right\} \tag{8),}$$

which taking the negative solution of the root leads to:

$$\Delta p = \frac{2 m_i m_j}{m_i + m_j} (v_i - v_j). \tag{9}$$

The DMD implementation used in this work runs in the isothermal ensemble, and the system is coupled to an external thermal bath using an Andersen thermostat[11]. Note that under the DMD paradigm no forces should be calculated, neither the equations of motion should be integrated. If an efficient algorithm for predicting collisions is used, the method can be extremely efficient allowing simulation of trajectories for very long time periods[12]. DMD has been shown very powerful to study protein flexibility[13,14], conformational transitions[15], ab initio folding[11], aggregation[16,17,18] and protein-protein docking[19].

The resolution level and the energy functional used in DMD simulations should balance accuracy with simplicity as complex potentials lead to many steps and accordingly to the increase in the number of potential collisions, making t_c (Eq. 1) small and the entire DMD calculation inefficient. Our PACSAB[9] approach uses a full description of the backbone, but compresses the side chain atoms into beads following MARTINI model[20] for proteins. The associated force-field consists of “bonded” and “non-bonded” terms. Chemical bonds and bond angles are fixed with narrow square well potentials whose width corresponds to 5% of the length of the bond/pseudobond distance[13]. We also use pseudobonds to fix the dihedral angle of the peptide bonds, in order to enforce its planar geometry. The interactions between non-bonded particles comprise hydrogen bonding between atoms in the amide groups of the backbone[11] and a discretized version of the interaction between the coarse-grained sidechain beads, constructed assuming pairwise additivity of the atomistic van der Waals and implicit solvation terms[9]. The atomistic implicit solvation term was defined with the EEF1 energy functional of Lazaridis and Karplus[11,13,21].

The original parametrization of PACSAB[9] was mostly directed towards representing folded proteins and shows slightly worse performance for disordered proteins, mimicking the situation found for currently available atomistic force-fields which tend to collapse unfolded proteins[22], due probably to an improper balance of solute-water interactions which leads to an unbalance in association/dissociation rates[10,23,24]. In order to correct these problems we implement here a dual description of non-bonded interactions by dividing them in “short-range” and “long-range”. Both non-bonded functionals have the same form, but different parameters, and are combined by:

$$V=\chi V_{\text{short-range}}+(1-\chi)V_{\text{long-range}} \qquad (10)$$

The switching function between two beads takes the form:

$$\chi(i,j) = \frac{1}{1+\exp((d(i,j)-\rho)/\eta)} \tag{11}$$

where d(i,j) is the distance between the two particles in the experimental reference structure (when available), or d(i,j)=|i-j|3 Å (i, j being the residue indexes) for unfolded proteins. After some initial tests the constants ρ and η were adjusted to 10 and 2 Å respectively.

The “short-range” potential was parametrized using three representative proteins (fasciculin (PDB id 1FAS), yeast copper transported (PDB id 1FVQ), and alcohol-binding protein LUSH (PDB id 1OOI), while the “long-range” potential was independently calibrated to reproduce the monomer/dimer ratio in an 8.5 mM solution of villin (PDB id 1VII). The parameters for the interpolation between the short-range and the long-range parametrizations were adjusted from simulations of the disordered protein ACTR. We have fitted ρ in Eq. 11 by searching the maximum value for which the radius of gyration of ACTR stays close to the experimental estimate from SAXS measurements (higher values of ρ give more strength to the short-range parametrization, reinforcing the stability of folded proteins but collapsing the structural ensembles of IDPs). To show the importance of using the dual parametrization for the simulation of unfolded proteins, we have plotted in Suppl. Figure S1 the radius of gyration of ACTR when using the short-range parametrization and the dual one. Once refined, the composite non-bonded term was tested without further correction in a variety of folded, unfolded, diluted and concentrated systems. In all cases we have run conventional DMD simulations at T=300K.

Analysis of the trajectories was performed using standard analysis tools in FlexServ[25] and MDWEB[26]. Atomistic trajectories for some proteins were extracted from the MODEL database[27]. The similarity between DMD/PACSAB and atomistic MD deformation spaces is computed by using Hess metrics on the essential spaces defined by the eigenvectors needed to represent 90% of variance[14]

$$\gamma_{XY} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}(\nu_i^X \bullet \nu_j^Y)^2 \tag{12}$$

where *X* and *Y* index the two methods to be compared, *i* and *j* index the eigenvectors (ranked on the basis of their contribution to structural variance), and *m* is the number of eigenvectors in the “important space” (that required to explain 90% of variance in our work). We have corrected the absolute similarity index for limited simulation time artifacts by including self-similarity terms, to provide a global estimate of the similarity as described elsewhere[14]:

$$\Gamma = \frac{2\sum_{i=1}^{m}\sum_{j=1}^{m}(\nu_i^X \bullet \nu_j^Y)^2}{\sum_{i=1}^{m}\sum_{j=1}^{m}(\nu_i^{Y'} \bullet \nu_j^{Y'})^2 + \sum_{i=1}^{m}\sum_{j=1}^{m}(\nu_i^{X'} \bullet \nu_j^{X'})^2} \quad (13),$$

where the self similarity products $(\nu_i^{Y'} \bullet \nu_j^{Y'})$ are obtained by comparing first and second halves of the trajectories.

RESULTS

PACSAB reproduces well the structure and dynamics of folded proteins: We have collected DMD/PACSAB simulations for 21 folded proteins: six used in the benchmark of the PRIMO coarse-grained model (1VII, 3BG1, 1FKS, 1BTA, 1CYE, 1D3Z), six used in the benchmark of the OPEP coarse-grained model (1AFP, 1B75, 1E0G, 1FCL, 1QHK, 2B86), and fifteen proteins of the MICROMODEL database[28] (1I6F, 1FAS, 1CSP, 1FVQ, 1PHT, 1CQY, 1OPC, 1KTE, 1JLI, 1OOI, 1BFG, 1CHN, 1PDO, 1LIT, 1BJ7). The average RMSDs found between the experimental structure and the conformations sampled along 500 ns (note that due to lack of collision with solvent molecules in our implicit solvent DMD simulations, simulation times are expected to represent much longer periods of “real time”[9,29]) are around 0.04 Å/residue (see figure 1), values which are very similar to those obtained in 50 ns simulations by PRIMO[30] and OPEP[31], parametrized to reproduce exclusively folded proteins, and not far away from those obtained using atomistic force-fields[27,28]. The analysis of RMSD profiles with time illustrates the stability of the trajectories without evidences of unfolding (see Suppl. Figure S2), suggesting that the similarity between DMD trajectories and experimental structures is not a simple equilibration artifact. The RMSD of each protein of the benchmark is reported in the Supplementary Table S1 (the RMSD has been calculated from the position of the Cα atoms). Finally, essential dynamics (ED) analysis[13] of the collected trajectories show that the type of movement sampled here is very close to those obtained by using atomistic simulations in the MICROMODEL dataset[28]. In summary, despite its extreme simplicity the dual PACSAB force-field seems able to reproduce well the structure and dynamic properties of folded proteins.

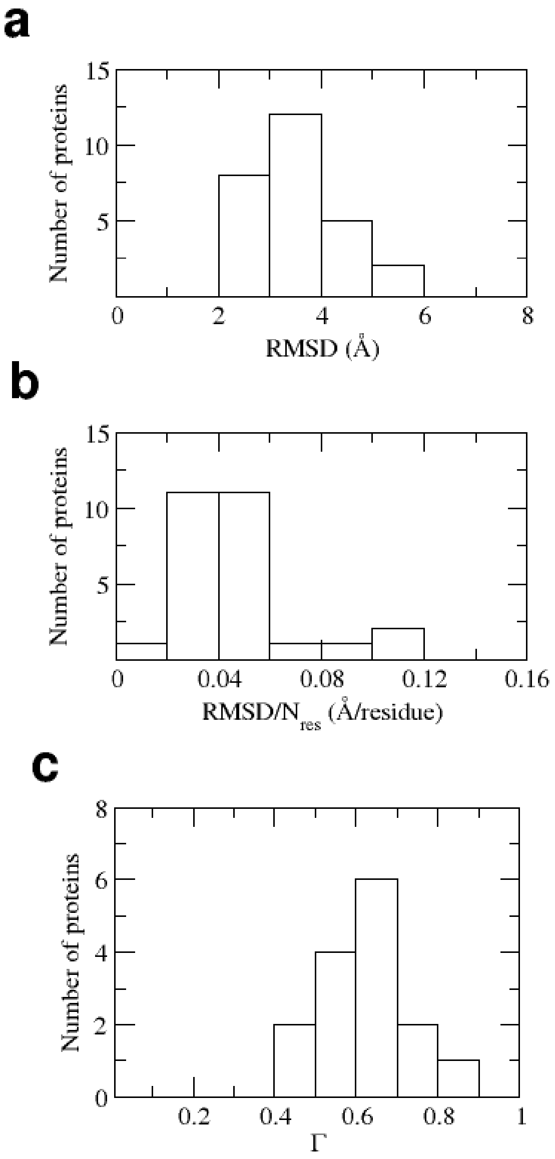

Figure 1. Results for the benchmark of folded proteins used to test the results of the force field. (a) RMSD after a simulation of 500 ns. (b) and RMSD per residue (middle) after a simulation of 500 ns. (c) Distribution of Γ (see main text) for the 15 proteins of MICROMODEL included in the benchmark.

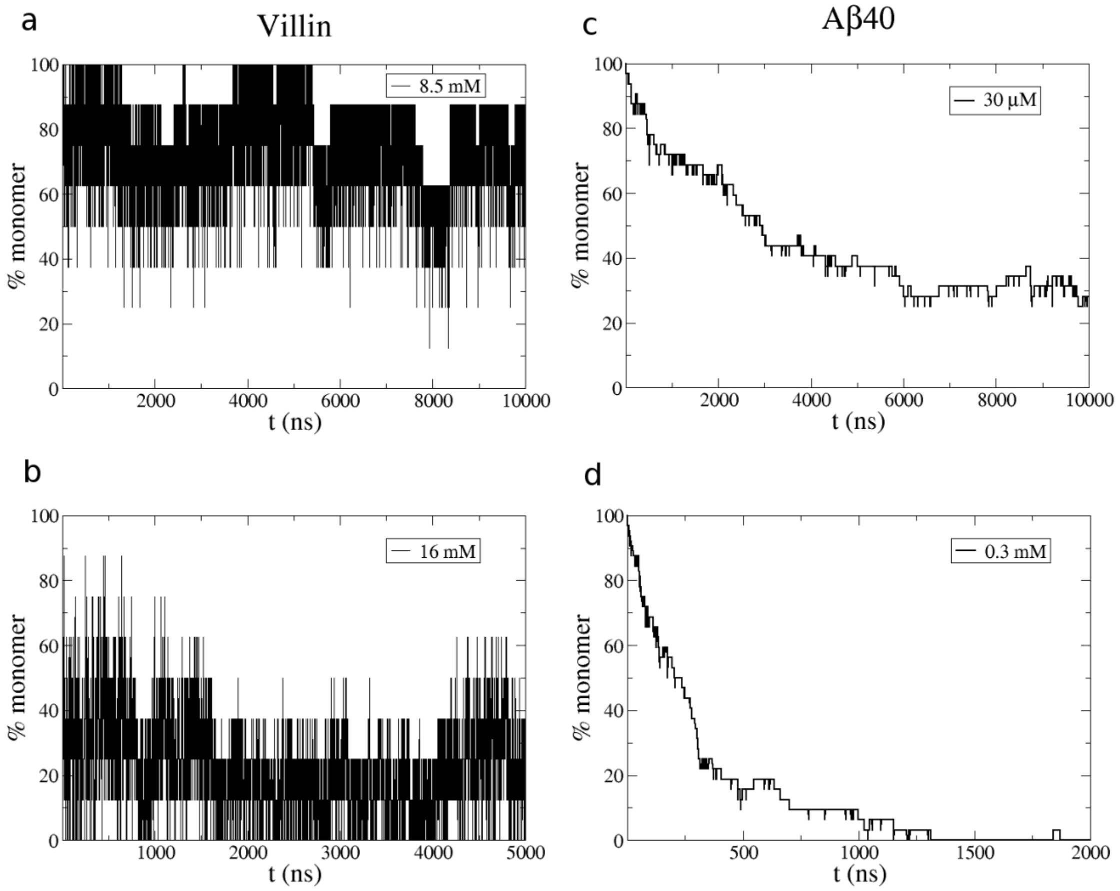

Figure 2. Percentage, averaged over all the trajectories, of the molecules that remain monomers in the simulations of: (a) villin at a concentration of 8.5 mM, (b) villin at a concentration of 16 mM, (c) Aβ40 at a concentration of 30 μM, (d) Aβ40 at a concentration of 0.3 mM.

PACSAB reproduces well dimerization processes: our simulations reproduce well the experimental ratio[32] of monomer/dimer in both 8.5 and 16 mM aqueous solution of villin (see figure 2). To model the solution with the minimal computational cost he have placed two molecules in random relative positions inside a cubic box with periodic boundary conditions, the size of the box being that corresponding to the concentration to be analyzed[9]. We show in Figure 2 the percentage of monomers averaged over the 8 trajectories that we have simulated for each concentration. Very interestingly, the DMD simulations show a good statistics of association/dissociation events, due to the nature of the sampling technique and the lack of solvent molecules, which allows us to reach the stationary state much faster than in explicit solvent atomistic molecular dynamics simulations[24]. As an additional test set we consider the disordered Aβ40 peptide[16], which is known to form in certain conditions amyloid fibrils linked to Alzheimer's disease. We simulated first a 30 μM aqueous solution of the Aβ40 peptide, finding in the stationary regime around 30% of monomer, a value compatible with that inferred from ESI-IM-MS experiments[33]. The size of the cubic box corresponding to this concentration for two molecules is 48 nm. We start the simulations from completely extended conformations. A reversible binding process happening in one of the trajectories is shown in Suppl. Figure S3. Due to the lower concentration that gives a lower collision frequency, for this peptide we have run 32 simulations in order to have a higher statistics of associations and dissociations. The increase of the concentration to 0.3 mM leads to the practical disappearance of the monomer. In summary, even if our force-field has not been created to simulate specifically aggregation of peptides or proteins, it has a reasonable ability to distinguish between monomeric and dimeric states.

PACSAB reproduces well a variety of intrinsically disordered proteins: Despite its simplicity DMD/PACSAB simulations produce good ensembles of a variety of unfolded proteins, which have been found challenging to reproduce by atomistic MD simulations[22,23]. One of this examples is **ACTR**, a 47 residue long intrinsically disordered protein (IDP) which considering SAXS data should adopt an extended (radii of gyration, R_g, around 24 Å[34]) conformation in aqueous solution, but presents very compact structures when studied with standard atomistic MD simulations, unless a refitting of the residue-water potentials is made[23]. DMD/PACSAB simulations provide a fully extended conformation (R_g = 21(1) Å), close to the SAXS estimates (the simulation of ACTR with the original version of PACSAB gives an R_g around 16(1) Å). No significant population of persistent secondary structure elements is found, in good agreement with NMR measurements[35]. Very interestingly, the ensemble collected from 8 DMD/PACSAB simulations starting from the NCBD-bound state of the ACTR protein (PDB id 1KBH), and that obtained from 8 independent trajectories starting from an extended conformation are nearly identical (see Figure 3), confirming the excellent sampling capabilities of DMD/PACSAB simulations.

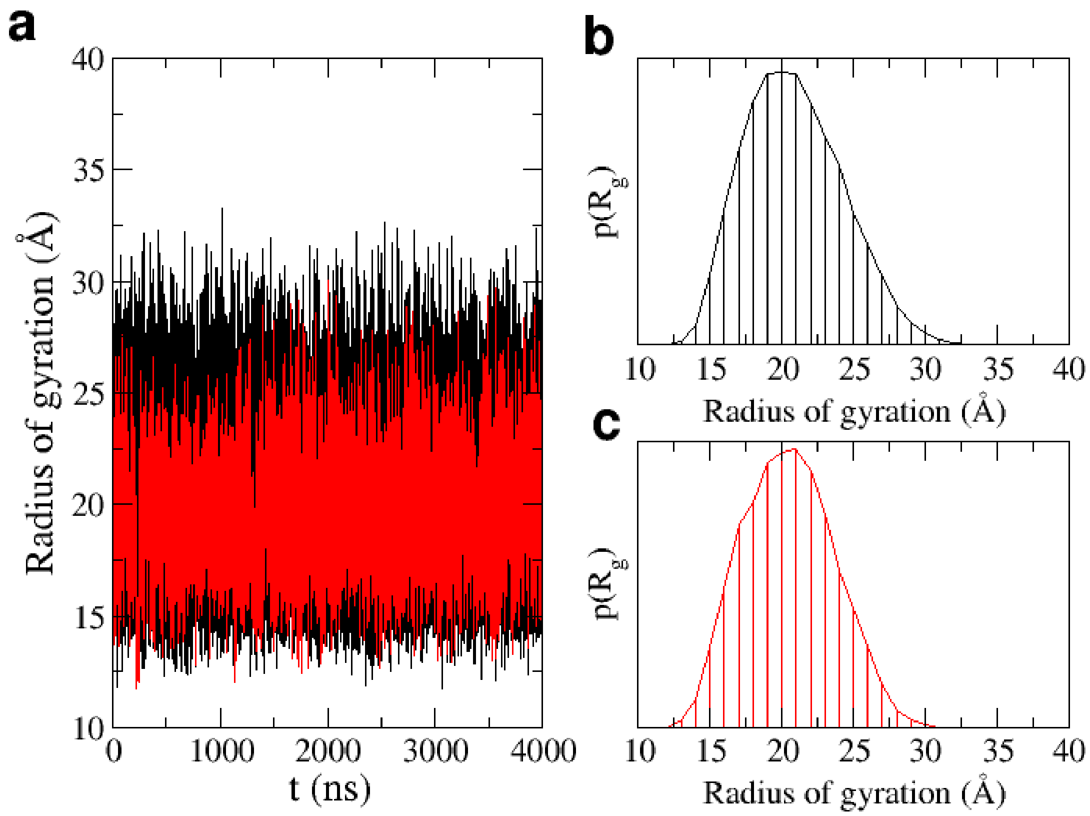

Figure 3. Simulations of ACTR. (a) Radius of gyration in all the trajectories (black lines, simulations starting from an extended coil conformation; red lines, simulations starting from its conformation when bound to NCBD) (b) Distribution of radius of gyration in the simulations starting from an extended coil conformation (c) Distribution of radius of gyration in the simulations starting from its conformation when bound to NCBD.

The larger (140 residues) amyloidogenic protein **α-synuclein** is another example of IDP largely studied due to its role in Parkinson's disease[36]. This protein presents a stable structure when embedded in a lipid environment (PDB id 1XQ8), but when solvated in physiological conditions the protein exists as a mixture of extended conformations[37]. SAXS experiments suggested an average R_g around 35 Å[38], and NMR experiments supported a more compact structure with R_g around 27 Å[39], while more recent paramagnetic relaxation enhancement (PRE) measurements of NMR spectra coupled with atomistic MD simulations favored a wide R_g distribution centered at around 32 Å[40]. Very extended (5-20 μs) atomistic unbiased MD simulations by Shaw's group[22] using a variety of force-fields provided in all cases unrealistically compact structures (R_g

in the range 15-18 Å), and it was necessary to create a new water model (TIP4P-D), where the dispersion interactions of water had been increased[22], to (in practice) reduce hydrophobic interactions and obtain more reasonable results (R_g between 25 and 30 Å). Eight unbiased DMD/PACSAB simulations starting from random extended conformations confirm here the complexity of the conformational ensemble of α-synuclein (we have simulated non-acetylated α-synuclein to compare with the previous simulations[22,40]). We obtain a distribution of radii of gyration with its maximum at 30(1) Å (see figure 4), but analysis of our DMD/PACSAB trajectories (Suppl. Figure S4) suggest the existence of two main states in slow equilibrium: one extended (R_g around 30 Å) and another very extended (R_g around 40 Å), without any evidence of significant population of the compact state suggested by atomistic MD simulations. In the simulations we made of this protein in a previous work with the original PACSAB force field we found a much more compact structural ensemble (R_g = 19(1) Å). Analysis of the inter-residue contacts (figure 5) reveals that the R_g-40 Å state is mostly extended with few persistent inter-residue contacts. On the contrary, the R_g-30 Å state displays a series of transient long range contacts (between the sequence range 40-60 and at the last 30 residues at the C terminal), which were already detected in PRE/NMR/MD studies by Vendruscolo and coworkers[40] and later in a more refined MD post-processing of NMR data by Salvatella's group[37]. We have also computed the mean inter-residue distances as a function of the sequence distance, finding results in very good agreement with the experimental results[41] (data shown in Suppl. Figure S5). Thus, it seems that DMD/PACSAB simulations are able to provide a reasonable description of the complex conformational landscape of the long α-synuclein protein. Focusing in more local structural characteristics, we have not found any persistent secondary structure along the protein chain, in agreement with the experimental observations obtained from NMR measurements[39].

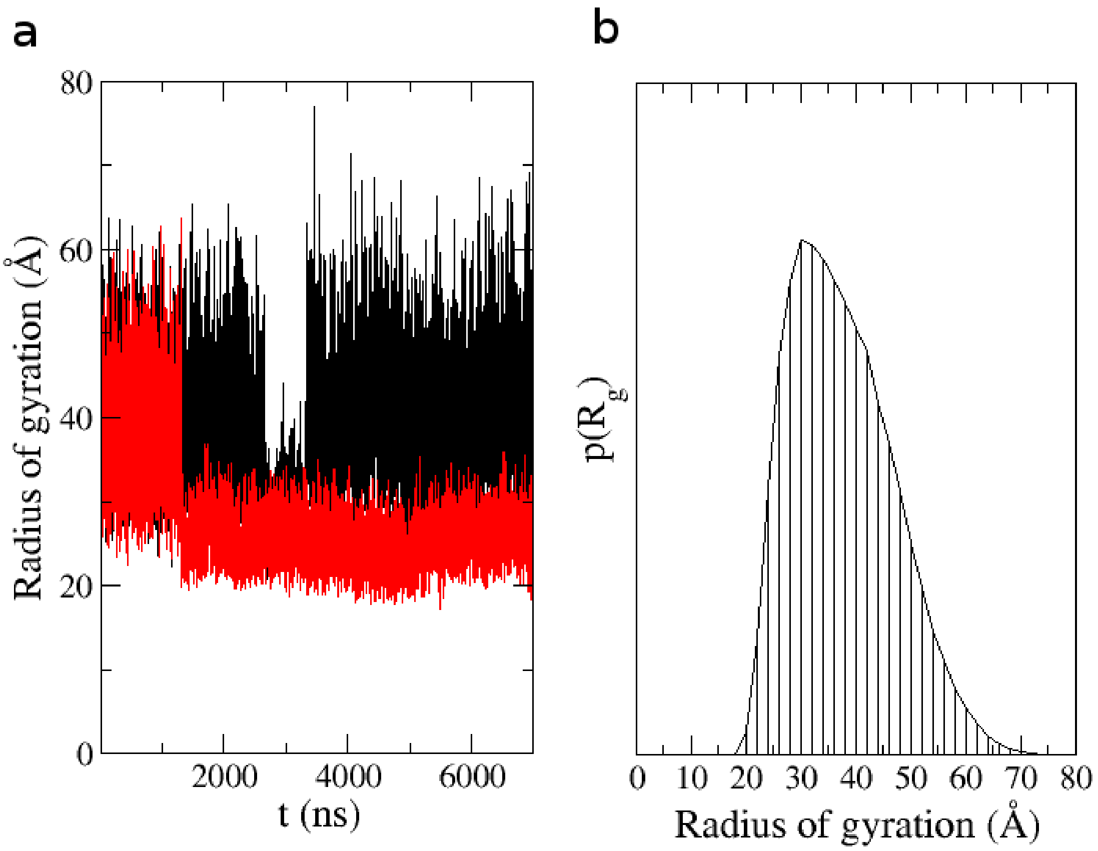

Figure 4. Simulations of α-synuclein. (a) Radius of gyration of two different trajectories (trajectories 4 (black line) and 8 (red line) of figure S2). (b) Histogram of the radius of gyration for the structural ensembles obtained from the 8 trajectories. All the simulations started from extended coil conformations.

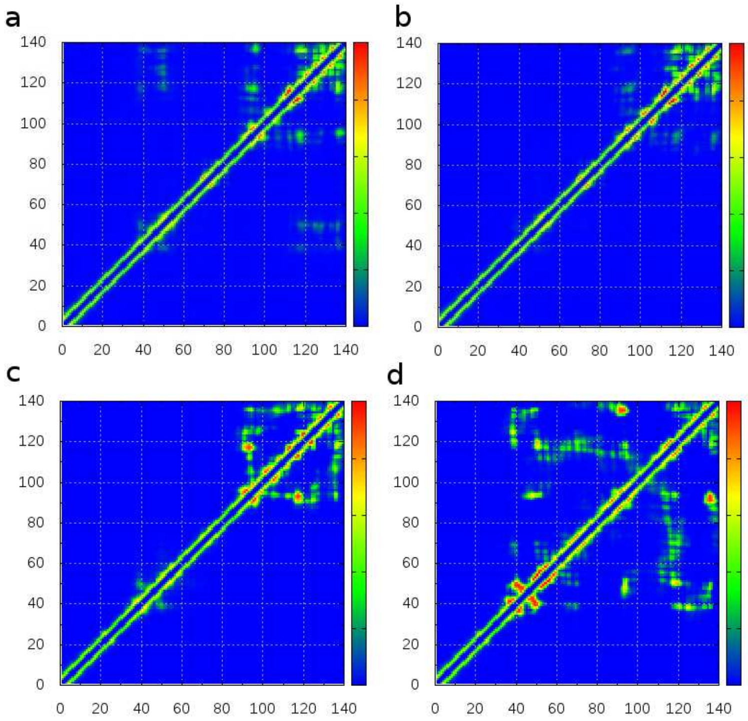

Figure 5. Contact maps in the two trajectories shown in the previous figure. (a) contact maps from trajectory 4 at 3000 ns (b) contact maps from trajectory 4 at 6000 ns. (c) contact maps from trajectory 8 at 1000 ns (d) contact maps from trajectory 8 at 4000 ns. The color scale (arbitrary units), goes from blue (no contacts) to red (many contacts).

The **RS peptide** (24 residues; sequence GAMGPSYGRSRSRSRSRSRSRSRS) is another challenging system, which has been experimentally characterized as a disordered peptide with R_g = 12.5 Å. This peptide was thoroughly studied with explicit solvent atomistic simulations[42], whose results were strongly dependent on the force-field and water model used. Like in the case of α-synuclein, standard water models give a radius of gyration lower than the experimental value, while simulations using the aforementioned new water model TIP4P-D give more extended structural ensembles. For this small peptide we were able to sample the full conformational space with a single 10 μs DMD/PACSAB trajectory, finding a small percentage of secondary structure (3.2% α-helix and 1.2% β-strand) due to the formation of short-lived secondary structure elements along the trajectory (see figure 6), and R_g = 12.3(1) Å, virtually identical to the experimental estimate.

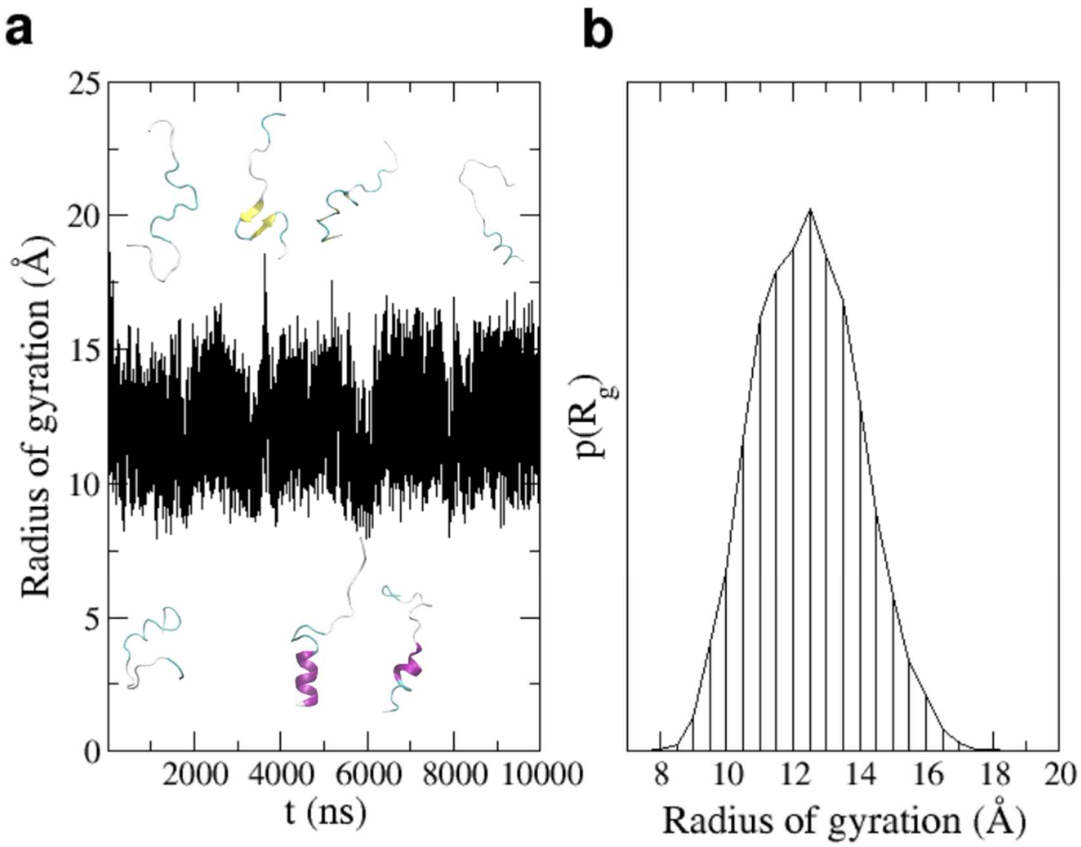

Figure 6. Simulation of the RS peptide. (a) Evolution of the radius of gyration along the trajectory (several snapshots along the trajectory shown). (b) Probability distribution of the radius of gyration.

We also tested our force field in two other unfolded proteins that have been studied with explicit solvent atomistic MD simulations: **IN** and **CspTM**. The N-terminal disordered domain of HIV-1 integrase (**IN**), a zinc-binding protein that is natively unfolded in the absence of zinc, has an experimental R_g around 24 Å[43], while explicit solvent atomistic MD simulations of this protein[22] with conventional water models provided R_g around 12 Å, the protein appearing less collapsed when the TIP4P-D water model was used in combination with last generation Amber and CHARMM force-fields, resulting in R_g around 20 Å. In our 10 μs simulations starting from a random extended conformation, we find an average R_g = 20(1) Å, quite close to the experimental value. We found the same good performance when applying our method to the study of unfolded cold-shock protein from *Thermotoga maritima* (**CspTm**), with an experimental R_g =

16(1) Å[43]. Also for this protein, only explicit solvent simulations with the TIP4P-D water model provide conformational ensembles where the protein displays the correct size[22]. In our DMD/PACSAB simulations starting from an extended random conformation, we find an average R_g=15 Å, almost coincident with the experimental estimate.

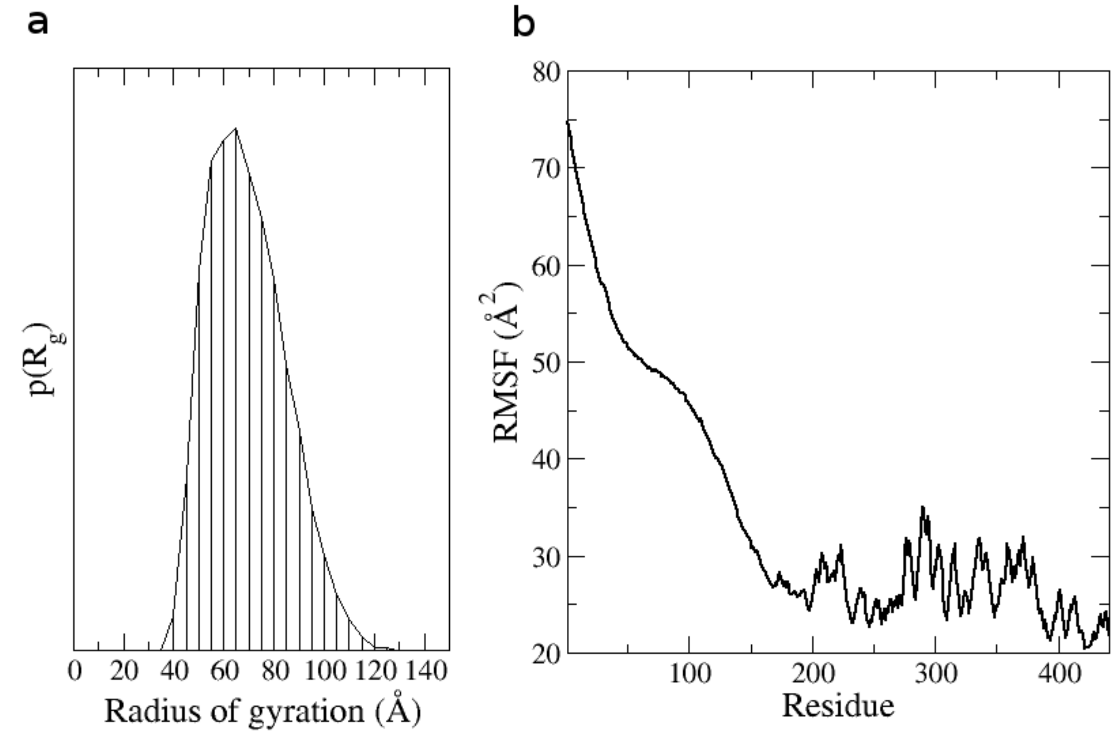

Figure 7. Simulation of hTau40. (a) Probability distribution of the radius of gyration. (b) Average RMS fluctuations from the trajectory.

Finally, we have applied our model to generate the conformational ensemble of a very large IDP, impossible to tackle with standard explicit solvent MD. We have chosen the 441-residues long Tau protein (**hTau40**), for which experimental information of its radius of gyration is available[44]. Tau is a highly disordered protein that binds to and stabilizes microtubules in nerve axons. In Alzheimer's disease Tau loses its ability to bind the microtubules and aggregates forming intracellular neurofibrillary tangles[45], but its theoretical study has been hampered by its large size, which precludes atomistic MD simulations with explicit solvent. We show in figure 7 the distribution of radius of gyration obtained from our 7 μs simulation, where we find R_g = 65(3) Å, in very good agreement with the value obtained from SAXS[44]. The mean inter-residue distances as a function of the sequence distance are in very good agreement with the valued found in experiments[41] (see Suppl. Figure S5). Remarkably, despite the simplicity of our model we find a very good agreement with experimental information about local characteristics of the protein. The protein shows a very high mobility in its N-terminal half, while the rest of the protein is less flexible (see RMS fluctuations per residue in figure 7). This distribution of mobility along the protein sequence is consistent with the estimation of the residue mobility from observed spin relaxation rates[45]. NMR measurements found a propensity to form α-helical structure around residue 120 and in the C-terminal. The prediction of the secondary structure propensity of each residue in our simulations is a very challenging test, since the PACSAB force field was calibrated essentially with just three parameters[9] (the strengths of the Van der Walls, the implicit solvation and the hydrogen bonding terms) to fit the association/dissociation probabilities of proteins, that depend on the average characteristics of the proteins rather than local sequence details. Very encouragingly, we found a region prone to α-helix structure around residue 120 (see Suppl. Figure S6), in good agreement with the NMR measurements[39]

PACSAB reproduces well proteins with dual folded/IDP nature: To test the performance of our force field to reproduce unfolded segments in generally folded proteins, we simulated pyridoxine 5'-phosphate oxidase, an enzyme whose structure is stable and known when bound to pyridoxal 5'-phosphate (PDB id 1G76), but when unbound, a region of 56 residues in the middle of the sequence becomes disordered, and does not give defined density maps (PDB id 1WV4). Our simulation, started from the fully folded 1G76 structure, reproduces correctly the disorder of this region, while keeping perfectly folded the rest of the protein (see figure 8).

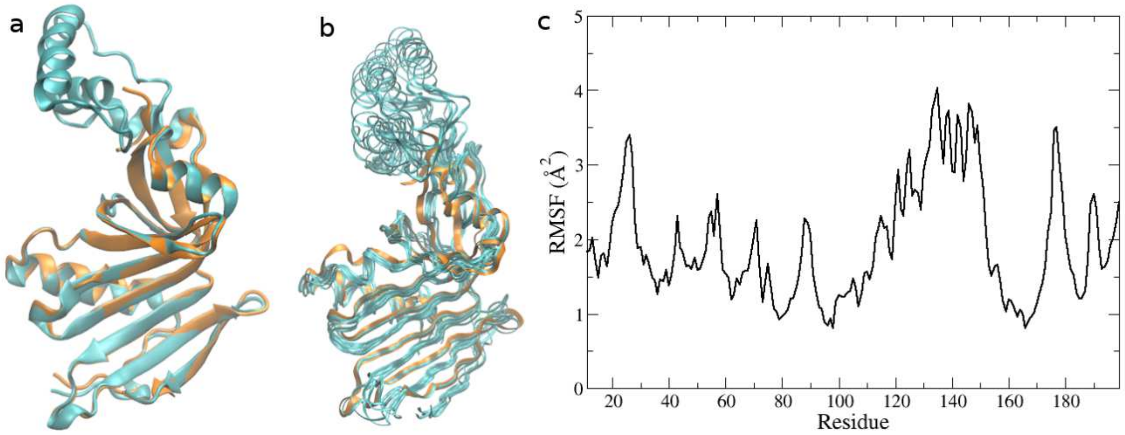

Figure 8. Conformational ensemble produced by our simulation of pyridoxine 5'-phosphate oxidase. (a) Crystal structure of the protein when bound to its ligand (cyan, PDB id 1G76) and of the unbound protein (orange, PDB id 1WV4), where the sequence region 111-157 is missing due to disorder. For the sake of clarity we have removed the N-terminal tail, which is different in the two PDB structures. (b) Several snapshots of the conformational ensemble of the simulation of 1G76, superimposed to the crystal structure 1WV4. (c) Average RMS fluctuations from the trajectory. The disordered region can be identified from its high RMSF, as also several loops and turns do.

CONCLUSIONS

We present here a refined version of our DMD/PACSAB coarse-grained force-field to be used to explore the structure and dynamics of both folded and unfolded proteins. Our refined DMD/PACSAB force field uses an effective non-bonded potential, constructed by interpolation between two parametrizations: one for the interaction between particles in close vicinity, and another one for distant particles (typically those that belong to different molecules or that are distant in sequence in an unfolded protein). This strategy improves the balance between association/dissociation rates and allows the accurate representation of both folded and unfolded proteins, while reproducing properly the reversibility of protein binding and protein dimerization, the first step of the aggregation process. Very interestingly, our simple implicit solvent model reproduces the correct thermodynamics of the system, while kinetics is largely accelerated due to the absence of solvent molecules. This enables us to make a faster conformational sampling of unfolded proteins, and explore efficiently the conformational space of large IDPs. The good performance of our model opens the prospect of generating good predictions of the conformational ensembles of large IDPs, impossible to study with standard explicit solvent simulations.

ASSOCIATED CONTENT

Supporting Information: Table showing the RMSD of the proteins in the benchmark, ordered by sequence length; Radius of gyration of ACTR using different parametrizations of the force field; RMSD obtained in the simulations of proteins in the PRIMO and OPEP benchmarks; illustration of a reversible binding event; RMSD obtained in the eight simulations of α-synuclein; mean value of the inter-residue distance as a function of the sequence separation for α-synuclein and tau; percentage of helical structure along sequence for the tau protein. The Supporting Information is available free of charge on the ACS Publications website

AUTHOR INFORMATION

Corresponding Authors

*E-mail: agusti.emperador@irbbarcelona.org (A.E.).

*E-mail: modesto.orozco@irbbarcelona.org (M.O.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work has been supported by the Spanish Ministry of Science (BFU2014-52864-R), the Catalan SGR, the Instituto Nacional de Bioinformática, the European Union's Horizon 2020 BioExcel and Elixir-Excellerate projects, the Biomolecular and Bioinformatics Resources Platform (ISCIII PT 13/0001/0030) co-funded by the Fondo Europeo de Desarrollo Regional (FEDER), and the MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona). M. O. is an ICREA (Institució Catalana de Recerca i Estudis Avançats) academia researcher.

REFERENCES

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

[2] Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5915.

[3] Chung, H. S.; Piana, S.; Shaw, D. E.; Eaton, W. E. Structural origin of slow diffusion in protein folding. *Science* **2015**, *349*, 1504-1510.

[4] Morriss-Andrews, A.; Shea, J. E. Computational studies of protein aggregation: methods and applications. *Annu. Rev. Phys. Chem.* **2015**, *66*, 643−666.

[5] Saunders, M. G.; Voth, G. A. Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **2013**, *42*, 73−93.

[6] Ingolfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The power of coarse graining in biomolecular simulations. *WIREs Comput. Mol. Sci.* **2014**, *4*, 225–248.

[7] Anandakrishnan, R.; Drozdetski, A.; Walker, R. C.; Onufriev, A.V. Speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations. *Biophys. J.* **2015**, *108*, 1153−1164.

[8] Alder, B.J.; Wainwright, T.E. Studies in Molecular Dynamics. I. General method. *J. Chem. Phys.* **1959**, *31*, 459-466.

[9] Emperador, A.; Sfriso, P.; Villarreal, M. A.; Gelpi, J. L.; Orozco, M. PACSAB: Coarse-grained force field for the study of protein−proteiniInteractions and conformational sampling in multiprotein systems. *J. Chem. Theory Comput.* **2015,** *11*, 5929-5938 (2015).

[10] Abriata, L. A.; Dal Peraro, M. Assessing the potential of atomistic molecular dynamics simulations to probe reversible protein-protein recognition and binding. *Sci. Rep.* **2015**, *5*, 10549.

[11] Ding, F.; Tsao, D.; Nie, H.; Dokholyan, N. V. Ab initio folding of proteins with all-atom discrete molecular dynamics. *Structure* **2008**, *16*, 1010−1018.

[12] Smith, W.S.; Hall C. K.; Freeman B. D. Molecular dynamics for polymeric fluids using discontinuous potentials. *J. Comput. Phys.* **1997**, 134, 16–30.

[13] Emperador, A.; Meyer, T.; Orozco, M. Protein flexibility from discrete molecular dynamics simulations using quasi-physical potentials. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 83−94.

[14] Emperador, A.; Meyer, T.; Orozco, M. United-atom discrete molecular dynamics of

proteins using physics-based potentials. *J. Chem. Theory Comput.* **2008**, *4*, 2001−2010.

[15] Sfriso, P.; Emperador, A.; Orellana, L.; Hospital, A.; Gelpi, J.-L.; Orozco, M. Finding conformational transition pathways from discrete molecular dynamics simulations. *J. Chem. Theory Comput.* **2012**, *8*, 4707-4718.

[16] Urbanc, B.: Cruz, L.; Yun, S.; Buldyrev, S. V.; Bitan, G.; Teplow, D. B.; Stanley, H. E. In silico study of amyloid beta-protein folding and oligomerization. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 17345−17350.

[17] Nguyen, H. D.; Hall, C. K. Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 16180–16185.

[18] Nedumpully-Govindan, P.; Kakinen, A.; Pilkington, E. H.; Davis, T. P.; Ke, P. Ch.; Ding, F. Stabilizing off-pathway oligomers by polyphenol nanoassemblies for IAPP aggregation inhibition. *Sci. Rep.* **2016**, *6*, 19463.

[19] Emperador, A.; Solernou, A.; Sfriso, P.; Pons, C.: Gelpi, J.-L.; Fernandez-Recio, J.; Orozco, M. Efficient relaxation of protein-protein interfaces by discrete molecular dynamics. *J. Chem. Theory Comput.* **2013**, *9*, 1222−1229.

[20] Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.

[21] Lazaridis, T.; Karplus, M. Effective energy function for proteins in solution. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 133−152.

[22] Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B* **2015**, 119, 5113-5123.

[23] Best, R. B.; Zheng, W.; Mittal, J. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **2014,** *10*, 5113-5124.

[24] Petrov, D.; Zagrovic, B. Are current atomistic force fields accurate enough to study proteins in crowded environments? *PLoS Comput. Biol.* **2014,** 5, e1003638.

[25] Camps, J.; Carrillo, O.; Emperador, A.; Orellana, L.; Hospital, A.; Rueda, M.; Cicin-Sain, D.; D'Abramo, M.; Gelpi, J.-L., Orozco, M. FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics* **2009**, *25*, 1709-1710.

[26] Hospital, A.; Andrio, P.; Fenollosa, C.; Cicin-Sain, D.; Orozco, M.; Gelpi, J.-L. MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. *Bioinformatics* **2012**, *28*, 1278-1279.

[27] Meyer, T.; D'Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Perez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; Gelpi, J.-L.; Orozco, M. MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure* **2010**, *18*, 1399−1409.

[28] Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Perez, A.; Camps, J.; Hospital, A.; Gelpi, J.-L.; Orozco, M. A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 796−801.

[29] Juneja, A.; Ito, M.; Nilsson, L. Implicit solvent models and stabilizing effects of mutations and ligands on the unfolding of the amyloid β-peptide central helix. *J. Chem. Theory Comput.* **2012**, *9*, 834-846.

[30] Kar, P.; Gopal, S. M.; Cheng, Y. M.; Predeus, A.; Feig, M. PRIMO: A transferable coarse-grained force field for proteins. *J. Chem. Theory Comput.* **2013**, *9*, 3769−3788.

[31] Chebaro, Y.; Pasquali, S.; Derreumaux, P. The coarse-grained OPEP force field for non-amyloid and amyloid proteins. *J. Phys. Chem. B* **2012**, *116*, 8741−8752.

[32] Harada, R.; Tochio, N.; Kigawa, T.; Sugita, S.; Feig, M. Reduced native state stability in crowded cellular environment due to protein−protein interactions, *J. Am. Chem. Soc.* **2013,** *135*, 3696-3701.

[33] Pujol-Pina, R.; Vilaprinyo-Pasqual, S.; Mazzucato, R.; Arcella, A.; Vilaseca, M.; Orozco, M.: Carulla, N. SDS-PAGE analysis of Aβ oligomers is disserving research into Alzheimer´s disease: appealing for ESI-IM-MS. *Sci. Rep.* **2015,** *5*, 14809 (2015).

[34] Kjaergaard, M.; Norholm, A. B.; Hendrus-Altenburger, R.; Pedersen, S. F.; Poulsen, F. M.; Kragelund, B. B. Temperature-dependent structural changes in intrinsically disordered proteins: Formation of a-helices or loss of polyproline II? *Protein Sci.* **19**, 1555-1564.

[35] Ebert, M.-O.; Bae, S.-H.; Dyson, H. J.; Wright, P. E. NMR relaxation study of the complex formed between CBP and the activation domain of the nuclear hormone receptor coactivator ACTR. *Biochemistry* **2008**, *47*, 1299-1308.

[36] Van Rooijen, B. D.; van Leijenhorst-Groener, K. A.; Claessens, M. M. A. E.; Subramaniam, V. Tryptophan fluorescence reveals structural features of α-Synuclein oligomers. *J. Mol. Biol.* **2009**, *394*, 826-833.

[37] Esteban-Martin, S.; Silvestre-Ryan, J.; Bertoncini, C. W.; Salvatella, X. Identification of fibril-like tertiary contacts in soluble monomeric α-synuclein. *Biophys. J.* **2013**, *105*, 1192−1198.

[38] Morar, A. S.; Olteanu, A.; Young, G. B.; Pielak, G. J. Solvent-induced collapse of

α-Synuclein and acid-denatured cytochrome c. *Protein Sci.* **2001**, *10*, 2195-2199.

[39] Schwalbe, M.; Ozenne, V.; Bilbow, S.; Jaremko, M.; Jaremko, L.; Gajda, M.; Jensen, M. R.; Biernat, J.; Becker, S.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α-Synuclein in solution from NMR and small angle scattering. *Structure* **2014**, *22*, 238-249.

[40] Allison, J. R.; Varnai, P.; Dobson, C. M.; Vendruscolo, M. Determination of the free energy landscape of α-synuclein using spin label nuclear magnetic resonance measurements. *J. Am. Chem. Soc.* **2009**, *131*, 18314-18326.

[41] Nath, A.; Sammalkorpi, M.; DeWitt, D. C.; Trexler, A. J.; Elbaum-Garfinkle, S.; O'Hern, C. S.; Rhoades, E. The conformational ensembles of a-synuclein and tau: combining single-molecule FRET and simulations. *Biophys. J.* **2012**, *103*, 1940-1949.

[42] Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmueller, H. Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. *J. Chem. Theory Comput.* **2015,** *11*, 5513-5524.

[43] Wuttke, R.; Hofmann, H.; Nettels, D.; Borgia, M. B.; Mittal, J.; Best, R. B.; Schuler., B. Temperature-dependent solvation modulates the dimensions of disordered proteins. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 5213-5218.

[44] Mylonas E.; Hascher, A.; Bernado, P.; Blackledge, M.; Mandelkow, E.; Svergun, D. I. Domain conformation of tau protein studied by solution small-angle X-ray scattering. *Biochemistry* **2008**, *47*, 19345-10353.

[45] Mukrasch M. D.; Bibow, S.; Korukottu, J.; Jeganathan, S.; Biernat, J.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. Structural polymorphism of 441-residue tau at single residue resolution. *PLoS Biology* **2009**, *7*, e1000034.