Published in final edited form as:
J Mol Biol. 2011 January 14; 405(2): 570–583. doi:10.1016/j.jmb.2010.10.015.

Atomic-level characterization of the ensemble of the Aβ(1-42) monomer in water using unbiased Molecular Dynamics simulations and spectral algorithms

Nikolaos G. Sgourakis[2,4,5], **Myrna Merced-Serrano**[6], **Christos Boutsidis**[3,5], **Petros Drineas**[3,5], **Zheming Du**[1,4,5], **Chunyu Wang**[1,4,5], and **Angel E. Garcia**[2,4,5]

[1] Department of Biology, 110 8th Street, Troy, NY 12180, USA

[2] Department of Physics, Applied Physics and Astronomy, 110 8th Street, Troy, NY 12180, USA

[3] Department of Computer Science, 110 8th Street, Troy, NY 12180, USA

[4] Center for Biotechnology and Interdisciplinary Studies, 110 8th Street, Troy, NY 12180, USA

[5]Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA

[6] Department of Mathematics, University of Puerto Rico at Humacao, Humacao, PR 00791

Abstract

Aβ(1-42) is the highly pathologic isoform of amyloid-β, the peptide constituent of fibrils and neurotoxic oligomers involved in Alzheimer's disease. Recent studies on the structural features of Aβ in water have suggested that the system can be described as an ensemble of distinct conformational species in fast exchange. Here, we use replica exchange molecular dynamics simulations (REMD) to characterize the conformations accessible to Aβ42 in explicit water solvent, under the ff99SB forcefield. Monitoring the correlation between J-coupling ($^3J_{H^N H^\alpha}$) and residual dipolar coupling (RDC) data calculated from the REMD trajectories to their experimental values, as determined by NMR indicates that the simulations are converging towards sampling an ensemble that is representative of the experimental data after 60ns/replica of simulation time. We further validate the converged MD-derived ensemble through direct comparison with $^3J_{H^N H^\alpha}$ and RDC experimental data. Our analysis indicates that the ff99SB-derived REMD ensemble can reproduce the experimental J-coupling values with high accuracy and further provide good agreement with the RDC data. Our results indicate that the peptide is sampling a highly diverse range of conformations: by implementing statistical learning techniques (Laplacian Eigenmaps, Spectral Clustering, and Laplacian Scores) we are able to obtain an otherwise hidden structure in the complex conformational space of the peptide. Using these methods we characterize the peptide conformations and extract their intrinsic characteristics, identify a small number of different conformations that characterize the whole ensemble, and identify a small number of protein interactions (such as contacts between the peptide termini) that are the most discriminative of the different conformations and thus can be used in designing experimental probes of transitions between such molecular states. This is a study of an important intrinsically disordered peptide

To whom correspondence should be addressed. angel@rpi.edu Tel. 518-276-6310 Fax. 518-276-6680.

system that provides an atomic-level description of structural features and interactions that are relevant during the early stages of the oligomerization and fibril nucleation pathways.

Keywords

molecular dynamics simulations; amyloid; peptide; Alzheimer's disease; Abeta monomers; ff99sb forcefield; conformational clustering; spectral clustering; normalized cuts; Laplacian eigenmaps; laplacian scores; contact maps; J-coupling constants; residual dipolar couplings; RDCs; NMR

Introduction

The amyloid-β (Aβ) peptides are the major constituents of amyloid plaques, the pathological hallmark of Alzheimer's disease (AD) and neurodegeneration in general 1. Aggregation of Aβ leads to various β-sheet rich conformers that are found in the brains of AD patients and correlate with the onset of AD 2. Moreover, Aβ oligomerization leads to the formation of soluble, neurotoxic oligomeric species that impair synapse transmission and eventually memory function 3; 4. Both the amyloidogenic and oligomeric pathways originate in the cell membrane: the different-length isoforms of Aβ are derived from the proteolytic processing of a transmembrane protein, the amyloid precursor protein (APP). Variability in the exact site of APP cleavage leads to the production of Aβ isoforms of different lengths (ranging from 39 to 42 residues), of which Aβ42 is a major isoform and has a high potential to elicit amyloigonesis and toxicity.

Despite significant advances in the structure determination of Aβ fibrils 5; 6 and their polymorphisms 7 in atomic detail, few studies have been performed to characterize the ensemble of the full-length Aβ(1-42) peptide at the monomeric level in water. An NMR-derived model of the average structure of the 26mer Aβ(10-35) in water has revealed a collapsed coil with little presence of regular secondary structural elements 8. However, several experimental and computational studies focusing on different fragments of Aβ and its mutants 9; 10 have indicated a highly dynamic, rugged energy landscape that is consistent with an ensemble of rapidly interconverting, iso-energetic (to a first approximation) conformational species in fast exchange 11; 12. Previous experimental results have suggested that the peptide displays structural features that deviate significantly from the random coil model indicated by local conformational preferences of the backbone 13; 14; 15. In a previous study, we used MD simulations validated by experimental NMR data to elucidate the conformations accessible to both *in vivo* isoforms of Aβ, A40 and 42 [16]. Our MD-derived molecular ensemble suggested that both peptides displayed unique structural features that were consistent with the experimentally measured J-coupling data. Moreover the mechanism of aggregation and the energetics of the transitions between monomers, oligomers and fibrils are yet to be characterized in atomic detail. Recent efforts to characterize the structure of important intermediates along the aggregation pathway including neurotoxic oligomeric species have resulted in the solution structure of a soluble Aβ oligomer by NMR [17]. To this extent, a detailed view of the solution conformation of Aβ at the monomer level and their dynamics is important towards modeling the aggregation pathways, as well as in rationally designing therapeutics that would selectively stabilize non-amyloidogenic conformations [18; 19] and inhibit oligomers and fibril formation [20].

Here we present a detailed characterization of the ensemble of Aβ42 that is obtained by all-atom molecular dynamics simulations in explicit solvent. We implement the same enhanced sampling protocols used previously16 that were extended to the μsec simulation timescale and used a recently improved forcefield 21 derived from the AMBER series of molecular mechanics forcefields [22]. Our simulation data are validated by direct comparison with three

bond J-coupling constants and residual dipolar couplings (RDCs), as measured experimentally by NMR for the backbone NH groups. These experimental observables, through their intrinsic dependence on the average backbone conformation and orientation relative to a molecular alignment frame respectively, provide a sensitive probe of molecular structure and have been recently used to model the conformations of unfolded, intrinsically disordered and chemically denatured proteins using biased ensemble-based approaches [23; 24; 25]. In addition, RDCs have been previously measured for both major isoforms of Aβ and interpreted on the basis of statistical coil models [26; 27]. Analysis of our unbiased REMD structural ensemble reveals the presence of distinct conformational species, which we identify and further analyze to obtain a small number of representative conformations. Our results indicate the presence of a highly diverse conformational ensemble that can be analyzed in terms of correlated patterns of interacting residues to yield conformational species of distinct structural features. To analyze the structural properties of the ensemble, we port non-trivial techniques from statistical learning. More specifically we are using the Laplacian eigenmaps approach 28 to visualize the conformations in a low-dimensional space, while the spectral clustering technique 29 is used to efficiently extract conformations that are representative of the ensemble. Finally, using Laplacian scores [30], we identify interactions (such as contacts between the peptide termini) that are highly effective in distinguishing between distinct conformational basins and can be thus used to design experimental labels that report on the transitions between these conformational species. This study augments on the existing knowledge of the conformations accessible to Aβ(1-42) monomers in water and further indicates a strategy to effectively identify key structural features and classify diverse ensembles of such conformations from MD simulation data for metastable and intrinsically disordered systems in general.

Results

Convergence of the REMD simulations

We have estimated the time it takes for the simulations to converge, according to the selected observables by monitoring the agreement of results calculated from our MD simulations to their experimentally determined values. To perform this task, in addition to ${}^{3}J_{H^{N}H^{\alpha}}$ J-couplings we have monitored the correlation of residual dipolar couplings to the experimental results [27]. The two observables give very similar results (figure 1a,b). We observe a first phase in the simulations (0-60ns /replica) where the correlation to experiments changes rapidly after which the simulations converge to showing only larger large-timescale fluctuations that are more pronounced for the RDCs. After roughly 60 ns/ replica, the two observables reach Pearson's correlation coefficients (P.C.C.) of approximately 0.4 - 0.5 (figure 1). These values are strikingly similar to the ones obtained previously [16] for the same system using the OPLS/AA forcefield [31] and TIP3p water model [32] (60ns/replica and P.C.C. of 0.48 according to ${}^{3}J_{H^{N}H^{\alpha}}$), indicating the robustness of the replica exchange algorithm for this intrinsically disordered system. The differences in calculated J-coupling and RDC values among three samples of equal length obtained from the production phase of our simulation trajectory (60-225ns/replica) are less that 10% in magnitude, indicating that the simulation is indeed well converged to sampling an ensemble that is representative of the ff99SB forcefield after this first equilibration phase.

Validation with experimental results

We have examined the use of different parameter sets for the Karplus equation in calculating J-coupling constants from the simulation coordinates for comparison with their experimental measurements, as described in methods [33; 34]. In general, we observe a correlation between the experimental and simulation dataset that is comparable to the correlation between the

two experimental datasets, as indicated by RMSD values below 1Hz (0.32Hz vs 0.73Hz) (figure 2). This agreement with the experimental J-coupling data is comparable to results recently obtained using the same forcefield for stable protein folds [35]. Nevertheless, we also observe several outliers for which the calculated values are not in agreement with the experimental results. In particular, for residues Glu 22, Asp 23 and His 13, the calculated values differ by more than 1Hz from both experimental datasets. For single conformations, this may amount to differences in the φ dihedral angle as small as 5-7°, or as large as 74° for selected values of $^{3}J_{H^{N}H^{\alpha}}$, by virtue of the fact that the Karplus equation is not a 1-to-1 function of φ. Analysis of the values of φ for these three residues in our simulation dataset indicates that all three allowed basins of the Ramachandran plot are being sampled in our REMD trajectories, with different population weights that are given by the relative free energies under the current forcefield (figure 2 in supplementary material). Therefore, the discrepancy with the experimental results could be attributed to incorrect weighting of the different basins. All three of these outlier residues are charged at neutral pH modeled in this simulation study. This finding suggests a potential strategy for improvement of the ff99sb forcefield that takes into account the interplay between the backbone dihedrals and charged sites.

We further validated our converged MD dataset (according to the J-coupling convergence shown in figure 1b) by comparison of calculated residual dipolar couplings with previously published experimental data [27]. RDCs report on the alignment of the amide bond vectors relative to a conformation-specific molecular alignment frame that was calculated based on the steric properties of each conformation using the method PALES [36]. The comparison between the experimental and calculated RDCs is shown in figure 2c, d. In general, good agreement between the two datasets is obtained (RMSD 1.49Hz), which is reflected in a Pearson's correlation of 0.30 to the experimental data. We observe three outliers for which the disagreement with experimental data is larger than twice the sum of the experimental and simulation errors. These are residues Phe 20, Val 36 and Ala 42. In contrast, for the remaining 19 residues of Aβ42 for which accurate experimental values were available, the agreement with the experimental data is significantly better, reaching a Pearson's correlation of 0.57. As the calculated RDCs report both on the alignment properties of the molecule and the orientation of individual amide bond vectors, the fact that the disagreement with the experimental data is limited to residues 20, 36 and 42 indicate that the forcefield and solvation model used is likely sampling the correct shape distribution of Aβ42, and the inconsistencies are due to local deviations in the backbone torsion angles.

Identification and Characterization of representative conformational species

We have implemented a spectral clustering approach to characterize the conformations sampled in the REMD ensemble (summarized in figure 6). In the case of the intrinsically disordered Aβ peptide, which samples a diverse range of conformations, common clustering strategies that rely on the calculation of geometrical distances such as RMSD are limited by the drastic change in the shape of the molecule across the conformational space that is accessible. In summary, this approach is based on the representation of each protein conformation as a contact map (figure 6a). This representation is then used in the calculation of a square affinity matrix *A*, whose elements are defined for each pair of conformations according to a distance kernel. The spectral clustering technique involves the diagonalization of this matrix to obtain singular values of high discriminative power in distinguishing between different points using a small number of linearly independent dimensions (figure 6b). This information is encoded within a small number of the lowest non-trivial eigenvectors and can be used to cluster the ensemble into groups of conformations that share common contact map patterns. A direct visualization of the REMD conformational ensemble in the space defined by the lowest 3 non-trivial eigenvectors is

shown in figure 3a. We observe good dispersion of the REMD data in the three eigenvectors. In general we observe a high degree of structural similarity for conformations that are near and different overall structural features for conformations belonging to different regions of the space (figure 3), which indicates the high discriminative power of this technique. Consequently, conformations belonging to the spatially distinct clusters display common structural features, as exemplified in figure 4. Furthermore, we observe frequent transitions between the clusters throughout the REMD trajectories (figure S3), which further supports the conclusion that we are sampling a structurally converged ensemble of conformations, as previously indicated by the convergence of the ensemble-averaged J-coupling and RDC data shown in figure 1.

The identified representative conformations illustrate a diversity of local structural features including regions with elements of regular secondary structure that were assigned using the DSSP algorithm [37]. In particular, we frequently observe the formation of β-sheets involving interactions of the C-terminus with other parts of the sequence, as shown in figure 5. In a conformationally distinct cluster, we observe the formation of a β-sheet involving strands at residues 4-6 and 38-40 (sequence GVV), as well as an α-helix at the sequence ^{8}SGYE^{12}V (figure 5a). Alternatively the stand at residues 38-40 can form a β-hairpin involving residues 33-35, as seen in another representative conformation (figure 5b) or a sheet with residues 18-20, as seen in a separate closely clustered group of conformations (figure 5c). This indicates that the region 38-40 may act as a conformational switch, whose interactions with various parts of the sequence dictate the conformational state of the peptide. A similar conformation for the C-terminus has been previously reported in results using the OPLS/AA forcefield [31], where it was found to form a β-sheet with residues 31-34. In a separate closely clustered group of conformations we observe the formation of a β-hairpin involving short strands at residues ^{3}EF^{5}R and ^{10}YE^{12}V towards the N-terminus of the sequence (figure 5d). In addition, a short α-helix spanning residues 20-23 can be seen. In the same cluster, a hydrogen bond of Arg 5 with Ser 26 is also observed. In a conformationally distinct cluster we observe the formation of a 3_{10} helix at the sequence ^{29}GAII^{33}G (figure 5e). This region has a high potential to form a 3_{10} helix, as also observed in other clusters (figure 5c). Finally, in a separate cluster that represents a large part of the population, the conformation of the peptide resembles a coiled-coil (figure 5f). A brief 3_{10} helix is observed at residues 21-24. The presence of turn-like structures at residues Y10, F19 and F20, as observed in some of the clusters, is corroborated by the analysis of RDC measurements in stretched polyacrylamide gels previously reported by Lim and coworkers [26].

Identification of interactions with high discriminative power

We have further explored the use of different interactions in Aβ to discriminate among distinct conformations in the ensemble. For this purpose, we have implemented the Laplacian scores technique [30]. This technique can be used to extract features that are optimal in describing the local structure of points in a dataset. In our case, the features correspond to residue contacts derived from the contact map representation of the REMD conformations described previously. An inspection of the 2D Laplacian score map relative to the raw probabilities of contact formation for different pairs of residues in Aβ indicates these regions of high discriminative power (figure 7). These interactions are formed between the N-terminal residues 2-5 with residues 24-26 or with residues 34-40 and between residues 25-28 with residues 36-40. All three regions have a relatively small probability of contact formation in the ensemble and would not be identified on the basis of the contact probabilities alone. However, the Laplacian score analysis suggests interesting features. One such region is for contacts between the N- and C-termini of the peptide (shown in the upper left part of figure 6). These regions have been observed to interact through a β-sheet in one of the representative conformations in the ensemble (figure 3) involving a strand at residues

38-40. Therefore, the Laplacian score in this case can be attributed to the formation of this long-range structure. In addition, high Laplacian scores are obtained for interactions between the C-terminus (residues 36-40) with residues 23-28, which have also been found to form β-sheets as well as α-helices in different conformations in the ensemble. This result is consistent with the picture obtained from the analysis of representative conformations, where the β-strand at position 38-40 was found to interact with several alternative partners, thus indicating the high discriminative power of this region to distinguish between different cluster conformations. Finally, for interactions between residues 2-5 with residues 24-27, we identify another region of contacts with high discriminative power, which again highlights the importance of long-range interactions in shaping the energy landscape of Aβ. Therefore, monitoring the state of these key contacts would be highly informative for the overall conformation of the peptide.

Discussion

We have performed REMD simulations using the ff99SB forcefield [21] for the full-length Aβ(1-42) monomer which constitutes a characteristic intrinsically disordered peptide system. Previous studies have indicated the merits [21; 38; 39; 40; 41; 42] and limitations [35; 43] of using this forcefield to obtain realistic ensembles of short peptides and proteins, relative to a variety of NMR data that report on both the average structural properties and dynamics of biomolecules. In two recently published studies Wickstrom and coworkers have shown that this forcefield, when used in combination with the TIP4P-Ew explicit solvation model 44, can reproduce experimental backbone J-couplings with reasonable accuracy for short Alanine polypeptides 42 and the chemically denatured state of the vilin headpiece 45. Notably, using this combination of forcefield and solvation model, Fawzi and coworkers [40] performed multiple microcanonical simulations for a smaller fragment of Aβ(21-30) that reproduced both J-coupling constants as well as ROEs and ^{13}C relaxation rates measured by NMR. Day and coworkers studied the unbiased folding/unfolding thermodynamics of the trp-cage miniprotein and found the ff99SB forcefield produced folded ensembles with distributions centered at 0.6Å RMSD from the NMR structure and a folding temperature that is comparable to the one that is determined experimentally for this system [38]. However, Lindorff-Larsen and coworkers observed significant deviations in the distributions of sidechain rotameric states in extensive ff99SB simulations, relative to statistics obtained from the protein data bank that influence the calculation of accurate J-coupling values in proteins [35]. Taken together, these results indicate the major areas of improvement towards the next generation of AMBER forcefields.

Here, we confirm these findings and further explore the generality of MD results obtained using ff99SB under extensive sampling conditions for an intrinsically disordered peptide. We have obtained a converged ensemble, from the point of view of correlation to experimental J-coupling and RDC values, after 60ns/replica. The convergence of this ensemble at the structural level is further confirmed by the observed global sampling of the accessible conformational space according to the time history of the assigned clusters (figure S3). Validation with both J-couplings and RDCs indicates good agreement with experiments for most sites for which experimental data were available, which is manifested in Pearson's correlation coefficient in the range of 0.4-0.5 for both observables. In a previous study for the same peptide system using a variety of forcefields we have found that the best-performing forcefield, OPLS/AA 31, reached a PCC of 0.48 to the same experimental J-coupling dataset [16], indicating a small improvement for the ff99SB forcefield, which is within the simulation error. Moreover, we observe similar convergence times (approx. 60ns/replica) for the replica exchange algorithm for both forcefields, which suggests the simulation time needed to obtain realistic ensembles of intrinsically disordered peptides is in this range, which is a promising result given the combinatorial explosion of the

conformational space that is accessible to systems of size comparable to Aβ42 and are of great medical and biological importance. Finally, the reported correlation to the experimental data is, in most cases, comparable if not higher to the one obtained using μsec-timescale simulations of stable-folded proteins using a recently proposed improvement of the ff99SB forcefield used here [35].

We demonstrate the high discriminative power of the spectral clustering method used here in identifying representative conformations towards a detailed characterization of the highly diverse ensemble of Aβ42. To date, several dimensionality reduction techniques have been employed to study biomolecular dynamics from MD simulation data, for the purposes of clustering, the identification of representative conformations, or transitions between distinct conformational states [46; 47; 48; 49]. The spectral clustering technique implemented here, although previously applied for the clustering of protein sequences [50] has not been previously used to address the problem of classifying conformational ensembles from MD simulation data. In this study, we show that this technique is highly efficient in deriving families of conformations that share distinct intramolecular interaction patterns, as shown in figures 3-5. Analysis of our simulation data using this method suggests that Aβ42 samples a highly diverse conformational ensemble that can be analyzed on the basis of a relatively small number of collective variables that report on medium to long range intramolecular interactions. A similar approach has been recently used by our group using MD simulation data to identify and characterize distinct intermolecular orientations in the Rhodopsin/Transducin complex [51]. Without loss of generality, this approach can be implemented for the visualization and clustering of conformations obtained via other computational and experimental methods such as results from *ab initio* protein folding calculations and protein structure calculations.

When compared with commonly employed clustering algorithms 52 that are based on the Root-mean square distance (RMSD) kernel 53, our method offers some attractive features. The main drawback of RMSD-based clustering methods is that conformations that are far apart in RMSD space will be classified in different clusters regardless of their contact map similarity. In systems with conformational flexibility this may result in a very large number of clusters that are hard to interpret manually. Our method overcomes this problem by looking for structures that may be far apart in RMSD but share common interactions. Furthermore, results using hierarchical RMSD-based algorithms in particular are highly dependent on the choice of the RMSD cutoff used in the clustering, a parameter that needs to be optimized in order to obtain meaningful results (see procedure in 16). We have repeated the clustering using the Daura algorithm 54 on our dataset of 11,570 Aβ conformations using a 0.2nm cutoff for the definition of neighbor lists (same parameters as in 54). In general, we obtain several small clusters (2,170) the majority of which have very small sizes. The six largest clusters have populations in the range 2-6%. Looking at the central conformations of the clusters we see a diverse group of structures, as expected. The conformations of the largest cluster are very similar to those obtained using our method in cluster 6 (figure 4). This confirms that structures that are close in RMSD space can also be close the space defined by the contact map definition. However, the opposite is not necessarily true.

Finally, we have made use of Laplacian scores [30] to identify pairwise residue interactions that can be used to discriminate between different conformational species, thus opening the possibility of designing experimental labels to study transitions among such conformations. The identified contacts, although not observed in the REMD ensemble with high probability, show significant differences (on average) between different families of conformations (figure 7). This analysis indicates that the short β-strand at residues 38-40 may act as a conformational switch whose alternative interactions with other strands along the sequence

of Aβ determine the conformational state of the peptide. To this extent, the REMD ensemble can provide valuable predictions to experimentalists towards the study of transitions between different conformations with distinct aggregation and oligomerization properties. If experimentally verified, this information is valuable in designing strategies to block transitions that lead to pathogenic conformations, thus suggesting a novel approach in Alzheimer's disease treatment at the molecular level.

Methods

Replica Exchange Molecular Dynamics (REMD) simulations

Molecular Dynamics simulations were performed using the Replica Exchange Molecular Dynamics algorithm. The REMD is a generalized ensemble method 55; 56 that involves several identical copies of the system, or replicas, that are simulated in parallel over a range of temperatures. At frequent intervals, trials to exchange the temperature of all adjacent replicas are performed, according to a Metropolis Monte Carlo criterion. To optimize the temperature spacing of the replicas, we performed 16 pilot constant temperature (and volume) simulations for 3ns each spanning different temperatures in the range 250-600K. The histograms of potential energy obtained from these short trajectories were then used to define the temperatures of the replicas, such that the average exchange ratio is constant throughout the temperature space and equal to 15%, according to the algorithm described previously 57. The range of temperatures used in the final REMD simulations was from 270.0 to 601.2K. A total of 52 replicas were used to optimally span the temperature space. Exchange moves in temperature were attempted every 4ps between all adjacent replicas in temperature space. A detailed structural analysis was performed only on conformations sampled by all replicas at 7 temperatures in the range 289-311K. For all calculations we used the FF99SB forcefield 21 in combinations with the TIP4P-Ew water model 44. Previous calculations focusing on Aβ(10-35) by Fawzi and coworkers 40 have shown that this combination of forcefield and water model produces an ensemble of configurations that is in good agreement with NMR data.

To build the peptide system, we started from a completely extended conformation of the full-length Aβ(1-42) peptide with sequence:

^{1}DAEFRHDSG^{10}YEVHHQKLVF^{20}FAEDVGSNKG^{30}AIIGLMVGGV^{40}VIA.

The following procedure was used to construct the system: First, we run a 1ns MD simulation of the peptide *in vacuo*, at high temperature (~700K) starting from a completely extended conformation, followed by an energy minimization of the system. The collapsed peptide was then solvated in a cubic box, whose dimensions were adjusted to accommodate 4,947 water molecules (total system size 20,415 interaction sites). We chose a system size that reduces short-range interactions between periodic images of the peptide and is computationally tractable. The solvated system was then equilibrated at constant temperature (300K) and pressure (1 atm) for 1 ns with a short integration time step of 1fs. This resulted in a cubic simulation box of side length 53Å in each dimension. Finally, REMD simulations at constant volume were run for 225 ns / replica in total (aggregate simulation time of 11.7μsec). At this stage, the application of the LINCS 58 and SETTLE 59 algorithms to constrain the bond lengths in the peptides and water molecules respectively allowed a relatively large integration step of 2 fs. We used a cutoff of 10Å for the evaluation of Lenard-Jones interactions, while pair lists were updated every 10 integration steps. We used the particle mesh Ewald method 60 with a 52×52×52Å cubic grid to evaluate long-range electrostatics. Charge neutrality of the system is implicitly treated by the used of the Ewald method for the computation of long-range electrostatics. This is closely mimicking the NMR experimental conditions, where the sample salt concentration was kept minimal

(20 mM potassium phosphate buffer with no other salt) 27. Ions, especially ones of the cationic series, have been shown to play important roles in the aggregation and fibril morphology of Aβ 61; 62, however this is a condition that was not explored in the current study.

The system was coupled to a Nose-Hoover [63] heat bath to maintain a constant temperature between exchanges. All simulations were performed at 204 CPUs of Linux-based clusters at Rensselaer, with the use of the GROMACS [64] simulation machine under a variety of domain decomposition schemes.

Comparison with experimental data

A variety of experimental data were used for the purposed of a) Assessment of simulation convergence and b) validation of the MD-derived conformational ensemble. In a manner similar to the approach used previously [16], we have monitored the correlation with experimental three-bond J-couplings as an indicator of convergence and as a measure of validity. The correlation between two datasets X, Y is quantified in terms of the Pearson's correlation coefficient:

$$P.C.C = \frac{Cov(X,Y)}{\sigma^2(X)^* \sigma^2(Y)}$$

Where $Cov(X,Y)$ is the covariance of the two variables and $\sigma^2(X)$, $\sigma^2(Y)$ are the corresponding standard deviations. In order to assess the reproducibility of the J-coupling data we used two independent datasets of measured ${}^3J_{H^N H^\alpha}$ data to compare with simulations, the first published in a previous study by our group 16, while the second was collected under identical sample conditions and the experimental protocols described by Yan and coworkers [27]. The two measurements of J-coupling constants were very similar for most residues (Pearson's correlation coefficient of 0.92), with the exception of Glu 11, which was found to differ significantly between the two experimental datasets. In total, 21experimental ${}^3J_{H^N H^\alpha}$ values were used of which 17 were redundant between the two datasets. We used the Karplus equation to predict J-coupling constants from our MD coordinates [65]:

$$J = a\cos^2(\theta) + b\cos(\theta) + b,$$

where a, b, c are semi-empirically derived coefficients and $\theta = \phi - 60$, where φ is the peptide dihedral angle. The use of various published datasets of Karplus coefficients was explored. We used coefficients previously determined by fitting to X-ray structures [34] , as well as a modified dataset that accounts for dynamics within a single harmonic well [33]. Finally, motional averaging effects within our MD dataset were explicitly taken into account by fitting the Karplus coefficients to the experimental data for Aβ. This resulted in a set of coefficients that optimally describes our data and is within previously published values, as reported by Bruschweiler and Case [33]. The fitted values were determined to be $a = 7.7, b = -1.9$ and $c = 0.06$, introducing a marginal change to a, b and a significant decrease in c relative to previously published values (reviewed in 33). Using this set of fitted parameters, the root-mean square deviation (RMSD) from the experimental data was reduced to 0.73Hz from 1.46Hz, for J-couplings calculated using the coefficients published by Vuister and Bax 34. When using the corrected coefficients 33 the RMSD was 0.96Hz (figure 1 in supplementary material). For all calculations we used the final 165ns/replica of our simulation trajectories, sampled every 100ps. The MD dataset was split into three samples of 3,857 conformations each, which were used to estimate the error in the calculated values.

In addition, we used Residual Dipolar Couplings (RDCs) measured in 10% polyacrylamide gels as an additional, independent measure of the validity of our simulations. Experimental RDCs were obtained from a previously published study for 30 amides in Aβ(1-42) under partial alignments conditions at 273.3K [27]. From this dataset, we extracted 22 RDCs for which the experimental error was less than 33%. The method PALES 36; 66 was used for the calculation of RDC values from the MD data. In summary, for each conformation in our MD ensemble the program calculates an alignment orientation due to steric properties of the molecule, which is subsequently used to calculate RDC values. This is done by diagonalization of the moment of inertia tensor. Finally, ensemble-averaged RDCs are computed according to the equation:

$$D_{ij}=\frac{-\mu_0\gamma_i\gamma_j h}{2\pi}\left\langle\frac{3\cos^2\vartheta_{ij}-1}{2r^3}\right\rangle=D_{\max}\sum_{k,l}\left\langle\frac{s_{kl}}{r^3}\cos\theta_{ij}^{k}\cos_{ij}^{l}\right\rangle,$$

where μ_0 is the permeability of empty space,γ_i,γ_j are the gyromagnetic ratios of the i and j nuclei, h is Planck's constant, r^3 is the length of the internuclear vector and ϑ_{ij} is the angle between the internuclear vector and the external magnetic field. Expressed in the molecular frame, all constants are absorbed in $D_{\max}$, which is the maximum possible value of the RDC for a particular nuclei pair, s_{kl} is a component of order tensor describing the alignment of the protein in the laboratory frame and $\cos_{ij}^{k}$, $\cos_{ij}^{l}$ are the orientation cosines of the internuclear bond vector in the molecular alignment frame. Ensemble-averaged RDC values were uniformly scaled to minimize the RMSD form the experimental data, due to the fact that the alignment tensor magnitude depends on the fraction of molecules that are in the aligned stated that depends on the experimental conditions.

A contact map-based representation of the configurational ensemble

We have used a numerical representation of the protein conformations of our MD simulations (11,564 in total). In our approach, every protein conformation is represented as a 2-dimensional table that we refer to as "contact map table" in the sequel. In principle, we represent a conformation as a binary table of residue-to-residue interactions. We focus on 42 residues and on interactions within a distance threshold of 4.5Å between any pair of heavy atoms in the residue. Each of the 11,564 contact map tables has dimensions 42×42, where for all i,j = 1,…,42 the (i,j)-th element of the table indicates the presence or absence of an interaction (contact) between the i-th and the j-th residues of the protein that corresponds to this particular table. We fill all elements of this table with zeros and ones such that a '0' implies a broken contact and a '1' implies a formed contact. We further simplify this contact map table by neglecting trivial short-range interactions between residues with less than 3 sequence separation; i.e. the main diagonal as well as the three sub-diagonals around the main diagonal of every contact map table are neglected. To organize the contact map tables in a more compact way, we first transform each of them to an 1-dimensional row vector. This vector has 741 dimensions, since we discarded the central diagonals kept only half of it due to its symmetry. That way, every element of this 741-dimensional vector corresponds to a unique residue-to-residue interaction in the protein. Finally, our MD ensemble can be represented as a 11,564×741 matrix, where each row corresponds to a snapshot (the vectorized contact map described above) and each column corresponds to a residue-to-residue interaction. This binary matrix is denoted as A in the sequel.

Spectral Analysis of Protein Conformations

The goals of our computational are three-fold:

1. Visualization: we want to visualize the conformations in a small number of dimensions such that to be able to quickly understand the hidden structure of the complex conformation space.
2. Clustering: identification of a small number of "representative" conformations: we want to find a small subset of conformations that efficiently summarize and characterize the protein ensemble.
3. Feature selection: identification of a small number of "representative" residue-to-residue interactions: we want to find a small subset of residue-to-residue interactions with high "discriminative power", i.e interactions that suffice to classify the conformations into different groups.

The above goals are achieved by employing techniques typically referred to as "spectral algorithms"; this characterization implies algorithms that use eigenvectors and eigenvalues of appropriate matrices [67]. All of our techniques employ an eigenvalue type analysis of the Laplacian matrix of a proper graph describing the matrix A. In more details we use:

1. The Laplacian Eigenmaps approach [28] to visualize the conformations in a three dimensional Euclidian space. This is described in Appendix A1.
2. The spectral clustering approach based on normalized cuts [29] to cluster the conformations into different groups and select representatives within each group. This technique is described in Appendix B.
3. The feature selection approach based on the Laplacian Scores [30] to identify contacts with high discriminative power (see Appendix C).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Drs. Paul Maragakis, Kresten Lindorff-Larsen and Xavier Salvatella for useful discussions. This work is supported by the NIH Molecular Libraries Roadmap Initiative (IP20HG003899-01), NSF (MCB0543769 and DMR0117792) and IBM.

References

1. Soto C. Unfolding the role of protein misfolding in neurodegenerative diseases. Nat Rev Neurosci. 2003; 4:49–60. [PubMed: 12511861]
2. Rauk A. The chemistry of Alzheimer's disease. Chemical Society Reviews. 2009; 38:2698–2715. [PubMed: 19690748]
3. Lesne S, Koh MT, Kotilinek L, Kayed R, Glabe CG, Yang A, Gallagher M, Ashe KH. A specific amyloid-beta protein assembly in the brain impairs memory. Nature. 2006; 440:352–7. [PubMed: 16541076]
4. Mucke L, Masliah E, Yu GQ, Mallory M, Rockenstein EM, Tatsuno G, Hu K, Kholodenko D, Johnson-Wood K, McConlogue L. High-level neuronal expression of abeta 1-42 in wild-type human amyloid protein precursor transgenic mice: synaptotoxicity without plaque formation. J Neurosci. 2000; 20:4050–8. [PubMed: 10818140]
5. Luhrs T, Ritter C, Adrian M, Riek-Loher D, Bohrmann B, Dobeli H, Schubert D, Riek R. 3D structure of Alzheimer's amyloid-beta(1-42) fibrils. Proc Natl Acad Sci U S A. 2005; 102:17342–7. [PubMed: 16293696]
6. Petkova AT, Ishii Y, Balbach JJ, Antzutkin ON, Leapman RD, Delaglio F, Tycko R. A structural model for Alzheimer's beta - amyloid fibrils based on experimental constraints from solid state NMR. Proc Natl Acad Sci U S A. 2002; 99:16742–7. [PubMed: 12481027]

7. Paravastu AK, Qahwash I, Leapman RD, Meredith SC, Tycko R. Seeded growth of beta-amyloid fibrils from Alzheimer's brain-derived fibrils produces a distinct fibril structure. Proc Natl Acad Sci U S A. 2009; 106:7443–8. [PubMed: 19376973]
8. Zhang S, Iwata K, Lachenmann MJ, Peng JW, Li S, Stimson ER, Lu Y, Felix AM, Maggio JE, Lee JP. The Alzheimer's peptide a beta adopts a collapsed coil structure in water. J Struct Biol. 2000; 130:130–41. [PubMed: 10940221]
9. Baumketner A, Krone MG, Shea JE. Role of the familial Dutch mutation E22Q in the folding and aggregation of the 15-28 fragment of the Alzheimer amyloid-beta protein. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:6027–6032. [PubMed: 18408165]
10. Wu C, Murray MM, Bernstein SL, Condron MM, Bitan G, Shea JE, Bowers MT. The Structure of A beta 42 C-Terminal Fragments Probed by a Combined Experimental and Theoretical Study. Journal of Molecular Biology. 2009; 387:492–501. [PubMed: 19356595]
11. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. Science. 1991; 254:1598–603. [PubMed: 1749933]
12. Yang MF, Teplow DB. Amyloid beta-Protein Monomer Folding: Free-Energy Surfaces Reveal Alloform-Specific Differences. Journal of Molecular Biology. 2008; 384:450–464. [PubMed: 18835397]
13. Hou L, Shao H, Zhang Y, Li H, Menon NK, Neuhaus EB, Brewer JM, Byeon IJ, Ray DG, Vitek MP, Iwashita T, Makula RA, Przybyla AB, Zagorski MG. Solution NMR studies of the A beta(1-40) and A beta(1-42) peptides establish that the Met35 oxidation state affects the mechanism of amyloid formation. J Am Chem Soc. 2004; 126:1992–2005. [PubMed: 14971932]
14. Riek R, Guntert P, Dobeli H, Wipf B, Wuthrich K. NMR studies in aqueous solution fail to identify significant conformational differences between the monomeric forms of two Alzheimer peptides with widely different plaque-competence, A beta(1-40)(ox) and A beta(1-42)(ox). Eur J Biochem. 2001; 268:5930–6. [PubMed: 11722581]
15. Schweitzer-Stenner R, Measey T, Hagarman A, Eker F, Griebenow K. Salmon calcitonin and amyloid beta: two peptides with amyloidogenic capacity adopt different conformational manifolds in their unfolded states. Biochemistry. 2006; 45:2810–9. [PubMed: 16503636]
16. Sgourakis NG, Yan YL, McCallum SA, Wang CY, Garcia AE. The Alzheimer's peptides A beta 40 and 42 adopt distinct conformations in water: A combined MD/NMR study. Journal of Molecular Biology. 2007; 368:1448–1457. [PubMed: 17397862]
17. Yu L, Edalji R, Harlan JE, Holzman TF, Lopez AP, Labkovsky B, Hillen H, Barghorn S, Ebert U, Richardson PL, Miesbauer L, Solomon L, Bartley D, Walter K, Johnson RW, Hajduk PJ, Olejniczak ET. Structural characterization of a soluble amyloid beta-peptide oligomer. Biochemistry. 2009; 48:1870–7. [PubMed: 19216516]
18. Sciarretta KL, Gordon DJ, Meredith SC. Peptide-based inhibitors of amyloid assembly. Methods Enzymol. 2006; 413:273–312. [PubMed: 17046402]
19. Soto C, Sigurdsson EM, Morelli L, Kumar RA, Castano EM, Frangione B. Beta-sheet breaker peptides inhibit fibrillogenesis in a rat brain model of amyloidosis: implications for Alzheimer's therapy. Nat Med. 1998; 4:822–6. [PubMed: 9662374]
20. Fradinger EA, Monien BH, Urbanc B, Lomakin A, Tan M, Li H, Spring SM, Condron MM, Cruz L, Xie CW, Benedek GB, Bitan G. C-terminal peptides coassemble into A beta 42 oligomers and protect neurons against A beta 42-induced neurotoxicity. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:14175–14180. [PubMed: 18779585]
21. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins. 2006; 65:712–25. [PubMed: 16981200]
22. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. Journal of the American Chemical Society. 1995; 117:5179–5197.
23. Esteban-Martin S, Fenwick RB, Salvatella X. Refinement of Ensembles Describing Unstructured Proteins Using NMR Residual Dipolar Couplings. J Am Chem Soc.

24. Jensen MR, Markwick PR, Meier S, Griesinger C, Zweckstetter M, Grzesiek S, Bernado P, Blackledge M. Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. Structure. 2009; 17:1169–85. [PubMed: 19748338]
25. Nodet G, Salmon L, Ozenne V, Meier S, Jensen MR, Blackledge M. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. J Am Chem Soc. 2009; 131:17908–18. [PubMed: 19908838]
26. Lim KH, Henderson GL, Jha A, Louhivuori M. Structural, dynamic properties of key residues in A beta amyloidogenesis: Implications of an important role of nanosecond timescale dynamics. Chembiochem. 2007; 8:1251–1254. [PubMed: 17549789]
27. Yan Y, McCallum SA, Wang C. M35 oxidation induces Abeta40-like structural and dynamical changes in Abeta42. J Am Chem Soc. 2008; 130:5394–5. [PubMed: 18376837]
28. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation. 2003; 15:1373–1396.
29. Shi J, Malik J. Normalized cuts and image segmentation. Ieee Transactions on Computers. 2000; 22:888–905.
30. He, X.; Cai, D.; Niyogi, P. Neural Information Processing Systems, Vancouver. 2006.
31. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. Journal of Physical Chemistry B. 2001; 105:6474–6487.
32. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of Simple Potential Functions for Simulating Liquid Water. Journal of Chemical Physics. 1983; 79:926–935.
33. Bruschweiler R, Case DA. Adding Harmonic Motion to the Karplus Relation for Spin-Spin Coupling. Journal of the American Chemical Society. 1994; 116:11199–11200.
34. Vuister GW, Bax A. Quantitative J Correlation - a New Approach for Measuring Homonuclear 3-Bond J(H(N)H(Alpha) Coupling-Constants in N-15-Enriched Proteins. Journal of the American Chemical Society. 1993; 115:7772–7777.
35. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins. 2010 In Press.
36. Zweckstetter M, Bax A. Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR. Journal of the American Chemical Society. 2000; 122:3791–3792.
37. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–637. [PubMed: 6667333]
38. Day R, Paschek D, Garcia AE. Microsecond simulations of the folding/unfolding thermodynamics of the Trp-cage miniprotein. Proteins. 78:1889–99. [PubMed: 20408169]
39. Showalter SA, Johnson E, Rance M, Bruschweiler R. Toward quantitative interpretation of methyl side-chain dynamics from NMR by molecular dynamics simulations. J Am Chem Soc. 2007; 129:14146–7. [PubMed: 17973392]
40. Fawzi NL, Phillips AH, Ruscio JZ, Doucleff M, Wemmer DE, Head-Gordon T. Structure and dynamics of the Abeta(21-30) peptide from the interplay of NMR experiments and molecular simulations. J Am Chem Soc. 2008; 130:6145–58. [PubMed: 18412346]
41. Maragakis P, Lindorff-Larsen K, Eastwood MP, Dror RO, Klepeis JL, Arkin IT, Jensen MO, Xu H, Trbovic N, Friesner RA, Iii AG, Shaw DE. Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins. J Phys Chem B. 2008; 112:6155–8. [PubMed: 18311962]
42. Wickstrom L, Okur A, Simmerling C. Evaluating the performance of the ff99SB force field based on NMR scalar coupling data. Biophys J. 2009; 97:853–6. [PubMed: 19651043]
43. Best RB, Buchete NV, Hummer G. Are current molecular dynamics force fields too helical? Biophys J. 2008; 95:L07–9. [PubMed: 18456823]
44. Horn HW, Swope WC, Pitera JW, Madura JD, Dick TJ, Hura GL, Head-Gordon T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. J Chem Phys. 2004; 120:9665–78. [PubMed: 15267980]

45. Wickstrom L, Okur A, Song K, Hornak V, Raleigh DP, Simmerling CL. The unfolded state of the villin headpiece helical subdomain: computational studies of the role of locally stabilized structure. J Mol Biol. 2006; 360:1094–107. [PubMed: 16797585]
46. Ekins S, Balakin KV, Savchuk N, Ivanenkov Y. Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and Kohonen and Sammon mapping techniques. J Med Chem. 2006; 49:5059–71. [PubMed: 16913696]
47. Garcia AE. Large-amplitude nonlinear motions in proteins. Phys Rev Lett. 1992; 68:2696–2699. [PubMed: 10045464]
48. Mesentean S, Fischer S, Smith JC. Analyzing large-scale structural change in proteins: comparison of principal component projection and Sammon mapping. Proteins. 2006; 64:210–8. [PubMed: 16617427]
49. Zhang Z, Wriggers W. Local feature analysis: a statistical theory for reproducible essential dynamics of large macromolecules. Proteins. 2006; 64:391–403. [PubMed: 16700056]
50. Paccanaro A, Casbon JA, Saqi MA. Spectral clustering of protein sequences. Nucleic Acids Res. 2006; 34:1571–80. [PubMed: 16547200]
51. Sgourakis NG, Garcia AE. The Membrane Complex between Transducin and Dark-State Rhodopsin Exhibits Large-Amplitude Interface Dynamics on the Sub-Microsecond Timescale: Insights from All-Atom MD Simulations. J Mol Biol.
52. Hartigan. Clustering algorithms. 1975.
53. Mclachlan AD. Gene Duplications in the Structural Evolution of Chymotrypsin. Journal of Molecular Biology. 1979; 128:49. &. [PubMed: 430571]
54. Daura X, Suter R, van Gunsteren WF. Validation of molecular simulation by comparison with experiment: Rotational reorientation of tryptophan in water. Journal of Chemical Physics. 1999; 110:3049–3055.
55. Hukushima K, Nemoto K. Exchange Monte Carlo method and application to spin glass simulations. Journal of the Physical Society of Japan. 1996; 65:1604–1608.
56. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chemical Physics Letters. 1999; 314:141–151.
57. Garcia AE, Herce H, Paschek D. Simulations of Temperature and Pressure Unfolding of Peptides and Proteins with Replica Exchange Molecular Dynamics. Annu Rep Comput Chem. 2006; 2:83–95.
58. Hess B, Bekker H, Berendsen HJC, Fraaije J. LINCS: A linear constraint solver for molecular simulations. Journal of Computational Chemistry. 1997; 18:1463–1472.
59. Miyamoto S, Kollman PA. Settle - an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. Journal of Computational Chemistry. 1992; 13:952–962.
60. Darden T, York D, Pedersen L. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. Journal of Chemical Physics. 1993; 98:10089–10092.
61. Klement K, Wieligmann K, Meinhardt J, Hortschansky P, Richter W, Fandrich M. Effect of different salt ions on the propensity of aggregation and on the structure of Alzheimer's A beta(1-40) amyloid fibrils. Journal of Molecular Biology. 2007; 373:1321–1333. [PubMed: 17905305]
62. Narayanan S, Reif B. Characterization of chemical exchange between soluble and aggregated states of beta-amyloid by solution-state NMR upon variation of salt conditions. Biochemistry. 2005; 44:1444–1452. [PubMed: 15683229]
63. Nose S. Constant Temperature Molecular-Dynamics Methods. Progress of Theoretical Physics Supplement. 1991:1–46.
64. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. J Comput Chem. 2005; 26:1701–18. [PubMed: 16211538]
65. Karplus M, Anderson DH. Valence-Bond Interpretation of Electron-Coupled Nuclear Spin Interactions - Application to Methane. Journal of Chemical Physics. 1959; 30:6–10.
66. Zweckstetter M. NMR: prediction of molecular alignment from structure using the PALES software. Nat Protoc. 2008; 3:679–90. [PubMed: 18388951]
67. Golub, GH.; Van Loan, CF. Matrix Computations. Third Edition. Press, TJHU., editor. 1996.

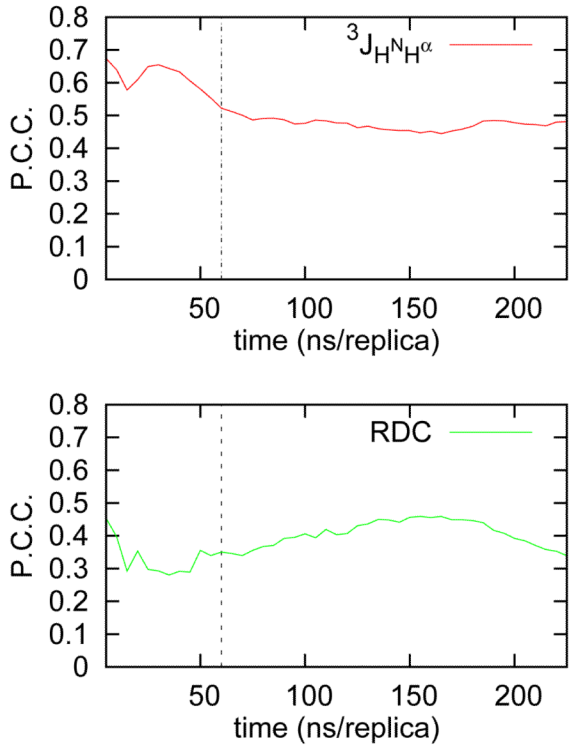

Figure 1. Simulation convergence through comparison with experimental data

Correlation between J-couplings and RDCs calculated from the MD ensemble with their experimentally determined values. Monitoring the Pearson's correlation coefficient (P.C.C) to the experimental data as a function of the total simulation time indicates that the ensemble is converged after approximately 60ns /replica. The dashed line indicates the start of the production phase of the simulation, during which the calculated J-coupling and RDC values are converged to their ensemble-averaged values, as dictated by the details of the forcefield and solvation model used.

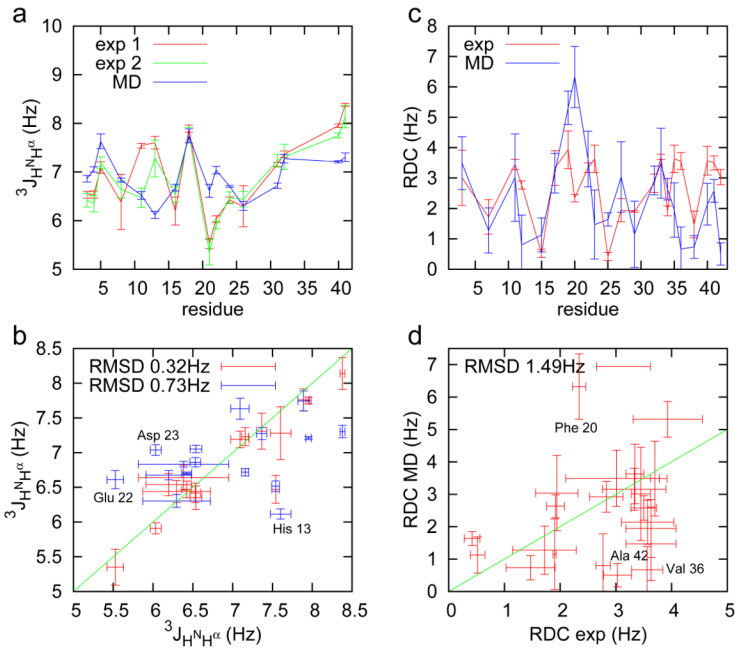

Figure 2. Validation with experimental data

32 experimental three-bond J-coupling values and 22 Residual Dipolar Couplings that report on the average conformation of the backbone and orientation of the amide bond vectors in the molecular alignment frame respectively are compared with results calculated from our REMD simulation trajectories. Two independent experimental measurements of $^3J_{H^NH^\alpha}$ were used [16; 27]. Results are shown in **a, c** along the sequence of Aβ and as correlation plots in **b, d**. Experimental J-coupling and RDC values were measured at 300K and 277.3K respectively while simulation values were calculated over the range of replica temperatures 280-310K. Simulation errors were estimated using block averages [59].

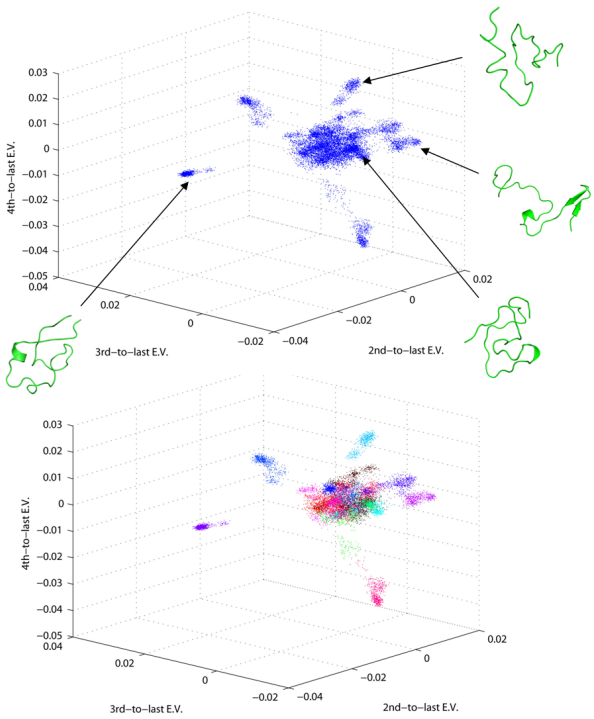

Figure 3. Spectral clustering of the conformational ensemble and identification of representative conformational species

(**a**): Visualization of the REMD ensemble of conformations in a space defined by the last 3 non-trivial eigenvectors of the Affinity matrix. Good dispersion of the data in this low-dimensional space is observed. Each region of the space contains conformations with distinct contact patterns, as shown in the structural diagrams belonging to individual conformations. (**b**): Example of the use of the k-means spectral clustering algorithm for k=20, operating in the space defined by the last 6 non-trivial eigenvectors of the Affinity matrix (see methods). The clustering results are visualized in 3 dimensions corresponding to the last 3 non-trivial eigenvectors, same as in (**a**). Using this technique we have identified several clusters of conformations that share common structural elements as discrete groups in the low-dimensional eigenspace. Representative (central) conformations for selected clusters are shown in the insets, and discussed in the main text. These results suggest that Aβ42 can sample a wide range of conformations with distinct features that can be analyzed using a relatively small set of collective variables.

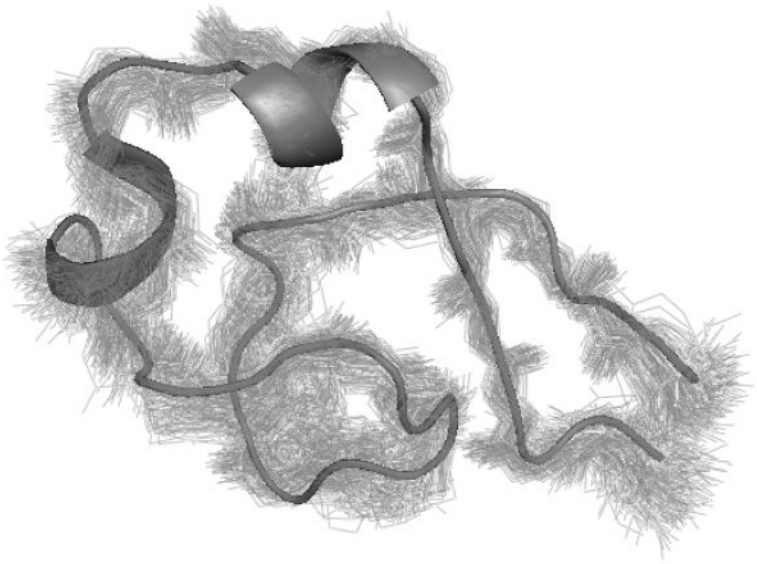

Figure 4. Structural precision in the clustering results

Overlay of all conformations within a single identified cluster according to the spectral clustering technique implemented here. Despite the high degree of structural heterogeneity in the REMD ensemble the contact map-based approach chosen here successfully identifies clusters of conformations with similar features. The central conformation of this cluster is shown in a cartoon representation, while the trace of the backbone is shown for every other member of the cluster. A high degree of structural similarity among conformations within the same cluster is observed, as indicated by an average pairwise RMSD of 1.33Å for the protein backbone among cluster members.

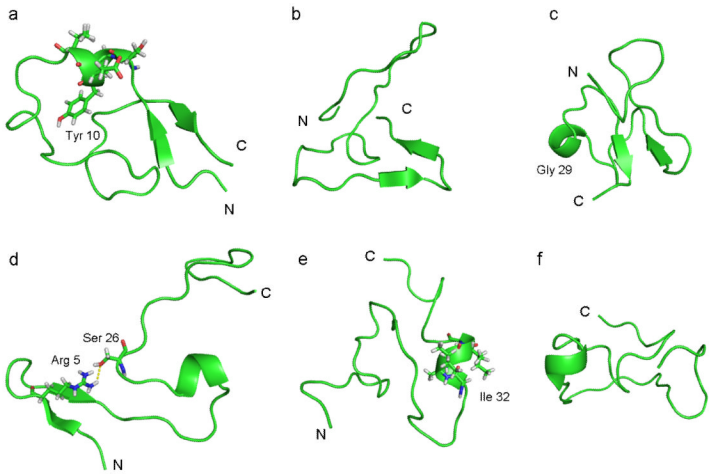

Figure 5. Conformations of Aβ in water

Representative conformations of Aβ42 from the REMD ensemble as identified using the spectral clustering technique. A diverse mixture of extended as well as collapsed coil conformations with secondary structural elements is observed.

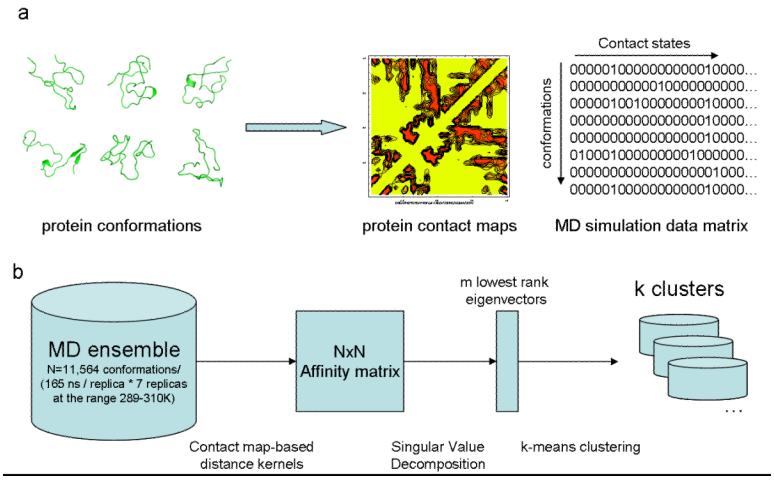

Figure 6. Flow diagram of the spectral clustering method

In **(a)** a diverse ensemble of conformations obtained from enhanced-sampling molecular dynamics is encoded as a binary distance matrix (contact matrix) where each column represents the state of a residue contact (i,j) defined according to a distance threshold of 4.5Å between any pair of heavy atoms belonging to residues i and j. In **(b)** the original MD dataset, in the contact matrix representation is used to calculate a square Affinity matrix, whose elements are given by $e^{-D_{ij}}$ where D_{ij} is the distance between conformations with indices i and j according to a chosen distance kernel. The singular value decomposition of the Affinity matrix yields eigenvectors of high discriminative power. In particular, the m lowest non-trivial eigenvectors (where m<<N) can be used as explicit coordinates to separate the MD ensemble into k clusters using the k-means clustering algorithm.

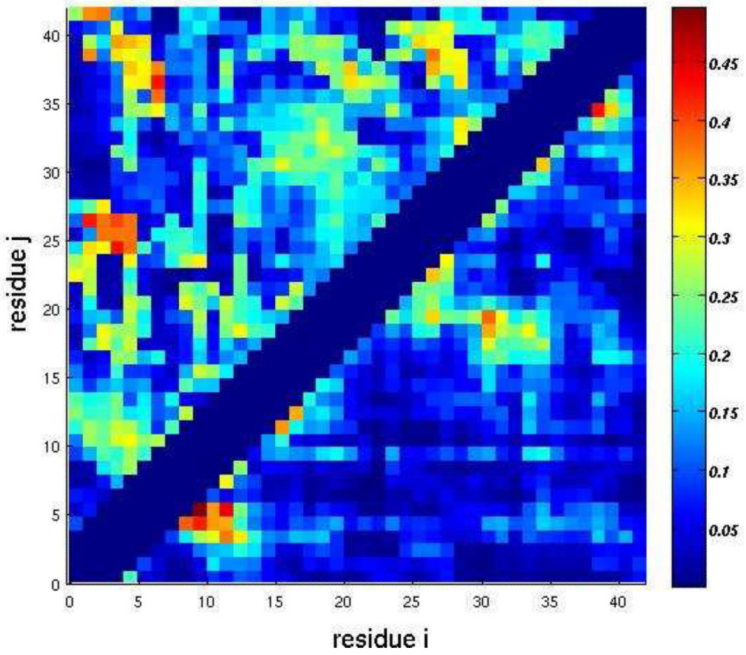

Figure 7. Identification of discriminative contacts using Laplacian scores

The raw probabilities of contact formation between all pairs of residues *i*,*j* in Aβ42 according to a 4.5Å distance threshold (**lower right quadrant**) are contrasted to the extracted Laplacian scores for the same residue pair (**upper left quadrant**). According to this analysis we identify several contacts of high power in discriminating between different conformational species that are not apparent from a simple inspection of the ensemble-derived statistics of contact formation, as discussed in the text. This information can be used to design experimental probes to investigate transitions between different conformations.