Exploring energy landscapes of protein folding and aggregation

Normand Mousseau¹, Philippe Derreumaux²

1. ABSTRACT

Human diseases, such as Alzheimer’s and Creutzfeldt-Jakob’s are associated with misfolding and aggregation of specific proteins into amyloid fibrils sharing a generic cross-beta structure (1). The self-assembly process is complex, but once a nucleus is formed, rapid fibril formation occurs. Insight into the structures of the oligomers during the lag phase, varying between hours and days, is very difficult experimentally because these species are transient, and numerically using all-atom molecular dynamics because the time scale explored is on the order of 10-100 ns (2, 3). It is therefore important to develop simplified protein models and alternative methods to sample more efficiently the conformational space. In the past few years, we have developed the activation-relaxation technique (ART nouveau) coupled to the OPEP coarse-grained force field. This review reports the application of ART-OPEP on protein folding and aggregation.

2. INTRODUCTION

A remarkable property related to amyloid plaque formation is that this process is shared by related-disease proteins (e.g. the Aβ protein for Alzheimer’s disease and the prion protein for Creutzfeldt-Jakob’s disease) and non-homologous peptides of variable length as short as four amino acids. We can therefore attack the complexity of the aggregation process on two fronts by simulating either low-order species (e.g. monomers and dimers) of full-length proteins or higher-order species (e.g. 10-mers) of short peptides (4). In this review, we report the application of the activation-relaxation technique (ART nouveau) coupled to the OPEP force field on both types of systems and focus specifically on the knowledge that can be obtained regarding the energy landscape of protein aggregation. Because ART-OPEP has advantages and limitations as any method, it is important to show that this protocol provides pertinent information on the structures and folding...
The difficulty of representing the energy landscape limits considerably the usefulness of this picture for understanding the dynamical properties of complex systems such as proteins and glasses, and most efforts in the last two decades have focused on developing clever projections that offer a usable description of the dominant properties of the energy landscape. Of course, no unique projection of a high-dimensional problem onto a one or two-dimensional space can provide the full picture; some can even lead to conflicting understanding if inappropriate coordinates are selected.

In spite of the challenge associated with finding a meaningful representation, the concept of energy landscape provides a very convenient basis for discussing the kinetic and dynamical properties of complex systems, such as proteins, that cannot easily be analyzed using more standard representations. This was the case, for example, for the solution of Levinthal's paradox, stated in the 1960's (11): how can a protein always fold into its native state if the number of conformations available reaches astronomical proportion? A first convincing answer was provided by Wolynes and collaborators (12), based in large part on theoretical studies of proteins on lattices: natural selection has chosen amino acid sequences that have a strong bias toward folding into their native state or, more concisely, proteins have a funnel-shape energy landscape that brings any initial state rapidly into the same minimum free-energy structure.

The standard funnel picture, presented in the top panel of Figure 2, shows a very deep and steep landscape. While the protein can be trapped briefly in some metastable states, that can be disordered or partially ordered (or even fully ordered but non-native (13)), the overall trend is strongly towards the native state. Of course, this representation is a strong simplification as in reality the funnel is 3N dimensional, where N is the number of atoms. For example, the two-dimensional funnel drawn in Figure 2 does not explain the two-state model observed in thermodynamical measurements of folding and unfolding of simple proteins. Moreover, pathways in a high-dimensional space are much more complex than in 2D and one would imagine an overall funnel where the peptide fails to reach a native state.

To circumvent this representation problem, Becker and Karplus (14) introduced the disconnectivity graph, a new projection from the energy landscape based on a distance both in configuration space (separated by a number of adjacent minima) and in energy. The typical funnel landscape can be represented here as in the bottom panel of Figure 2. The bottom of each leaf indicates the minimum-energy of a given configuration and the position of the branch, at the top of the leaf, the relative activation-barrier to go from one minimum to another connected to the same branch. Although clear spatial information is now lost, information regarding the dynamics is highlighted through the clear appearance of relative barriers, helping identify efficiently the nature of the dominant dynamics with a detailed knowledge of the various activated mechanisms.
Exploring energy landscapes of protein folding and aggregation

Figure 2. The bottom of each leaf indicates the minimum-energy of a given configuration and the position of the branch, at the top of the leaf, the relative activation-barrier to go from one minimum to another connected to the same branch. Although clear spatial information is now lost, information regarding the dynamics is highlighted through the clear appearance of relative barriers, helping identify efficiently the nature of the dominant dynamics with a detailed knowledge of the various activated mechanisms.

Recently, Rao and Caflisch introduced yet another graph (15) that goes beyond the funnel picture with a representation in terms of a network representing the connectivity between various basins, defined using a standard Cartesian metric. Instead of focusing solely on the native state, this representation helps recreate the ensemble of folding trajectories and emphasizes the existence of a number of quasi-essential intermediate states for folding. While this approach has not been used much since its introduction (16), it provides direct kinetic information that can be missed in simpler representations.

3.2. The free-energy landscape

It is possible to use similar concepts for the free-energy landscape. However, one must be careful because the free energy involves an integral to define the entropy. In the phase space, which defines the free energy, there is no landscape since every point has the same probability (17). As such, the concept of free-energy landscape is only valid when projected onto a relatively low-dimensional space, with the other degrees of freedom integrated, providing sufficient information for defining the entropy. As was pointed out by Krivov and Karplus (18), however, the relevant space of reaction coordinates can still be much larger than two dimensions so that the planar projection, necessary for the visualization, can introduce artifacts leading to an inaccurate description of the dominant free energy barriers.

4. EXPLORING THE LANDSCAPE

One of the major difficulties associated with the use of the concept of energy landscape is the need for constructing it, sampling at least the most important parts. For fast systems, such as liquids, molecular dynamics simulations are sufficient to sample the relevant configurational space. The problem is more challenging for proteins since the fastest ones fold on a µs timescale, and more sophisticated approaches are needed.

These methods can be classified into three groups: (1) biased sampling, (2) reweighting methods and (3) activated approaches. The first set of methods introduces biases in the sampling to force the system to visit less frequent parts of the phase space. These methods include transition path sampling (19), metadynamics (20) and milestoning (21), for example. The challenge here is to find the appropriate reaction coordinate to drive the bias efficiently. In the absence of such a coordinate, these methods become increasingly expensive with more complex systems, when the relevant part of the phase space becomes too small or when the barriers are important compared with the melting temperature. The second set of methods, such as Wang-Landau algorithm (22), replica exchange molecular dynamics (REMD) (23) and, to a lesser degree, the weighted-histogram method (WHAM) (24), focuses on producing a more balance sampling of the energy landscape or, in the case of WHAM, on extracting as much information as possible from independent simulations. While these approaches can help the sampling, they are limited by the efficiency of moving from one state to another and have not delivered as much as we could have expected initially for large and complex systems.

The final set of methods is appropriate for systems with a dynamics dominated by activated mechanisms (21, 25-31). This is the case, for example, of proteins, at least in implicit solvent. Activated methods generate pathways connecting adjacent local minima via first-order saddle points, reconstructing physical pathways as defined by the transition state theory (6). The phase space is therefore reduced to a discrete network composed of local minima connected via activated states, decreasing considerably the number of states to sample (see Figure 3). This network can be generated through continuous trajectories such as with the activation-relaxation technique (ART nouveau) (27) or assembled through a random and biased search (32). Activated methods are not sensitive to the height of the barriers and can therefore move rapidly.
Exploring energy landscapes of protein folding and aggregation

Figure 3. Activated methods, such as ART nouveau, sample a much reduced configurational space consisting only of local minima (large blue dots) connected by pathways defined by the position of the transition state (small purple dots) explaining, in part, the efficiency of these methods.

Figure 4. Cartoon representing the first steps of an ART nouveau event. Starting from a local minimum, the structure is pushed in a random direction until a direction of negative curvature appears. The structure is brought along this direction, minimizing its energy in the perpendicular hyperplane until it reaches an adjacent first-order saddle point.

Through the configurational space. They lack a proper thermodynamical basis, however, which restricts their application. In spite of these limitations, activated approaches have provided many insights in protein folding.

4.1. The activation-relaxation technique

There exist many methods for identifying transition states. Most of them, such as the rubber band method of Elber and Karplus (28), the variational Verlet algorithm (29), the nudged elastic band method (NEBM) (30) or milestoning (21) require the knowledge of both the initial and final state in addition to a rough guess of the transition. Moreover, the initial and final states must be nearby, separated by at most a few intermediate states. Longer pathways are unreliable since the resulting trajectory is strongly dependent on the initial guess.

Of course, for proteins, we do not know the intermediate states and we must turn to methods that can identify transition states without a priori knowledge of the final state, such as ART nouveau (27) or similar methods such as the dimer (31) and the eigenvector-following methods (26, 33). While these methods are basically equivalent (34), we focus here on ART nouveau that was developed by one of us.

An ART nouveau (simply called ART, for shortness, in the rest of this review) event can be divided in four steps (Figure 4):

(i) starting from a local minimum, the system is pushed slowly in a random direction; at each step, we evaluate the value of the direction of lowest curvature (the lowest eigenvalue of the Hessian matrix); when the lowest curvature becomes negative, we stop the push in this random direction;

(ii) the configuration is then displaced along the direction of negative value while the energy is minimized in the orthogonal directions;

(iii) when the total force reaches a value close to zero, the system has converged onto a transition state (first-order saddle point);

(iv) the system is then moved over the saddle point and relaxed into a new minimum, completing the event defined by the initial minimum-transition state-final minimum.

Let us review these steps in some details. For a small peptide, up to about 15 amino acids, the initial displacement is taken as a random vector in the \((3N-6)\) dimensional space where rotational and translational degrees of freedom are removed; all atoms are therefore displaced, each in a random direction. For a larger protein or ensemble of peptides, the initial random displacement is generally made on a subset of all atoms, to allow for a faster sampling. In all cases, the initial displacement excludes moves along the strong forces associated with covalent bonds and bond angles since these directions are strictly harmonic by definition.

Once the initial direction is set, the protein is then deformed very slowly, displacing all atoms by a total of about 0.4 Å at each step. After each displacement, the protein conformation is relaxed in the hyperplane perpendicular to the move away from the initial minimum.
Exploring energy landscapes of protein folding and aggregation

This step prevents building too much stress in the system during the first phase of activation for a better control of the activation. After each move, the value of the lowest curvature is calculated, from the force field, using the Lanczos algorithm. Starting with a random vector, we apply the algorithm to construct a small tridiagonal matrix that we can diagonalize, obtaining value of the lowest curvature and the corresponding eigenvector. Reusing the result as a starting point for another application of the Lanczos method, we converge rapidly onto a stable eigenvector that evolves slowly as iterations progress.

The presence of a first-order saddle point is signaled by the appearance of a negative eigenvalue in the Hessian. To prevent the system from going back into the harmonic basin, with only positive curvature, we set a threshold below zero for starting the convergence to the transition state. This threshold depends on the potential and the system under study. The choice of this parameter is not very critical and it affects the success rate of the activation process but not the type of events found (27).

Once the edge of the harmonic basin is reached, the system is pushed along the direction of negative curvature, away from the initial minimum. While the initial push might involve only a subset of all the atoms, here, the whole system is free to move. After each step along the eigenvector, the Hessian is diagonalized again, with Lanczos algorithm, and the system is relaxed in the hyperplane perpendicular to this direction using an adaptive steepest-descent algorithm. Iterations stop when the absolute value of the force perpendicular to the ridge falls below typically 0.1 kcal/ (mol Å) and the total force, below 1.0 kcal/ (mol Å), meaning that a new transition state has been identified, or when the lowest curvature become positive, in which case we start again from the initial minimum.

From the saddle point, the structure is pushed away from the initial minimum along the direction of negative curvature and then relaxed, using damped molecular dynamics, into a minimum. Most of the time, the final state differs from the initial configuration and an event has been generated. In some case, however, the saddle point identified corresponds to a shoulder in the potential energy landscape and the structure goes back to the initial minimum.

Once an event is generated, it can be used in multiple ways. If one is interested in characterizing the energy landscape around a single minimum, then this event is simply stored and a new event is started from the same point (27, 35). It is also possible to perform an accept/reject, using the Metropolis criterion, for example,

\[ p_{\text{accept}} = \min \left[ 1, \exp \left( -\Delta E / k_B T \right) \right] \]

with either the activation barrier (16) or the asymmetry energy (the energy of the final state minus that of the initial minimum) (27). In general, we have used the latter criterion as we would need to know the complete distribution of barriers around a local minimum in order to use the activation barrier meaningfully. By using only the asymmetry between adjacent minima, the method ensures that at least each configuration is weighted appropriately with respect to the others.

Wales and collaborators have used events generated with a similar method to construct a sparse network similar to that of Figure 3, from which it is possible to extract the fastest folding pathway, for example (36, 37). This approach, called discrete path sampling, has been used to study folding of a β-hairpin, for example (37). While discrete path sampling provides very useful kinetic information, the numerical effort needed to extract pathways makes it difficult to assess the impact of a missed barrier or varying prefactors that could modify the dominant folding pathway. Moreover, the quantity of events needed to construct a sufficiently connected network increases rapidly with the number of degrees of freedom so that this method is difficult to apply to larger systems.

In this Review, we focus on the more straightforward approach to sampling the landscape using a Metropolis criterion based on the energy difference between the initial and the final minima at each event. By doing so, we generate a fully connected walk through the energy landscape. Because this constructed landscape is reduced to a discrete set of local minima, ART does not include entropic contributions. The temperature used in the Metropolis accept/reject procedure is therefore unphysical and is generally set in such a way that the trajectories do not remain trapped in metastable states.

The main reserve about this approach is that the ART-constructed trajectories do not respect detailed balance nor belong to a well-defined thermodynamical ensemble. This is due to the fact that we do not know what the bias of ART is for finding one nearby saddle point instead of another. This criticism is valid. However, numerous tests and comparisons, many described below, find a strong similarity between the ART folding trajectories and those obtained from more standard techniques such as molecular dynamics.

We have investigated this question in a comparative ART study of the β-hairpin, using two different Metropolis criteria: one based on the barrier height and the other on the standard asymmetric energy (16). As can be seen from Figure 5, which superposes some of the intermediates explored by two runs using different Metropolis criteria, we find results in agreement with what can be expected from the network picture of folding observed by Rao and Caflisch (15). Specifically, there exist a number of well-connected intermediate structures, or hubs, that seem to connect various parts of the configurational space. Irrespective of the method used, these intermediate structures will be found as long as the pathway generated is continuous. Even though the kinetic varies considerably from one specific folding trajectory to another, the overall picture remains unchanged.

ART-generated trajectories are possible pathways that should capture the qualitative feature of folding or
Exploring energy landscapes of protein folding and aggregation

Figure 5. Top: Two folding trajectories using a different Metropolis accept/reject criterion based on either the energy barrier or the energy difference between the final and the initial minima. Bottom: Superposition of conformations found in these two sets of simulations. For each pair of conformations, we indicate the energy and the RMSD. Even though the detail of folding depends on the criterion used, we find the same intermediate structures.

aggregation since these are dominated by well-connected hubs. Because the relative weight of one hub over the other is subtle, the relative probability of the various folding mechanisms identified by ART cannot be used. While this might be construed as somewhat of a drawback, comparisons with various experiments and other simulations show rather that the trajectories identified with this method cover a wide range of specific conditions and can therefore be used to understand experimental setups that cannot easily be simulated, for example.

4.2. The OPEP potential

The OPEP (Optimized Potential for Efficient prediction of Protein Structure) is a generic force field that can be used for any amino acid sequence, with L- or D-amino acids. It is based on a coarse-grained representation of the amino acids, where all backbone atoms are included and all side chains are represented by a specific bead (38). While such a reduction in the number of degrees of freedom was proposed in 1976 by Levitt and Warshel (39), the originality of OPEP lies in a good compromise between energy accuracy, structural precision and computational cost (40). A force field with implicit solvent and coarse-graining of the side-chains cannot capture all the details of the interactions between the side-chains and between the side chains and the solvent, as the all-atom molecular mechanics AMBER force field (41) or all-atom spectroscopic force fields (42). However, we find it possible to design an energy function that discriminates native from non-native structures on an ensemble of 30 proteins (43), and predict, using Monte Carlo simulations, lowest-energy structures consistent with experimental data (44, 45). Full details of the energy function balancing side-chain – side-chain interactions (with both hydrophobic and hydrophilic components), dihedral angles of the backbone and hydrogen-bonding interaction can be found in Ref. (43). As discussed in this review and the next article of this series, OPEP has been coupled to ART nouveau (46), molecular dynamics simulations (47) and REMD simulations (48). Of interest for this review is that the OPEP-generated free energy surface of small proteins is fully consistent with experimental data, independently of the starting conformation used, providing further support of the capability of the OPEP force field to reproduce thermodynamical data as well (48).

5. THE β-HAIRPIN

The first step towards developing a full understanding of protein folding is to identify the mechanisms by which the secondary elements, such as α-helices and β-sheets, form. With dominant local interactions, α-helices form rapidly, within a few hundred ns for a 21-residue model (49-51). Beta structures are more difficult to create as they involve non-local interactions. This difficulty is observed even in β-hairpins, where the folding time can be 10 or more times slower than the helices, reaching many µs.

One of the most studied model is the second β-hairpin (GEWTYDDATKTFTVTE) of the domain B1 of protein G, which has been the subject of an extended range of theoretical (52, 53) and experimental studies (54, 55). With a melting point near room temperature, this peptide is not stable enough in solution for high-resolution NMR determination but it adopts, with a significant probability, a β-hairpin conformation near that found within the full-length protein G (54). Moreover, early fluorescence experiments show that this β-hairpin folds, in isolation, with a time constant of 6 µs with a kinetic consistent with that of the two-state folding model, i.e., with only two dominant states: folded and unfolded (55).

If experiments can often offer the details of the native structure, it is much more difficult to obtain information about the folding mechanisms and, from there, a clear explanation of the important folding time difference between the hairpin and a helix (56, 57). There has been significant progress, however, in the recent years. Ising-like models derived from NMR and temperature-jump
Exploring energy landscapes of protein folding and aggregation

experiments (55, 56), that consider native interactions only, suggest that the zipper-mechanism explains much of the experimental data (56). In agreement with this picture, mutations and ϕ-value analysis also point to the importance of forming a loop as the rate-limiting step in the folding of a β-hairpin (58, 59). Dyer et al., using FTIR and T-jump relaxation, conclude that the loop rearrangement could be the rate-limiting (60).

Much of the experimental picture depends finely on the models used to explain the data. To validate those, we must turn to simulations in order to follow precisely the sequence of events leading to the folded state. But 6 μs is a long time for computer simulations, and many new approaches, attempting to overcome the time limitation, have been used to try to solve this problem, transforming this fundamental question into a test-bed of new computational methods applied to the protein folding problem. The hairpin was therefore studied under a wide range of representations that go from all-atomic empirical force-field with explicit (61-73) and implicit solvent (18, 71, 74-76) to coarse-grained off-lattice potentials with implicit solvent (52, 53, 77-79), biased-potentials (80, 81) and minimal models (82).

The simulation methods are as varied as the molecular descriptions with straightforward unfolding (83-85) and folding molecular dynamics (66, 74, 86), activated methods (52, 53, 73, 75), replica exchange molecular dynamics (64, 72, 87-91), transition path sampling (92), distributed computing (71), self-guided molecular dynamics (93), multicanonical Monte Carlo (79) and other free-energy calculation methods (65, 94, 95), and statistical models (82).

Initial simulation work had identified two dominant folding mechanisms: the zipper out and the zipper in models. In the zipper out mechanism, the turn forms first, with the H-bonds propagating from there. This mechanism was first proposed based on an Ising-like model (55, 82) and observed in lattice Monte Carlo simulations as the dominant model (80). A similar mechanism, controlled by the hydrophobic core near the turn, which forms first and propagates from there, was identified by the same group (80) as well as by off-lattice model Langevin (81) and all-atom MD simulations (96).

The zipper in mechanism is dominated by the hydrophobic core: the two extremities come in proximity pulled in by the hydrophobic core, the H-bonds form nearby and propagate inward, towards the turn. This mechanism was identified in multicanonical MC (79), all-atom unfolding MD (83-85), Gō-based (86), REMD (64) and distributed MD simulations with implicit solvent (71).

ART-OPEP, while suffering from the limitations described in section 4.1, could provide a more comprehensive picture of folding (53, 97). Eighty-two folding simulations were launched at 300 K, from a fully extended conformation, using different initial random seeds. Three sets of parameters were used: the standard OPEP force field (52 runs), a modified OPEP potential (20 runs) and a Gō-like potential (10 runs). From all these 82 trajectories, counting between 4000 and 9000 trial events, 36 reached the native state, including 16 with the standard OPEP parameters. Here the definition of the native structure is very strict: less than 2.0 Å RMSD from the hairpin structure within protein G (PDB code 2GB1), a well defined hydrophobic core and six H-bonds formed. Focusing on the trajectories leading to the native state, Wei and collaborators identified three folding pathways (53, 97): the two pathways already observed, the zipper out (see Figure 6 (a) and zipper in (6b) mechanisms, and a third one, the reptation mechanism. This latter move occurs when a loop forms at the wrong place and the peptide chain is characterized by a network of non-native H-bonds. Slowly, however, following fluctuations, one strand “walks” over the other, in a reptation move, until the two strands are aligned and the peptide forms the native state (6c).

These results were important for two reasons. First, they demonstrated, for the first time, the predictive power of ART-OPEP, showing that even though the algorithm does not have detailed balance, the folding trajectories are physically possible and overlap with other better characterized algorithms such as molecular dynamics. Second, the reptation mechanism underlines the importance of non-native interactions during folding, and puts into question the use of unfolding and Gō-models where non-native interactions play either a small role or none at all.

In the recent years, experimental and numerical analysis of the folding pathway of various β-hairpins has continued. All-atom MD and REMD simulations found that asymmetric hairpins, stabilized by non-native H-bonds, are common in hairpins varying in length between 9 and 16 residues (73, 93, 95, 98, 99). The identification of the reptation move is more difficult, especially with techniques such as REMD that provide only thermodynamical information. Many works have studied shorter β-hairpins, counting as little as 9 residues. Because of the tight turn and the shorter branches, these hairpins seem to show a narrower range of mechanisms. A detailed analysis of an extensive set of MD folding trajectories of the trpzip2 peptide (12 residues) failed to reveal any reptation moves (70). Chen and Xiao, however, studied the 9-residue peptide designed by Blanco et al. (100) with an implicit solvent and found folding mechanisms that incorporated a mixture of the mechanisms discussed above, including reptation (74). Imamura and Chen studied the impact of the position of the hydrophobic interactions on the various folding mechanisms using a coarse-potential with MC (101). They found that if the hydrophobic core is placed at the turn, then the dominant mechanisms should be symmetric, including the zipping out, zipping in and middle out. According to these authors, the reptation move, which they found in their simulations, does not depend on the position of the hydrophobic core.

It is not clear how to reconcile all these different results. Restricting the discussion to the β-hairpin of protein G, the exact nature of the folding pathways remains debated today in part because it is still difficult to obtain

4501
Exploring energy landscapes of protein folding and aggregation

Figure 6. (a) A representative folding trajectory of the zipper out mechanism (mechanism I) (only accepted events are counted): The turn forms first (event 53), which forces the two branches to come together (event 80) and allows the formation of the H-bonds from the center to the end points (events 109 to 471). (b) A representative folding trajectory of the zipper in mechanism (mechanism II): the peptide collapses in a random coil (event 55), then forms partially a helix (events 169 and 280), as it stretches out, the end points meet and form native H-bonds (events 528), the H-bond network propagates towards the turn (events 567 and 600). (c) A representative folding trajectory of the reptation mechanism (mechanism III): the peptide collapses into a random coil with two turns (event 55), a short non-native hairpin forms (event 84), one branch slides over the other, moving the turn in the direction of the center (events 99 and 301), the two branches align themselves (event 302) and the native hairpin forms (event 334).
good statistics with all-atom potentials and explicit solvent. For example, a recent all-atom MD simulation, using GROMOS96, found only 3 folding events in a 278-ns simulation (72) while a recent REMD study with AMBER and implicit Poisson-Boltzmann solvent model identified a single pathway corresponding to the zipper in mechanism (76). While ART-OPEP could not provide the definitive answer, its efficiency allowed enough statistics to identify a new folding mechanism that was observed directly and indirectly afterwards by standard simulation techniques. These results on the β-hairpin demonstrate that the network nature of the protein energy landscape makes ART-OPEP valuable to generate relevant folding and certainly aggregation pathways.

6. AGGREGATION OF SHORT PEPTIDES

The discovery of short peptides with 4-8 amino acids forming amyloid fibrils in vitro (102-104) was a major advance for the computational community as it allows the study of the first steps of amyloid-forming protein aggregation. Simulations remain challenging but we are beginning to see various approaches converging towards a unified picture.

The favorite peptide for computational studies is Aβ (16-22), of sequence KLVFFAE, corresponding to the central hydrophobic core of the amyloid-β protein associated with Alzheimer’s disease. Following the pioneering solid-state NMR study carried out by Balbach et al. (105), this peptide was studied using MD with explicit (106-109) and implicit (110) solvent, REMD with explicit solvent (111), Monte Carlo with reduced representations (87, 112, 113) and ART-OPEP (114, 115). Stability MD simulations support the experimentally-derived anti-parallel organization (108). This orientation is also confirmed in the biased MD aggregation study of three monomers (109). Interestingly, the latter study suggests the obligatory presence of α-helical intermediates on the aggregation pathways (109).

We have generated ART-OPEP simulations on the monomer, dimer (115) and trimer of Aβ (16-22) (114), without introducing any bias towards a specific structure. While the structure of the monomer has not been characterized experimentally, ART-OPEP simulations suggest a random coil structure, in agreement with other simulations (109, 113). For the dimer, simulations are started from either a parallel β-sheet or two α-helices with a Metropolis temperature selected such that the ordered structures are visited while ensuring an overall satisfactory sampling of the conformational space. The lowest-energy minimum is anti-parallel in character with a H-bond network matching the solid-state NMR pattern of the fibril (pattern I in Figure 7) Analysis of the low energy structures shows that the system can populate many local minima.
Figure 8. Basic rearrangement mechanisms observed by ART-OPEP for motion without full detachment. (a) The transition between patterns I and III (previous figure) takes place by a rotation of one chain with respect to the other following by a reptation step. (b) The transition between patterns I and IV was seen to occur by a two-step reptation move. (c) We also observe pathway connecting pattern I to pattern II going through a rotation of both chains. Figure 9. All-atom $\beta$-sheets of higher energies formed by the trimer of A$\beta$ (16-22), as determined by ART simulations. Registries A to E are antiparallel in character, whereas registry F mixes parallel and antiparallel strands. The superposed structures in A, B, and F result from independent runs indicating that these structures can be found through many trajectories.

with anti-parallel orientations of the chains and distinct H-bonding patterns as is shown in Figure 7, and notably the second NMR pattern of Balbach et al. (82). The existence of these metastable dimeric states with various H-bond registries was confirmed by all-atom MD simulations (110).

As seen in Figure 7, ART-OPEP trajectories provide information on the various pathways. The dimer moves through intermediate anti-parallel states with various registries as well as from anti-parallel to parallel organization without the need for $\alpha$-helical intermediates. Even though the parallel $\beta$-sheet, corresponding to the conformation adopted by the full-length A$\beta$ in the fibril, has an energy 3 kcal/mol higher than the anti-parallel sheet, it is sampled in many runs, independently of the initial conformation. This result has also been observed on the same dimer using REMD simulations with the AMBER94 force field (111) and on the trimer of GNNQNY, using CHARMM 19 where various orientations appear as metastable states (116). The reference in the orientation is therefore simply a matter of balancing the various contributions to the configurational energy.

Analyzing the rearrangement mechanisms within the two-stranded $\beta$-sheet, we find that the two chains can detach and reattach. But they can also move by a reptation mechanism similar to that of the $\beta$-hairpin, a rotation of one chain with respect to the other or by a combination of both rotation and reptation moves (see Figure 8). Interestingly, the reptation mechanism was also observed on aggregates of A$\beta$ (16-22), by FTIR experiments (117), confirming our numerical results.

ART-OPEP simulations of the A$\beta$ (16-22) trimer support the results observed on the dimer: the NMR-derived anti-parallel $\beta$-sheet with the 16+k+22-k registry is one of the three lowest-energy structures, (see Figure 9 (a)). The two other lowest-energy structures have one strand shifted with respect to the native state (Figure 9 (b) and (c)). (105). It is important to note that these three predicted conformations are also found to be stable in explicit solvent MD simulations at 330 K for at least 7 ns, using the GROMOS96 force field.

Other groups have also obtained results similar to those of ART-OPEP. Favrin et al. manage to observe the aggregation of dimers, trimers and hexamers using Monte Carlo with moves restricted to $\phi$ and $\psi$ angles, reducing considerable the phase space available, but by a method completely different from ART-OPEP, and yet obtaining very similar results (112). The structures observed for the dimer and trimer are not confined to A$\beta$ (16-22), however. Lei et al., in all-atom MD simulations of the dimer of NVHTLSQ, a seven-residue peptide from human $\beta$2-microglobulin, also find many metastable $\beta$-sheets with various orientations and bonding networks (118).

But ART-OPEP simulations also sample higher energy structures, characterized by very diverse H-bonding patterns and monomer orientations (see Figure 9 d-f). Since the fibril-compatible state is already populated to some extent within a trimer, it is possible that some of the other higher energy trimeric states might share structural similarities with the off-pathway intermediates observed experimentally. This point remains to be determined, however.
Exploring energy landscapes of protein folding and aggregation

Figure 9. All-atom β-sheets of higher energies formed by the trimer of Aβ (16-22), as determined by ART simulations. Registries A to E are antiparallel in character, whereas registry F mixes parallel and antiparallel strands. The superposed structures in A, B, and F result from independent runs indicating that these structures can be found through many trajectories.

As more chains are added it becomes much more difficult to generate structured aggregates for this sequence. For example, recent replica exchange Monte Carlo (REMC) simulations using an all-atom implicit solvent potential on six Aβ (16-22) peptides only find amorphous aggregates with little β-sheet contents (113) even though ordered structures remain stable over many hundred ns (106, 107).

Even with a coarse-grained force field such as OPEP, it is difficult to form well-ordered structures for six Aβ (16-22) peptides because the number of attractive intermolecular interactions (due essentially to the presence of five hydrophobic amino acids) can stabilize many disordered structures. To reduce this number, we have turned to KFFE, the shortest peptide known to form amyloid fibrils in vitro (102), and studied tetramers (119), hexamers (120), heptamers (121), and octamers (122). As we discuss below, the characterization of these four oligomeric sizes provide a fairly complete picture of the first steps of aggregation for short chains, up to maybe 10 residues.

For simplicity, we focus here on the hexamer of KFFE. As for Aβ (16-22), unbiased ART-OPEP simulations are launched from a random orientation of isolated, unfolded monomers, ensuring that the initial state does not play any role in the formation of specific ordered structures. Since the peptide is very short, it cannot collapse upon itself, leaving the side-groups and H-bonds available for intermolecular interactions. This accelerates the formation of ordered structures. For example, the hexamer assembles into ordered structures within typically less than 20,000 attempted events, fewer than 10,000 accepted events.

All 10 runs locate ordered structures which can be categorized into three generic families, following relatively well-defined aggregation pathways: 1) Two runs lead to a double-layer three-stranded β-sheet (see Figure 10, panel a). 2) Four runs lead to a tetramer-plus-dimer structure with the dimer often positioned in a direction orthogonal to the axis of the tetramer (panel b). 3) Four runs go to a six-stranded β-sheet folding onto itself into an open β-barrel (panel c and d). Although the detailed structure of the KFFE amyloid fibril has not been characterized experimentally, it is interesting to note that structure 1, a double-layer three-stranded β-sheet, displays intra-sheet and inter-sheet distances, respectively 4.5 Å and 10-12 Å (Figure 11), in agreement with the x-ray diffraction measured distances of 4.7 Å between strands and 10-11 Å between sheets in amyloid fibrils (105). While the intra-sheet orientation is mixed, suggesting that more ordered states are possible, these results indicate that
even the hexamer of KFFE is compatible with the amyloid structure. This is not so surprising given the presence of the two phenylalanine amino acids (104).

While ART does not provide any information regarding the dynamics and does not have detailed balance, our experience with the hairpin and the Aβ (16-22) dimer and trimer shows that the aggregation trajectories generated with ART-OPEP are qualitatively representative of what is observed by other methods, when these results are available. For the hexamer of KFFE, these trajectories are most interesting. Figures 12 and 13 show the aggregation trajectory leading to the formation of the double-layer three-stranded β-sheet. Figure 12a follows the energy as a function of the number of accepted events. The energy drops rapidly, within 1000 events, before reaching a plateau (between event numbers 1000 and 2000) where a number of important rearrangements take place, as can be observed in the the relative orientation of a few dimers (panel b). Once all chains are well-stretched, they only have to align themselves to fall into place and the energy drops rapidly between events 2000 and 3000. Because ART continues to search for new local conformations, however, the system is not confined to this low-energy structure and we see that it continues to evolve after reaching the run’s energy minimum. Figure 13 shows the same trajectory through a sequence of cartoons that help picture the aggregation process. In particular, we see that once the two three-stranded β-sheets are formed, there is still some considerable motion of one sheet with respect to the other, including a orthogonal positioning (panel e) before the two fall into place.

Similar intermediates were observed by our group in the simulations of tetramers (119), heptamers (121) and octamers (122) of KFFE, and can be reduced to three dominant ordered structures in addition to the amorphous state: (1) the bi-layer β-sheet; (2) the β-barrel, which protects its hydrophobic residues from contact with the solvent; and (3) the single-layer β-sheet stabilized by a monomer, dimer or more, laying perpendicular to the sheet, such as is observed in Figure 13 (e). These three structures have been observed, all or partially, in other simulations. Stability studies, by Nussinov et al. and other groups, showed that the multi-layer β-sheet is very stable for a large number of sequences (108, 123-126). Many other
Exploring energy landscapes of protein folding and aggregation

Figure 11. Two different views of one double-layer three-stranded β-sheet hexamer of KFFE generated from a random conformation using ART-OPEP. The two layers are perfectly parallel and show a parallel inter-plane organisation with a mixed in-plane orientation. (a) View perpendicular of the fibril axis; (b) view perpendicular to the fibril axis. The dotted lines indicate the position of H-bonds. The N-terminal of each chain is indicated by a larger sphere. The distances indicated are in full agreement with the x-ray diffracted characteristic lengths associated with an amyloid fibril.

Figure 12. Characterization of the self-assembly pathway as a function of accepted event leading to the formation of the double-layer hexamer shown in Figure 16. (a) Evolution of the energy as a function of accepted event. (b) The intersheet orientation, calculated using the scalar product between vectors defined between the two end points of each monomer, for three pairs of chains found in the final structure.

Comparing various simulations, it is clear that although similar ordered structures are found, the probability of these varies strongly with the sequence, but also the force field and the simulation conditions. For example, Favrin et al. observed structure 3, in two forms, in their simulation of Aβ (16-22) using a simplified model (112). Röhrig et al. also find, in simulations of Aβ (16-22) ranging from a dimer to 32-mer, that the tetramer adopts structures similar to structures 1 and 3 identified by ART-OPEP (the single-layer tetramer cannot bend into a barrel, of course) (107).
Exploring energy landscapes of protein folding and aggregation

Figure 13. A generic assembly pathway constructed from the aggregation trajectories observed in the ART-OPEP simulation of an hexamer of KFFE. (a) Starting from six monomers with random orientations, (b) a first dimer forms, with the four other chains in a random conformation, (b) a third chain joins the original dimer, (c) while the three remaining random chains also assemble into a trimer, (d) the two trimers rotate and move around each other, adopting a perpendicular orientation, (e) finally, the two trimers organize themselves into a parallel interstrand arrangement with a mixed β-sheets. Dotted lines represents H-bonds between chains.

Figure 14. Low-energy structure of tetramer of Aβ (11-25). This structure is obtained through an ART-OPEP simulations starting from a well-formed antiparallel dimer and two monomers placed at 30 Å from the dimer. This structure, a three-stranded β-sheet stabilized by a fourth strand laying across the sheet is typical of what is observed in these simulations even though its lies more than 25 kcal/mol above the native state.

example, while the tetramer of Aβ (16-22) has a non-negligible probability to visit a four stranded β-sheet (47), Wu et al. and Colombo et al. could not manage to obtain ordered β-sheet of NFGAIL using MD with AMBER (127) or GROMOS force fields (128).

In spite of the variability in the ART-OPEP trajectories (see also Wei et al. in this issue) seen between different sequences, we start to have a good understanding of the structure of small oligomers formed by short peptides and the paths leading to ordered β-sheet structures. We find that various ordered topologies with β-sheet content are in equilibrium with amorphous aggregates, the early steps of aggregation are dominated essentially by side-chain - side-chain interactions, while the late steps are dominated essentially by H-bonds, through, notably, reptation moves. These findings are consistent with IR spectroscopy (117) and were confirmed numerically by a recent MD study on the dimer of human transthyretin (105-115) peptide (129).

We must now turn to studying larger oligomers of the same short peptides and understanding the oligomeric structure of longer chains, more relevant from a biological point of view. ART-OPEP trials on dodecamers of NFGAIL have led mostly to amorphous structures, even using fast activated methods (130) and new approaches are clearly needed. As for longer chains, they are the topic of the next section.

7. MONOMERS AND DIMERS OF LONGER PEPTIDES

Longer chains of 10 residues or more show a much richer aggregation behavior than the short peptides discussed in the previous section. They are also much more difficult to simulate as the space of conformation becomes very costly to explore in detail for most methods. Activated techniques, such as ART-OPEP, can therefore play an important role in identifying possible intermediates and aggregation pathways, where sampling with standard methods becomes problematic. However, even with an activated approach, the aggregation of long chains remains challenging.

This is clearly observed in the aggregation simulations of a tetramer of the 15-residue fragment Aβ (11-25) which provides an ideal length midway between the short fragments presented in the previous section and the full-length Aβ peptide. Various experiments indicate that this peptide assembles in vitro into amyloid fibril with anti-parallel orientation of the chains (131, 132). ART-OPEP simulations of the monomer (133) suggest, in agreement with experiments on closely related sequence (134), that random coil, α-helices and β-sheets are present in similar proportions except in very polar solvent. Stability ART-OPEP study of the four-stranded β-sheet (131, 132) also fits experimental results. However, ART-OPEP simulations are unable to lead to ordered structures when starting from initial random structures. The monomers come to together and form amorphous aggregates with 30% β-sheet content, at best. Even simulations starting from a preformed dimer fail to reach the ordered anti-parallel tetramer, converging instead onto a well-formed anti-parallel three-stranded beta-sheet stabilized by a fourth chain laying across the β-sheet (see Figure 14). These results indicate that the enhanced flexibility of monomers of about 10 residues or more, compared to that of shorter chains, changes qualitatively the aggregation process. If a tetramer of Aβ (16-22) explores well-formed β-sheets, a longer peptide will favor less-ordered structures, mostly for entropic reasons, as it can adopt many more disordered conformations while protecting its hydrophobic core.

It appears therefore that the current numerical tools are not powerful and rapid enough to simulate, from a random solution, the aggregation process of peptides of 10
Table 1. Interproton distances for Aβ (21-30).

<table>
<thead>
<tr>
<th>ROEs</th>
<th>21α-23N</th>
<th>22α-24N</th>
<th>24α-26N</th>
<th>28α-30N</th>
<th>22α-30N</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC1-ART</td>
<td>3.12</td>
<td>4.04</td>
<td>4.46</td>
<td>4.42</td>
<td>4.77</td>
</tr>
<tr>
<td>SC2-ART</td>
<td>3.31</td>
<td>3.59</td>
<td>4.64</td>
<td>4.94</td>
<td>7.42</td>
</tr>
<tr>
<td>SC3-ART</td>
<td>4.12</td>
<td>5.03</td>
<td>4.69</td>
<td>4.69</td>
<td>4.22</td>
</tr>
<tr>
<td>NMR-1</td>
<td>4.79</td>
<td>3.64</td>
<td>4.63</td>
<td>4.55</td>
<td>5.38</td>
</tr>
<tr>
<td>NMR-2</td>
<td>4.92</td>
<td>3.54</td>
<td>4.33</td>
<td>5.07</td>
<td>4.84</td>
</tr>
<tr>
<td>C1-REMD</td>
<td>4.97</td>
<td>6.36</td>
<td>5.38</td>
<td>4.36</td>
<td>6.53</td>
</tr>
</tbody>
</table>

Weighted-averaged interproton distances for the three families of structures of the fragment Aβ (21-30) identified in ART-OPEP simulations (140) compared to NMR-derived distances (136) and the best structure of all-atom REMD simulations with implicit solvent (138). Distances larger than 5Å, which are considered to violate the constraints, are indicated in bold.

A first test of the approach was performed on the monomer of Aβ (21-30) (AEDVGNKGA), a segment of the amyloid-β peptide which is known to be part of a loop in the fibril and was thought to play a nucleation role in the Aβ aggregation process (135). The interest for this fragment came from a partial proteolysis experiment by Lazo et al., who identified it as protease resistant for both Aβ (1-40) and Aβ (1-42) sequences (136). The high-resolution NMR analysis of the peptide Aβ (21-30) that followed identified two dominant structural families for the 10-residue fragment both characterized by a turn at residues Val24-Lys28 (136). Such a structuration is very rare for 10-amino acid sequences and is a useful test for various simulations methods and potentials. These included standard all-atom explicit solvent MD (137), all-atom REMD simulations with an implicit solvent (138), coarse-grained implicit-solvent discontinuous MD (139) and ART-OPEP simulations (140).

As usual, the ART-OPEP simulations (using the OPEP version 3.0 parameters (43)) where started from a fully stretched conformation and run with a Metropolis accept/reject criterion based on the energy difference between the final and initial minimum for each event. In agreement with the other simulations, clustering analysis revealed that the peptide displays a strong bias towards forming a loop at Val24-Lys28. There are many ways to form such a loop and ART-OPEP identified three dominant clusters, but only one overlapped closely with a structure found in the REMD simulations (138). While ART-OPEP results differ in the details from other simulations, they provide the best agreement of all simulations with NMR distances (140): two of the three structural families identified in the simulation respect the 5 NMR-derived constraints while the third family violates only one (See Table 1). In addition to providing new atomistic models respecting the NMR data, this study also demonstrates that ART-OPEP can also provide quantitative and not only qualitative information on the structural properties of monomers in solution.

The theoretical work on longer peptides is much more diverse and so it is not easy to compare the various methods. There has been a number of MD and REMD studies on the 11-residue peptide Aβ (25-35) (141), and the longer peptides Aβ (10-35) (142-144) and Aβ (12-36) (145, 146). Due to sampling limitations, these works tend to focus on the monomer, providing mostly hints as to the dominant contacts observed on these essentially disordered, but not fully random structures.

In this context, we have launched ART-OPEP simulations on both the monomer and dimer of Aβ (1-28), a peptide that has been extensively studied experimentally (147-152). This peptide has received very little attention numerically (153-155). ART-OPEP simulations starting from a fully stretched state show, in agreement with experiments, that the monomer is more than 50% in
Contact maps for the monomer (top) and dimer (bottom) of Aβ (1-28). In both cases, only the lower-energy structures are considered, i.e. below a standard deviation from the averaged energy sampled.

While results on the monomer are in qualitative agreement with experiment, it is the conformational change from the monomer to the dimer that is most interesting. The bottom panel of Figure 15 presents one of the lowest-energy structures obtained for the dimer. As we can see, dimerization, that leads to 50 kcal/mol gain in energy over two well-relaxed isolated monomers, produces a much more ordered structure. While the dominant contacts remain in the same regions as for the isolated monomer, the two chains are now fully intertwined with a locally anti-parallel organization but a globally parallel alignment.

Figure 16. Contact maps for the monomer (top) and dimer (bottom) of Aβ (1-28). In both cases, only the lower-energy structures are considered, i.e. below a standard deviation from the averaged energy sampled.
range of structures. In some cases, we recover the C-terminal hairpin observed by Skourakis et al. (161) for Aβ42 (see Figure 17), but we also observe a tendency for the N-terminal to assemble into a hairpin, a behavior similar to that observed in the monomer of Aβ (1-28), albeit with a lower probability. If we only consider the lowest-energy state for each of the runs, we see that the monomer of Aβ40 is almost as likely to form β-strands as Aβ42, with a percentage of β-strand, calculated over the lowest-energy state of each of 10 runs, of about 25% for the former, compared to about 28% for the latter. Representative structures for both Aβ40 and Aβ42 are shown in Figure 17. We see that both sequences visit some very similar structures. We also note also that the central hydrophobic core, residues 17-21, marked in red, is very disordered in agreement with NMR study (135). Clearly, the conformations sampled by the monomer do not seem sufficient to explain the kinetic differences observed in the aggregation of Aβ40 and Aβ42 (165).

8. SUMMARY AND PERSPECTIVES

The simulation of protein aggregation is a major challenge that cannot be tackled head on with current computing facilities. It is necessary to approach this problem from multiple angles making various approximations, in order to develop a complete and coherent picture of the aggregation process.

In this review, we have focused on an activated approach, ART nouveau, coupled with the generic OPEP coarse-grained force field. This approach can go beyond qualitative prediction for the monomer of Aβ (21-30) (141). This demonstrates that in spite of its limitations, ART-OPEP can also provide quantitative match to experiments, with structures in full agreement with NMR measurements. Both components were developed for protein folding of non-amyloidogenic sequences and are therefore not biased for following the aggregation of amyloid-forming proteins. ART-OPEP has proven especially useful in identifying the richness in the structure of the small aggregates of short chains, a richness that was confirmed by many other works. Of course, any simulation method is only as good as its two components: the model — determined by the force field — and the sampling method — MD (47), REMD, MC (38) or ART. Since ART and OPEP are independent, both can be used with other methods. OPEP, for example, was also used with MC, MD and REMD. On proteins, ART has also been used with internal coordinates using FLEX and AMBER (166). While ART can work with any derivable potential, it must be used with implicit solvents since it only samples local-energy minima.

Activated simulations on amyloid-forming peptides of various length indicate that there is an important difference in the early steps of aggregation between short (less than 10 amino acids) and long peptides. Results on the monomer and dimer of Aβ (1-28), for example, indicate that their populated structures are very far from the amyloid fibril structure, while the dimer of KFFE, Aβ (16-22) and other related sequences already visit fibril-competent structures, albeit with a probability of 10-20%. Similar conclusions can be drawn for the monomer (this work) and the dimer of Aβ40 and Aβ42 (in preparation). At this point, as was shown in this review, activated methods with unbiased coarse-grained potentials, i.e., not biased towards the formation of amyloid fibrils, are certainly a very promising approach to explore these questions.

9. ACKNOWLEDGEMENTS

The results discussed in this review were obtained in collaboration with a number of people, including G. Boucher, W. Chen, X. Dong, A. Melquiond, S. Santini, and G.-H. Wei. We thank them for their essential contributions to this project over the years. We would also like to acknowledge support from the Alzheimer Society of Canada (2005-2007). N.M. is grateful to NSERC, the Canada Research Chair Foundation and the Fonds québécois de recherche sur la nature et les technologies. P.D. acknowledges financial support from the Centre National de la Recherche Scientifique and the University of Paris 7 Denis Diderot.

10. REFERENCES

4. Ma, B. & R. Nussinov: Simulations as analytical tools to understand protein aggregation and predict amyloid
Exploring energy landscapes of protein folding and aggregation

5. Wei, G., W. Song, P. Derreumax & N. Mousseau: Self-assembly of amyloid-forming peptides using molecular dynamics simulations and the OPEP coarse-grained force field *Frontiers in biosciences*, this issue,
Exploring energy landscapes of protein folding and aggregation


4513
Exploring energy landscapes of protein folding and aggregation

Exploring energy landscapes of protein folding and aggregation


Exploring energy landscapes of protein folding and aggregation


Key Words: Amyloids, Activation-Relaxation Technique, OPEP, Molecular Dynamics, Protein Dynamics, Beta-Amyloid, Alzheimer, Computer Simulation, Beta-Hairpin, Coarse-Grained Potential, Review

Send correspondence to: Dr Normand Moussseau, Departement de physique, Universite de Montreal, c.p. 6128, succ. centre-ville, Montreal (Quebec), Canada, H3C 3J7, Tel: 514-343-6614, Fax: 514-343-02071, E-mail normand.moussseau@umontreal.ca

http://www.bioscience.org/current/vol13.htm