Chapter

STRUCTURAL ANALYSIS OF INTRINSICALLY DISORDERED PROTEINS: COMPUTER ATOMISTIC SIMULATION

Anna Battisti[1]***, Gabriele Ciasca***[2] ***and Alexander Tenenbaum***[3,*]
[1]International School for Advanced Studies (SISSA), Trieste, Italy
[2]Physics Institute, Catholic University, Roma, Italy
[3]Physics Department, Sapienza University, Roma, Italy

Abstract

Intrinsically disordered proteins (IDPs) are biomolecules that do not have a definite 3D structure; their role in the biochemical network of a cell relates to their ability to switch rapidly among different secondary and tertiary structures. For this reason, applying a simulation computer program to their structural study turns out to be problematic, as their dynamical simulation cannot start from a known list of atomistic positions, as is the case for globular proteins that do crystallize and that one can analyze by X-ray spectroscopy to determine their structure.

We have established a method to perform a computer simulation of these proteins, apt to gather statistically significant data on their transient structures. The only required input to start the procedure is the primary sequence of the disordered domains of the protein, and the 3D structure of the ordered domains, if any. For a fully disordered protein the method is as follows.

(a) The first step is the creation of a multi-rod-like configuration of the molecule, derived from its primary sequence. This structure evolves dynamically *in vacuo* or in an implicit model of solvent, until its gyration radius - or any other measure of the overall configuration of the molecule - reaches the experimental average value; at this point, one may follow two different paths.

(b1) If the study focuses on transient secondary structures of the molecule, one puts the structure obtained at the end of the first step in a box containing solvent molecules in explicit implementation, and a standard molecular dynamics simulation follows.

(b2) If the study focuses on the tertiary structure of the molecule, a larger sampling of the phase space is required, with the molecule moving in very large and diverse regions of the phase space. To this end, the structure of the IDP is let evolve dynamically

[*] E-mail address: alexander.tenenbaum@roma1.infn.it

in an implicit solvent using metadynamics, an algorithm that keeps track of the regions of the phase space already sampled, and forces the system to wander in further regions of the phase space.

(c) One can increase the accuracy of the statistical information gathered in both cases by fitting, where available, experimental data of the protein. In this step one extracts an ensemble of ’best’ conformers from the pool of all configurations produced in the simulated dynamics. One derives this ensemble by means of an ensemble optimization method, implementing a genetic algorithm.

We have applied this procedure to the simulation of tau, one of the largest fully disordered proteins, which is involved in the development of Alzheimer’s disease and of other neurodegenerative diseases. We have combined the results of our simulation with small-angle X-ray scattering experimental data to extract from the dynamics an optimized ensemble of most probable conformers of tau.

The method can be easily adapted to IDPs entailing ordered domains.

PACS: 87.14.E-, 87.15.ap, 87.15.bd, 87.15.bg

Keywords: Intrinsically Disordered Proteins; tau protein; transient secondary structures; transient tertiary structures; molecular dynamics; metadynamics

1. Introduction

Intrinsically disordered proteins (IDPs) do not have an average stable structure in their native state; they are similar to a random coil fluctuating in an ensemble of conformations, and resemble highly denatured proteins [1, 2]. Due to their flexibility, and at a variance with the well-known lock-and-key biomolecular paradigm, they perform tasks that globular proteins cannot perform [1, 3, 4, 5, 6]. IDPs entail at least one extended disordered region, and can entail globular domains alternating with flexible linkers or disordered domains. These proteins are therefore characterized by different degrees of disorder, from those formed by globular domains connected by disordered segments to those totally disordered [1, 2]. Even the latter may entail segments endowed, albeit temporarily, with secondary structures such as α-helices, β-sheets or PPII helices [7].

The main functions of IDPs are not structural, but regulatory: control, modulation and signalling. A characteristic feature of their biological function is an interaction energy among residues that is significantly lower than for globular proteins [1, 6]; this favors fast shifts between extended conformations, which generally accompany the binding to other molecules, and disordered molten globule-like conformations. The formation and dissolution of bound states is probably faster than in the case of globular proteins [1, 4].

The biochemical functions of IDPs relate to their secondary and tertiary structures, which vary in time. Given the speed at which transitions between different conformations are supposed to take place, a computer simulation of their dynamics seems to be a promising tool to characterize their time-dependent structure and to understand their behavior. The dynamical simulation of an IDP is a computational challenge, because by definition there are no experimentally determined 3D structures of the whole molecule, such as a Protein Data Bank file, from which to start. We have developed a method to perform computer simulations of IDPs, apt to overcome this obstacle and to gather statistically significant data on the transient structures of the proteins.

The simulation of an IDP confronts a second problem, namely the choice of a suitable force field. Molecular mechanics force fields have been parametrized on folded protein structures, and therefore may not correctly reproduce the structure of disordered proteins. On the other hand, there is no alternative to the use of one of the known force fields, because an *ab initio* calculation of a large disordered molecule would be unfeasible. Our simulations indicate the known force fields are apt to represent - with some caution - also the dynamics of IDPs.

The only required input to start the procedure is the primary sequence of the protein. We will illustrate the method by applying it to protein tau, a large, fully disordered protein[1].

2. The Tau Protein

The tau protein, one of the largest fully disordered IDPs [3], is involved in the nucleation and stabilization of the microtubules (MTs) in the cytoskeleton of the axons of the neurons. Protein tau achieves stabilization through the bonding of its repeats domain to the α- and β-tubulines forming the MTs [8]. But the same tau can aggregate in paired helical filaments (PHFs) and form fibrils which, in their turn, form insoluble tangles [1, 8]; this pathological deviation from its physiological function together with other factors triggers the development of Alzheimer's disease and of other neurodegenerative diseases [8].

Tau exists in several isoforms[1]; we have chosen to simulate its htau40 isoform, which is located in the human central nervous system; it has 441 residues and a molecular weight of 45.85 kDa. One can distinguish in its primary sequence four domains, corresponding to morphologically different sections of the molecule: the N-terminal projection domain (residues 1-150); a proline-rich segment (residues 151-243); a domain entailing four repeats (residues 244-368); the C-terminal domain (residues 369-441).

The process of formation of the PHFs is not entirely known, but some of its factors and stages have been investigated. A precursor stage of tau's polymerization has been related to specific transient global folds of the protein, in which the N-terminal is folded near the repeats domain in a hairpin conformation [7, 9, 10], or the C terminal is in proximity of the repeats domain and the N-terminal is folded near the C-terminal in a paperclip conformation [5, 11, 12]. The pathological aggregation in the form of insoluble tangles has been attributed to a local transition from the unfolded state to a β-structure [13, 14, 15, 16]. The aggregation process is supposed to start from a nucleus entailing the VQIVYK motif, a segment with high propensity for a β-structure [10], or the VQIINK motif. These two hexapeptides are, respectively, at the beginning of the third and of the second repeat [7, 9, 10], and have been identified as components of steric zippers formed by β-sheets parallel to the axis of the fibril [16]. The propensity of a polyproline-II motif toward the formation of β-sheets has also been listed among the possible causes of the aggregation of tau proteins, and hence of the origin of tau-pathologies [4, 17].

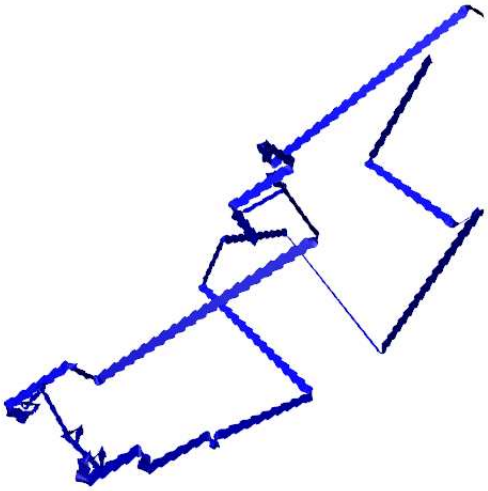

Figure 1. Initial shape of tau protein, as produced by the VMD program from the primary sequence [19].

3. Initial Configuration and Dynamical Evolution

In order to perform a molecular dynamics (MD) simulation of tau, we start from its primary sequence of amino acids. We use this sequence as an input to the visual MD (VMD) program [18]; this program lines up the amino acids following their primary sequence, departing from a straight line only when obliged by stereochemical incompatibility of neighboring amino acids. The output of VMD is thus a 3D sequence of straight segments of amino acids that bears little resemblance to a real protein, as shown in Fig.1. (There are other programs similar to VMD).

We use the MD simulation program GROMACS[2] to evolve this multi-rod-like structure *in vacuo* at $T = 300$ K [19]. The molecule's configuration collapses in a short time. This can be monitored by measuring the gyration radius $R_g = (\sum_i r_i^2 m_i / \sum_i m_i)^{1/2}$, where $\boldsymbol{r}_i$ are the positions of the atoms with respect to the center of mass of the molecule, and m_i are their masses; R_g measures the average size of the overall conformations of a molecule. Curve #1 in Fig.2 shows that the collapse of R_g, from an initial value of 10.4 nm to a value of 2.5 nm, takes place in about 100 ps and yields a very compact and entangled configuration. This fast evolution of the molecule is due to the absence of the solvent, which would prevent the collapse and the formation of a high number of intramolecular H-bonds as can be seen in Fig.3 (curve #1). We therefore stop the evolution when the configuration has reached a value of the gyration radius equal to the experimental average $R_g = 6.57$ nm [5]. The structure obtained in this way is then embedded in water, using an explicit or an implicit model.

The gyration radius is known for many molecules, being measured either by light scat-

[1] www.uniprot.org/uniprot/P10636; www.disprot.org/protein.php?id=DP00126

[2] GROMACS release 4.5.3, www.gromacs.org; box volume = 15253 nm3; ffamber99 force field; time step 2 fs; modified Berendsen thermostat, Parrinello-Rahman pressure coupling.

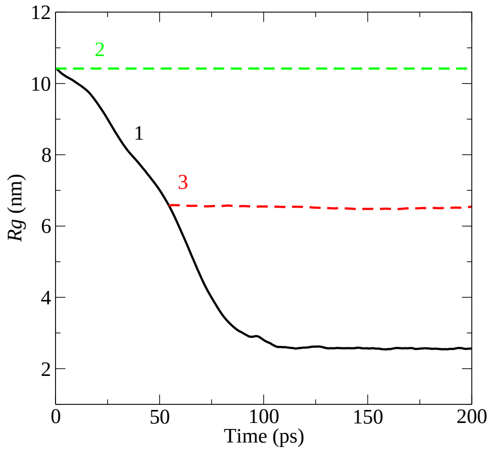

Figure 2. Evolution of the gyration radius at $T = 300$K in the first 200 ps, starting from the configuration of Fig.1. The continuous black line (1) shows the rapid collapse of the protein *in vacuo*. The dashed green line (2) shows the evolution after addition of water molecules to the initial configuration of tau. The dashed red line (3) shows the evolution after addition of water molecules to the stucture extracted at $t = 56$ ps and $R_g = 6.57$ nm [19].

tering, or by SAXS (small-angle X-ray scattering), or by small angle neutron scattering. One could instead use any other measure of the overall configuration of the molecule to monitor its evolution *in vacuo*, like asphericity, which combines shape and compactness, and is measured by fluorescence microscopy [20].

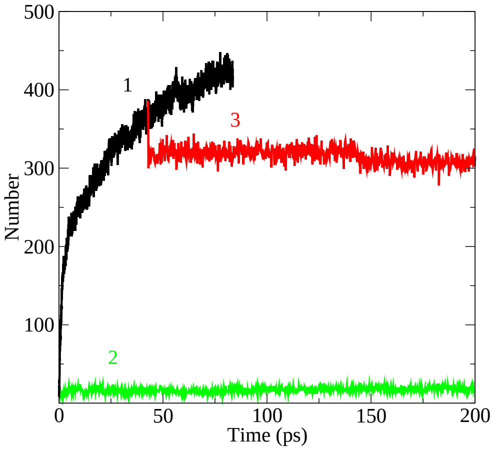

Figure 3. Time evolution of the number of intramolecular H-bonds at $T = 300$ K, first 200 ps. The black line (1) shows their rapid increase during the evolution *in vacuo*. The green line (2) shows their slow change after addition of water molecules to the initial configuration of tau. The red line (3) shows a sharp drop when water molecules are added to the structure extracted at t = 56 ps and the stable subsequent evolution [19].

An alternative way to reach in a short time a configuration of a large molecule from

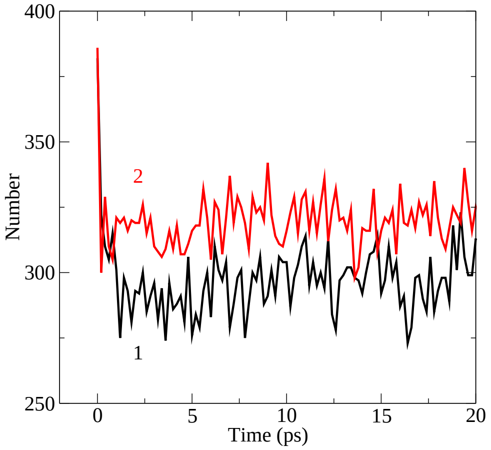

Figure 4. Time evolution of the number of intramolecular H-bonds at $T = 300$ K, beginning at the time when solvent is introduced in the simulation. Black line (1): explicit water molecules. Red line (2): implicit water solvent [19].

which to start the dynamical simulation is to embed the initial multi-rod-like molecule in an implicit solvent. When one immerses the initial VMD structure in implicit water [21], it rapidly evolves to a natural conformation, which can then be used to start a simulation. The time of this initial evolution is similar to the time of the evolution *in vacuo*; for tau, it takes about 50 ps.

A straightforward way of letting the initial multi-rod-like configuration evolve towards a native-like state of the molecule would be to embed its initial structure in explicit solvent. The configuration produced by VMD from the primary sequence is extended, and the simulation box has to be accordingly large, as a fully disordered protein (like tau) fluctuates in an ensemble of very different conformations. Therefore, in a MD simulation of such an IDP, one has to pay attention to the flexibility of the molecular structure: when periodic boundary conditions are used, the box must be large enough to avoid that during the dynamics the protein interacts with one of its periodic images, extending its shape to the region bordering the walls of the box. A large box entails a very large number of solvent molecules; this is a relevant obstacle to make such a simulation workable. Using the box-to-molecule size relation usually adopted in this kind of MD simulation, in the case of tau one would have to use a box filled with about $1.5 \cdot 10^6$ water molecules.

Moreover, the evolution of the molecule from the initial configuration is very slow, as shown by curves #2 in Fig.2 and Fig.3. Taking into account the experimental value for the average gyration radius $R_g = 6.57$ nm, one can foresee for the overall configuration of the molecule, and for R_g, an equilibration time in excess of several tens of nanoseconds. All in all, this would be a computationally very expensive procedure for a molecule as large as tau; but it could be feasible for a small molecule.

4. Secondary Structures

The initial evolution, either *in vacuo* or in implicit solvent, lasts few tens of ps; at the end one gets a structure of the molecule of a size comparable to the average experimental one. For large molecules (like tau) the size is significantly reduced with respect to the one produced by VMD; its structure can thus be put in a simulation box much smaller than the initial one, and the number of solvent molecules in the box is significantly reduced with respect to the number needed to embed the initial extended state produced via VMD.

At this stage the simulation run can begin, embedding the molecule's structure in a suitable solvent. Let us assume that the computer experiment aims at gathering information on the existence and probability of segments of the molecule endowed, albeit temporarily, with a secondary structure. As we show below, to achieve the best simulation of these structures the solvent must be represented by explicit molecules, in order to have a realistic competition between intramolecular and intermolecular (that is, between protein and water molecules) H-bonds.

Tau is a good candidate for this kind of simulation. Its value of R_g is large due to the molecule's non-globular structure, and is about that of a random coil of the same length (6.9 nm). Nevertheless, when R_g is measured in partial domains of tau that entail the repeats, its value turns out to be larger than the value estimated for a random coil; this hints at a propensity of these domains to form secondary structures [5, 11]. Because these structures would very likely be transient, there is a definite interest in a dynamical simulation of tau, in order to acquire a detailed knowledge of these transient patterns.

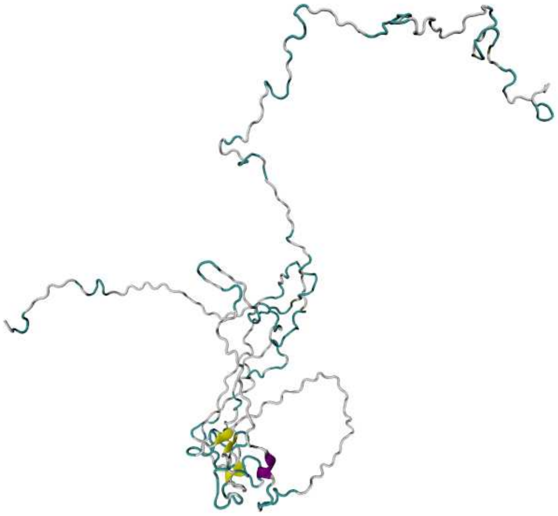

Figure 5. Shape of the tau protein after a 100 ps evolution, 56 ps *in vacuo* and 44 ps in explicit water, at $T = 300$ K. Two transient short β-sheets (yellow) and a transient short α-helix (purple) are highlighted [19].

After a short minimization of the total energy, the system (tau + water molecules [3]) is

[3] spce water model in our simulation.

ready to start a dynamical evolution in a region of the phase space corresponding to realistic conformations of the molecule. The stabilizing effect of the introduction of explicit water on the dynamical evolution of the protein implies a sudden interruption of the collapse of the molecule, and the beginning of a slow fluctuation of the structure; one can see this in Fig.2, curve #3. As for the intramolecular H-bonds, curve #3 of Fig.3 shows that the introduction of the water molecules causes a sudden decrease in the number of those bonds, about a fourth of which is replaced by H-bonds between tau and the water molecules. Fig.4 shows more in detail this instantaneous decrease (curve #1), and the first 20 ps of a NVT (constant number of molecules, volume, and temperature) simulation. After the introduction of the solvent, the molecule assumes in a short time a native-like configuration, as shown in Fig.5; the structure displays short transient secondary structures like β-sheets and α-helices.

One could follow a similar procedure by embedding in implicit water [21] the molecule's structure extracted midway during the collapse *in vacuo*, when the molecule has shrunk to a natural size, or leaving it in the implicit solvent, if one used this from the beginning. The use of an implicit solvent would allow a much faster simulation run, if compared to the implementation of an explicit water model. On the other hand, the two solvent types are known to operate differently in the prediction of secondary structures [22, 23]. Indeed, as shown in Fig.4, the effect of implicit water on the replacement of intramolecular H-bonds is not the same as that of explicit water molecules: the latter seem to be more efficient in competing with intramolecular H-bonds and replacing them with solvent-solute ones. After an evolution *in vacuo* of the molecule, possible spurious H-bonds, which would not form if a solvent had been included in the simulation from the beginning of the dynamics, are better removed by putting the molecule in explicit solvent.

We used the initial conformation of tau produced by VMD to start a simulation at constant temperature and pressure [24]. For this simulation we have chosen the ffG53a6 force field, implemented in the GROMACS package 3 [4]. The simulation has been carried out at neutral pH (pH = 7), close the physiological value (that is in the 7.2 - 7.4 range). Accordingly, amino acids were set to their default protonation states at pH = 7, with Lys, Arg carrying a +1 and Glu, Asp a -1 net charge.

In this first dynamical simulation of the complete tau (htau40), we have studied the time evolution of the molecule in water over a time of 30 ns. Fig.6 shows the gyration radius; R_g is not stable around its experimental value, as it progressively decreases to about 4.3 nm. Even though the latter value is within the range of values computed from a set of static conformers of tau produced by the EOM method [5, 14], the continuous decrease of R_g hints at a possible shortcoming of the force field in reproducing the overall shape of the molecule. In order to clarify this dynamical behavior we have computed the time evolution of the gyration radius of the four domains corresponding to morphologically different sections of the molecule: the N-terminal projection domain, the proline-rich segment, the repeats domain, and the C-terminal domain. We report the results in Fig.7; they show that all four domains reach an equilibrium stage: first the C-terminal domain, after about 10 ns; second the repeats domain, shortly before 20 ns; then the proline-rich segment and the

[4] GROMACS release 4.5.3, www.gromacs.org; box volume = 15253 nm3; ffG53a6 force field; spce water model; time step 2 fs; modified Berendsen thermostat, Parrinello-Rahman pressure coupling.

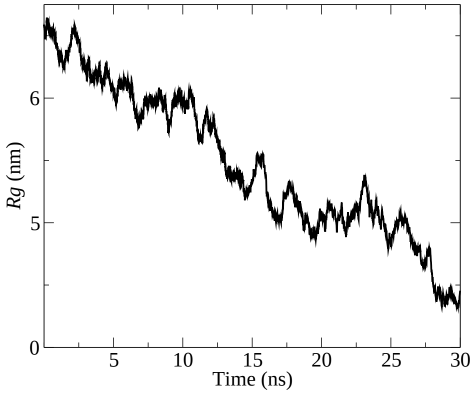

Figure 6. Time evolution of the gyration radius of protein tau during the standard MD dynamics at $T = 300$ K [24]. The experimental average value of R_g is 6.6 nm and the standard deviation is 0.3 nm [13].

N-terminal domain, after 22 ns. The final decrease of the total R_g visible in Fig.6 after an apparent stabilization between 18 and 24 ns has thus tobe attributed to a reduction of the distances among domains, rather than to a further shrinking of one or more of them.

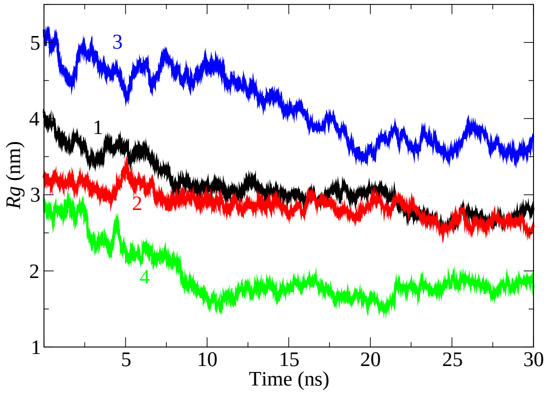

Figure 7. Time evolution of the gyration radius of four domains of tau at $T = 300$ K, in a standard MD simulation. Curve #1 (black): residues 1-150, N-terminal domain; curve #2 (red): residues 151-243, proline-rich segment; curve #3 (blue): residues 244-368, repeats domain; curve #4 (green): residues 369-441, C-terminal domain [24].

One can notice that the stabilized value of R_g found for the repeats domain in our simulation (3.7 nm) almost coincides with the experimental value of 3.8 nm found for the K18 construct of tau, the latter almost coinciding with the repeats domain [5].

In order to assess the propensities of various domains of tau to form temporary secondary structures, we have extracted information on the formation of temporary secondary structures from the whole 30 ns dynamics. We have measured the time evolution of the

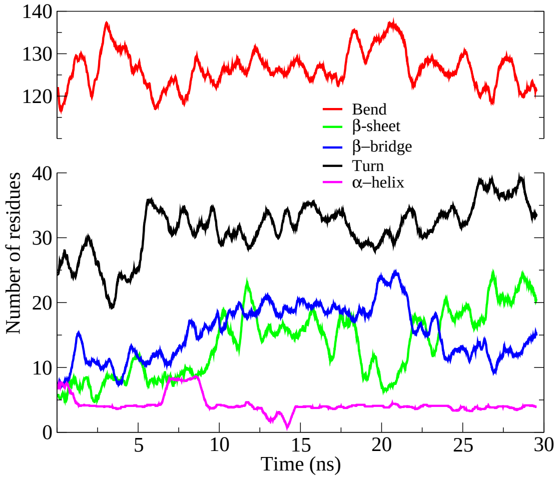

Figure 8. Number of residues found in secondary structures during the 30 ns dynamics. The curves have been smoothed by averaging the data over a sliding 1 ns interval [24].

number of residues found in coils, β-sheets, β-bridges, bends, turns, and α-helices; we show these quantities are shown in Fig.8. While the majority of residues are in a coil-like conformation and in bends, there is a significant presence of secondary structures like turns, β-bridges, β-sheets, and α-helices.

The number of residues forming bends oscillates in a stable way during the dynamics; the number of residues forming turns stabilizes after about 6 ns; the number of residues forming a helix oscillates during the whole dynamics (mostly an α-helix, with some short shifts to a 3-helix or a 5-helix). Fig.8 shows that the formation of temporary secondary structures does not depend significantly on the overall shape of the molecule, due to their localized nature. The pattern of extension and time dependence of temporary secondary structures in tau shown in Fig.8 should thus be representative of the equilibrium state.

We have compared the average values of the extension and frequency of temporary α- and β-structures measured in this simulation with propensities to form α-helices or β-structures, assigned to various segments of tau using experimental NMR data [7]. Weighing the number of residues entailed in each of these segments with its propensity (fraction of time spent in the secondary structure), one finds an average number of 12 residues in β-structures and of 4 residues in an α-helix. These results are compatible with the results of our MD simulation, namely 26 $\pm$ 14 residues and 5 $\pm$ 2 residues for β-structures and α-helices, respectively [24].

5. Tertiary Structures

As mentioned before, particular tertiary structures, albeit transient, are supposed to play a key role in the pathological evolution of protein tau. These structures are likely to evolve over regions of the phase space that are much larger than those characteristic of temporary secondary structures. It is therefore important, in order to gather information on the existence and probability of configurations endowed with statistically significant tertiary

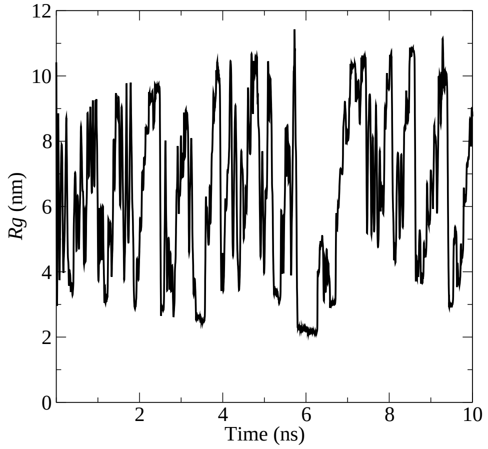

Figure 9. Time evolution of the gyration radius of protein tau during the metadynamics at $T = 300$ K [27].

structures, to achieve a dynamical simulation able to sample the overall structure of the molecule over large regions of the phase space.

The previous procedure, that has been proved to be effective in sampling local temporary secondary structures, would not be as effective in sampling larger regions of the space, as required to detect statistically relevant, albeit transient, tertiary structures. Metadynamics is a simulation tool that greatly expands the explored region of the phase space of an IDP, allowing a significant sampling of its fluctuating tertiary structure.

The metadynamics algorithm keeps track of the regions of the phase space already sampled by the molecule in its dynamics, recording a collective dynamical variable, and forces the system to leave those regions and to wander in other regions of the phase space [25]. The system thus samples a portion of the phase space much larger than the one sampled in an equal time of standard MD computation. Because of the heavier computation implied by keeping the dynamical track, the simulation of a large system, like protein tau, dictates the use of an implicit solvent model [26].

We have performed 10 ns of metadynamics simulation of protein tau [27]. The ffG53a6 force field we used in the previous molecular dynamics simulation of tau provides a statistical measure of local transient secondary structures; but the force field is less effective in maintaining over long times an extended conformation of the molecule. An advantage of the use of metadynamics is that it overcomes this problem, as its algorithm avoids the risk of a shrinking of the molecule by forcing its structure up and down the R_g range, as shown in Fig.9. We used in this simulation the ffamber99 force field, because the implicit solvent model could not be implemented with the ffG53a6 force field.

To implement this method we have chosen the collective dynamical variable R_g, the gyration radius. We have fixed the parameters of the metadynamics algorithm in such a way that R_g oscillates around its experimental value [5]. We have monitored the time evolution of tau by computing R_g; Fig.9 shows the gyration radius during a 10 ns evolution. The

[5] CV = R_g; deposition stride $\tau = 10$ ps; height $W = 0.5$ kJ/mol; Gaussian width $\sigma = 0.35$ nm; limits on R_g: upper UWALL = 7.0 nm, lower LWALL = 5.5 nm.

metadynamics algorithm induces large structural changes in the molecule, with the gyration radius spanning the range between 2.5 nm and 11 nm, centered near the experimental value: its average value over this evolution is $R_g = 6.3$ nm. The large oscillations of R_g hint at the variety of configurations sampled by the system in different regions of the phase space.

One can better understand the way in which the metadynamics algorithm acts on the system by separately computing the time evolution of the gyration radius of the N-terminal domain, of the proline-rich segment, of the repeats domain, and of the C-terminal domain. We report the results in Fig.10, which shows that all four domains undergo very strong modifications. The algorithm shakes them alternatively, leaving from time to time one of the domains in what appears to be a local equilibrium well. As an example, the proline-rich domain is almost stable during the first 2 ns, while the other domains show strong changes in their overall configuration; the first domain reaches again a relative stability between 3.5 and 6.3 ns, and between 9.0 and 9.6 ns. The other domains also stay for some time in a quasi-equilibrium state: the C-terminal between 5.2 and 6.3 ns, the N-terminal between 5.8 and 6.8 ns, and the repeats domain between 5.2 and 6.4 ns. The large amplitude of the oscillations of the N-terminal are due to its higher flexibility in comparison to the other domains, in particular the domain entailing the repeats [7].

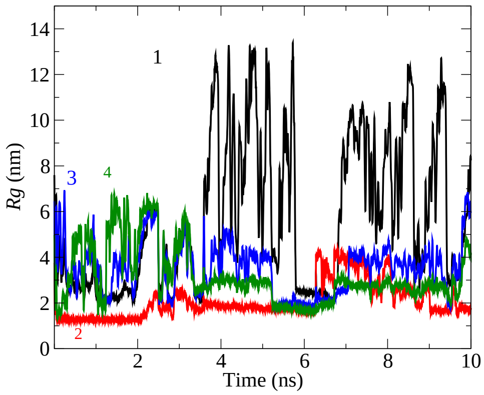

Figure 10. Time evolution of the gyration radius of four domains of tau at $T = 300$ K, in a metadynamics simulation. Curve labels and colors as in Fig.7 [27].

The metadynamics algorithm drives the system to distant points of the space phase, improving the statistical sampling by producing likely and less likely configurations. Due to this drive, transient tertiary structures last short times, probably shorter than their natural lifetime in an unbiased dynamics. Therefore, a contact map of the molecule averaged over all configurations produced by the metadynamics simulation could highlight only resilient tertiary structures. A contact map of tau computed during the 10 ns metadynamics trajectory clearly shows a statistically relevant tertiary pattern entailing a long segment encompassing the N-terminal and the proline-rich domain (residues 120-190) in anti-parallel proximity of a segment encompassing the proline-rich domain and the repeats domain (residues 200-280) [27]. There must be a turn joining these two segments around residue 195, right in the middle of the proline-rich domain. Experiments have shown that the approach of the N-terminal to the central region of the molecule is involved in the aggregation process leading to the formation of PHFs [29]. It is also noteworthy that the end of this hairpin structure (residues 275-280) is the hexamer VQIINK, also known to be involved in this aggregation process [13, 14, 15].

6. Fit of Experimental SAXS Data

We have improved the information on the equilibrium behavior of tau, obtained through the data produced by the computer simulation, by comparing them with experimental SAXS results obtained from a specimen of tau in solution. The SAXS experiment has been performed using full-length htau40. SAXS measurements were acquired on the BioSAXS beamline (ID 14-3) at the Synchrotron Radiation Facility ESRF (Grenoble, France) [30], at the constant temperature of 303 K. Solvent scattering was measured to allow an accurate subtraction of the background scattering. Fig.11(a) shows the result of this experiment. More details on the experiment are given in [28].

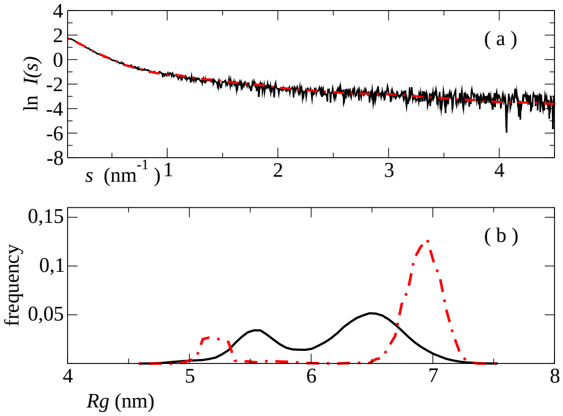

Figure 11. Panel (a): experimental SAXS curve (black continuous line); fit by an ensemble of conformers produced in a 30 ns standard MD simulation and selected by the genetic algorithm GAJOE (red dashed line). Panel (b): distribution of R_g values. Conformers produced by the simulation (black continuous line) and conformers selected by the genetic algorithm (red line), after addition of the coordinated water layer [24].

NMR could provide an alternative set of experimental data to compare with the results of the computer simulation. The program PALES, an atomic resolution approach to alignment tensor prediction, allows the computation of alignment tensors for a given configuration of the molecule; the same program calculates residual dipolar couplings that can be compared to the experimental ones [14].

We have used the SAXS curve to extract from our standard MD simulation [24] and from the metadynamics simulation [27] an ensemble of conformers that give the best fit of the experimental data.

The fitting procedure is as follows: (i) we extract about 9000 regularly spaced conformers of tau from our 30 ns standard MD simulation, or 10000 conformers of tau, regularly spaced by a 1 ps interval, from our 10 ns metadynamics simulation; (ii) this pool of conformers is processed by the program CRYSOL [31] to obtain the theoretical SAXS pattern of each conformer, taking properly into account the scattering from the hydration shell of the water layer coordinated with the molecule; (iii) the ensemble optimization method EOM with the genetic algorithm GAJOE [32] is then employed to select from the pool of theoretical SAXS curves an ensemble of conformers (162 from the standard MD run, 194 from the metadynamics run) that provide with their averaged theoretical scattering intensity the best fit of the experimental SAXS data; each conformer is weighed with its genetic multiplicity [32, 33].

The application of the genetic algorithm GAJOE over 1000 cycles to select the best ensemble of conformers progressively yields an improved fit of the SAXS data [6]. The final values of χ^2 attest the accuracy of the fit: 1.4 for the standard MD run, 1.0 for the metadynamics run. The statistical sampling of tau's phase space achieved by metadynamics is more accurate than the one achieved by standard molecular dynamics. This is an expected result, due to the larger portion of the phase space sampled by the metadynamics simulation.

As shown in Fig.11(a) and Fig.12(a), the selected ensembles fit quite well the experimental SAXS results, when each conformer is weighed with the appropriate multiplicity determined by the genetic algorithm. We show in Fig.11(b) and Fig.12(b) the distributions of R_g values of both the original pool (black line) and of the selected ensemble (red line), respectively. It may be noted that most selected conformers in the standard MD run belong to the first 5 ns of the trajectory, where the value of R_g is near to its initial equilibrium value; but there is also a significant presence of conformers with R_g values between 5.1 and 5.4 nm, belonging to the temporarily stabilized trajectory stretch between 18 and 24 ns (Fig.6). In the metadynamics run the R_g values of the ensemble selected by the genetic algorithm show a distribution peaked near the experimental value $R_g = 6.6$ nm, with a significant presence of conformers with R_g values between 3.0 and 10.5 nm. The radius of gyration averaged over this ensemble is $R_g = 6.8$ nm, with a standard deviation of 1.7 nm. It is in very good agreement with the theoretical value expected for a 441 amino acids random coil in solution, which is 6.9 nm [34].

The distribution of the R_g values before and after the selection performed by means of EOM with the genetic algorithm reflects the higher efficiency of metadynamics in sampling different regions of the phase space, in comparison to a standard MD simulation of similar duration. This can be seen comparing panels in Fig. 11(b), standard MD, and in Fig. 12(b), metadynamics. The original pool in panel 12(b) encompasses a much broader range of R_g values than in panel 11(b). The selected ensemble distribution shown in Fig.12(b) appears to be broader and therefore statistically more significant than the one shown in Fig.11(b), and it is also approximately centered on the experimental value $R_g = 6.6$ nm.

[6] GAJOE parameters were set as follows: number of generations 1000; number of ensembles 50; number of curves per ensemble 20; number of mutations per ensemble 10; number of crossings per generation 20.

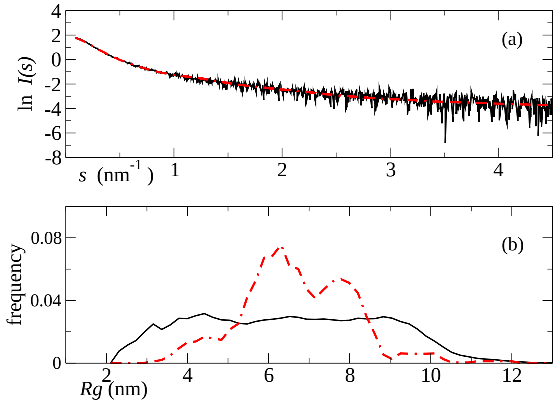

Figure 12. As in Fig.11, but for an ensemble of conformers produced in a 10 ns metadynamics simulation [27].

7. Transient Structures

The ensemble of 162 conformers selected from the standard MD run has been analyzed with the DSSP program [35, 36], as implemented in GROMACS, to identify secondary structures like coils, β-sheets, β-bridges, bends, turns, and α-helices. The propensity of the molecule to form these secondary structures, measured by the number of residues in each structure, turns out to coincide with the average number measured during the whole 30 ns MD run, within one standard deviation of the latter. This confirms the validity of the dynamical simulation as far as local secondary structures are concerned, notwithstanding a possible shortcoming of the force field with regard to the overall shape.

Comparing again our results with those obtained by Mukrasch and coworkers [7], we find a better agreement than in the standard MD run: 15 residues in β-structures and 6 residues in α-helices, to compare with the experimental results: 12 residues and 4 residues, respectively.

Fig.9 and Fig.10 clearly show the thorough shaking of the molecular structure produced by the metadynamics algorithm. This dynamics drives the system to distant points of the phase space and thus to very different global folds. Fig.13 displays the instant contact maps ($C_\alpha - C_\alpha$ distance between all pairs of residues) of four configurations, chosen among the pool of 194 conformers selected by the genetic algorithm. Three are among those selected with highest frequency (panels (a), (c), and (d)); the fourth (panel (b)) further illustrates the variety of global folds. Only distances smaller than 1.5 nm are on display.

The hairpin pattern described before is embedded in various hairpin or paperclip transient tertiary structures, as shown in panels (b), (c), and (d) of Fig.13. While all those configurations are transient, they share this common, persistent tertiary motif: a hairpin folding encompassing part of the N terminal, the proline-rich domain, the first repeat, and a functionally relevant part of the second repeat. As mentioned before, hairpin configurations [7, 9, 10] and paperclip configurations [5, 11, 12] are believed to be a precursor stage of the polymerization of tau leading to the formation of fibrils and to the onset of neurodegenerative diseases.

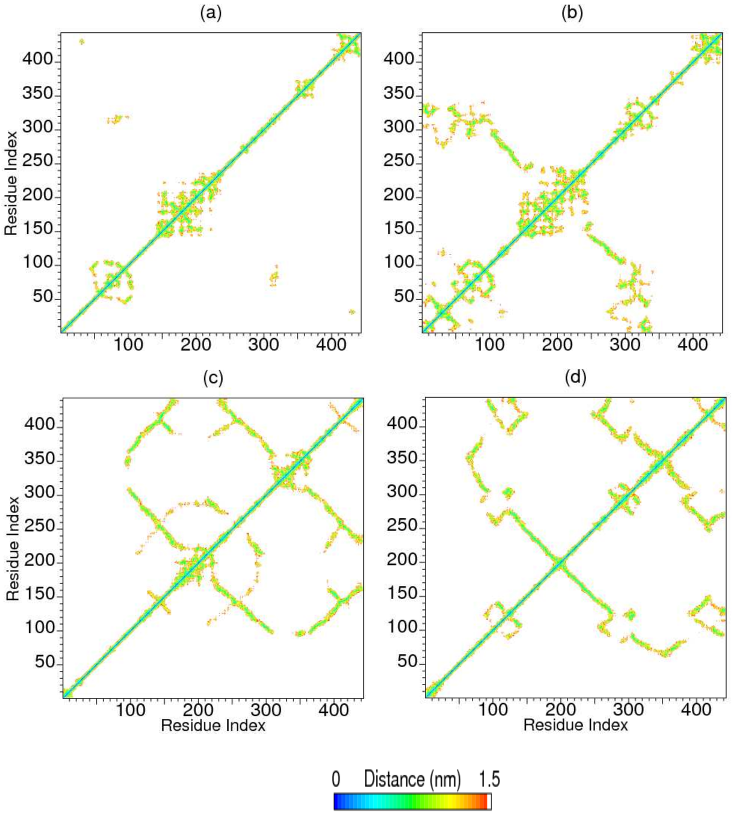

Figure 13. Instant contact maps ($C_\alpha - C_\alpha$ distance between pairs of residues) at $t = 41$ ps (a), $t = 1027$ ps (b), $t = 4228$ ps (c), $t = 7567$ ps (d).

8. Conclusion

Given the speed at which transitions between conformations of an IDP are supposed to take place, the computer simulation of their dynamics seems to be a promising tool to understand their behavior. We present a method that can be used for any IDP. In order to start a MD simulation of a fully disordered protein of unknown 3D structure, one begins from its primary sequence of amino acids. One then implements the following procedure to produce a 3D structure to start the simulation. (i) One creates a first 3D structure by feeding the VMD program (or a similar program) with the primary sequence of the whole protein. (ii) The resulting structure - a multi-rod-like sequence of amino acids - is put in a large box with periodic boundary conditions; after a short energy minimization this structure is taken as the initial one for a dynamical evolution *in vacuo* or in implicit solvent at the chosen temperature, performed with the package GROMACS or with any other simulation program; this step produces a rapid contraction of the protein. (iii) The evolution is stopped when the decreasing gyration radius R_g (or a similar shape-dependent variable) has reached its average experimental value. This yields a starting point for the simulation in a more realistic environment, i.e. with the addition of solvent. (iv) The simulation box is reduced to fit the reduced size of the protein and filled with solvent, either explicit or implicit.

This significant reduction of the volume of the box greatly reduces the number of solvent molecules needed to fill it in the case of explicit solvent. (v) The energy of the system (protein + solvent) is then minimized in the case of explicit solvent, to allow the solvent molecules to adapt to the shape of the solute molecule. (vi) A short equilibration (about 100 ps) is performed at constant temperature. (vii) Another short equilibration (about 100 ps) is performed at constant temperature and pressure. (viii) One uses the last conformation of the previous step to start an extended simulation at constant temperature and pressure.

If the simulation focuses on the analysis of secondary structures, the best simulation is a standard MD one in explicit solvent. If the simulation focuses on the analysis of tertiary structures, the best simulation is metadynamics in implicit solvent.

The statistical information gathered in both cases can be refined by fitting, where available, experimental data of the protein, like small-angle X-ray scattering results. In this step one extracts an ensemble of 'best' configurations from the pool of all configurations produced in the simulated dynamics. This ensemble is produced by means of an ensemble optimization method, implementing a genetic algorithm. This set of conformers, selected from the simulated dynamics by fitting the experimental data, is the best approximation of an equilibrium ensemble that can be extracted from the simulated dynamics; it provides a significant observation of secondary structures and of the overall fold of the molecule. Furthermore, the selected ensemble of conformers represents a 3D data basis that one can use to start further simulations.

When applied to protein tau this method shows that the protein samples a limited number of almost stable secondary structure motifs (mainly short α and β structures), whereas the overall preferred conformation is a random coil entailing a hairpin motif. The transient nature of these structures in protein tau is relevant, and has biochemical implications *in vivo*: the hairpin motif is supposed to be involved in the early stages of tau's polymerization, leading to the formation of pathogenic paired helical filaments.

If the protein under study is only partially disordered, entailing regions of stable and known 3D structure, the method can be implemented by modifying only step (i): the program VMD is used to create multi-rod-like 3D structures of the disordered regions, which are then connected to the ordered regions. The following steps are unchanged,

Aknowledgment

This chapter derives in part from three articles published in Molecular Simulation on 1/2/2012, 26/4/2012, and 24/10/2013, copyright Taylor & Francis, available online at:
http://www.tandfonline.com/10.1080/08927022.2011.608671
http://www.tandfonline.com/10.1080/08927022.2011.633347
http://www.tandfonline.com/10.1080/08927022.2013.794275

References

[1] Tompa, P. Intrinsically disordered proteins. In: Sussman J, Silman I editors. *Structural Proteomics and its Impact on the Life Sciences.* Singapore: World Scientific; 2008; pp. 153-180.

[2] Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, and Obradovic Z. Intrinsic Disorder and Protein Function. *Biochemistry*, 2002, 41, pp. 6573-6582.

[3] Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, and Dunker AK. DisProt: the Database of Disordered Proteins. *Nucl. Acids Res.*, 2007, 35, pp. D786-D793.

[4] Tompa P. Intrinsically Unstructured Proteins. *Trends Biochem. Sci.*, 2002, 27, pp. 527-533.

[5] Mylonas E, Hascher A, Bernadó P, Blackledge M, Mandelkow E, and Svergun DI. Domain Conformation of Tau Protein Studied by Solution Small-Angle X-ray Scattering. *Biochemistry*, 2008, 47, pp. 10345-10353.

[6] Tompa P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.*, 2005, 579, pp. 3346-3354.

[7] Mukrasch MD, Bibow S, Korukottu J, Jeganathan S, Biernat J, Griesinger C, Mandelkow E, and Zweckstetter M. Structural Polymorphism of 441-Residue Tau at Single Residue Resolution. *PLoS Biol.*, 2009, 7, pp. 399-414.

[8] Avila J, Lucas JJ, Pérez M, and Hernández F. Role of Tau Protein in Both Physiological and Pathological Conditions. *Physiol. Rev.*, 2004, 84, pp. 361-384.

[9] Carmel G, Mager EM, Binder LI, and Kuret J. The structural basis of monoclonal antibody Alz50s selectivity for Alzheimers disease pathology. *J. Biol. Chem.*, 1996, 271, pp. 32789-32795.

[10] Gamblin TC, Berry RW, and Binder LI. Tau Polymerization: Role of the Amino Terminus. *Biochemistry*, 2003, 42, pp. 2252-2257.

[11] Jeganathan S, von Bergen M, Brutlach H, Steinhoff HJ, and Mandelkow E. Global Hairpin Folding of Tau in Solution. *Biochemistry* 2006, 45, pp. 2283-2293.

[12] Bibow S, Mukrasch MD, Chinnathambi S, Biernat J, Griesinger C, Mandelkow E, and Zweckstetter M. The dynamic structure of filamentous tau. *Angew. Chem. Int. Ed.*, 2011, 50, pp. 11520-11524.

[13] von Bergen M, Barghorn S, Biernat J, Mandelkow E-M, and Mandelkow E. Tau aggregation is driven by a transition from random coil to beta sheet structure. *Biochimica et Biophysica Acta*, 2005, 1739, pp. 158-166.

[14] Mukrasch MD, Markwick P, Biernat J, von Bergen M, Bernadó P, Griesinger C, Mandelkow E, Zweckstetter M, and Blackledge M. Highly Populated Turn Conformations in Natively Unfolded Tau Protein Identified from Residual Dipolar Couplings and Molecular Simulation. *J. Am. Chem. Soc.*, 2007, 129, pp. 5235-5243.

[15] von Bergen M, Friedhoff P, Biernat J, Heberle J, Mandelkow E-M, and Mandelkow E. Assembly of tau protein into Alzheimer paired helical filaments depends on a local sequence motif ((306)VQIVYK(311)) forming beta structure. *Proc. Natl. Acad. Sci. USA*, 2000, 97, pp. 5129-5134.

[16] Sawaya MR, Sambashivan S, Nelson R, Ivanova MI, Sievers SA, Apostol MI, Thompson MJ, Balbirnie M, Wiltzius JJW, McFarlane HT, Madsen AØ, Riekel C, and Eisenberg D. Atomic structures of amyloid cross-spines reveal varied steric zippers. *Nature*, 2007, 447, pp. 453-457.

[17] Friedhoff P, von Bergen M, Mandelkow E-M, and Mandelkow E. Structure of tau protein and assembly into paired helical filaments. *Biochimica et Biophysica Acta*, 2000, 1502, pp. 122-132.

[18] Humphrey W, Dalka A, and Schulten K. Visual Molecular Dynamics. *J. Mol. Graph.*, 1996, 14, pp. 33-38.

[19] Battisti A and Tenenbaum A. Molecular dynamics simulation of intrinsically disordered proteins. *Mol. Simulat.*, 2012, 38, pp. 139-143.

[20] Arteca GA, Reimann CT, and Tapia O. Proteins *in vacuo*: Denaturing and folding mechanisms studied with computer simulated molecular dynamics. *Mass Spectrosc. Rev.*, 2001, 20, pp. 402-422.

[21] Still C, Tempczyk A, Hawley RC, and Hendrickson T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.*, 1990, 112, pp. 6127-6129.

[22] Roe DR, Okur A, Wickstrom L, Hornak V, and Simmerling C. Secondary Structure Bias in Generalized Born Solvent Models: Comparison of Conformational Ensembles and Free Energy of Solvent Polarization from Explicit and Implicit Solvation. *J. Phys. Chem. B*, 2007, 111, pp. 1846-1857.

[23] Tan C, Yang L, and Luo R. How Well Does Poisson-Boltzmann Implicit Solvent Agree with Explicit Solvent? A Quantitative Analysis. *J. Phys. Chem. B*, 2006, 110, pp. 18680-18687.

[24] Battisti A, Ciasca G, Grottesi A, Bianconi A, and Tenenbaum A. Temporary secondary structures in tau, an intrinsically disordered protein. *Mol. Simulat.*, 2012, 38, pp. 525-533.

[25] Laio A and Gervasio FL. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog.Phys.*, 2008, 71, 126601 (22 pp).

[26] Still WC, Tempczyk A, Hawley RC, and Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc., 1990, 112, pp. 6127-6129.

[27] Battisti A, Ciasca G, and Tenenbaum A. Transient tertiary structures in tau, an intrinsically disordered protein. *Mol. Simulat.* 2013, 39, pp. 1084-1092.

[28] Ciasca G, Campi G, Battisti A, Rea G, Rodio M, Papi M, Pernot P, Tenenbaum A, and Bianconi A. Continuous thermal collapse of the intrinsically disordered protein tau is driven by its entropic flexible domain. *Langmuir*, 2012, 28, pp. 13405-13410.

[29] Gamblin TC, Berry RW, and Binder LI. Tau Polymerization: Role of the Amino Terminus. *Biochemistry* 2003, 42, pp. 2252-2257.

[30] Pernot P, Theveneau P, Giraud T, Nogueira RF, Nurizzo D, Spruce D, Surr J, McSweeney S, Round A, Felisaz F, Foedinger L, Gobbo A, Huet J, Villard C, and Cipriani F. New beamline dedicated to solution scattering from biological macromolecules at the ESRF. *J. Phys.: Conf. Ser.*, 2010, 247, 012009 (8 pp).

[31] Svergun DI, Barberato C, and Koch MHJ. CRYSOL a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.*, 1995, 28, pp. 768-773.

[32] Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, and Svergun DI. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *J. Am. Chem. Soc.*, 2007, 129, pp. 5656-5664.

[33] EOM manual http://www.embl-hamburg.de/biosaxs/eom.html

[34] Kohn JE, Millett IS, Jacobs J, Zagrovic B, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TS, Hasan MZ, Pande VS, Ruczinski I, Doniach S, and Plaxco KW. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. USA*, 2004, 101, pp. 12491-12496.

[35] Kabsch W and Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 1983, 22, pp. 2577-2637.

[36] Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, and Vriend G. A series of PDB related databases for everyday needs. *Nucleic Acids Res.*, 2011, 39, pp. D411-D419.

MA