Application of Discrete Molecular Dynamics to Protein Folding and Aggregation

S.V. Buldyrev

Abstract With the rapid increase in computational speed and memory, simulations of proteins and other biological polymers begin to gain predictive power. However, in order to simulate a folding trajectory of a moderate size protein or an aggregation process of a large number of peptides, traditional molecular dynamics methods based on explicit solvent and accurate force field models still must gain several orders of magnitude in speed. Under these circumstances, simplified models which capture the essential features of the system under study may shed light on the problem in question. One of these simplified methods is discrete molecular dynamics (DMD). DMD replaces the interaction potentials between atoms and covalent bonds by discontinuous step functions. This simplification as well as coarse graining of the model (replacing groups of atoms by one effective bead) and replacing the effect of solvent by varying the strength of inter-bead interactions can speed up simulations sufficiently to generate many folding–unfolding events and to track the aggregation of many peptides. This increase in speed is gained mainly due to the ballistic motion of either secondary structures of the protein or individual peptides. This ballistic motion is a characteristic feature of the DMD method. This chapter will review successes and failures of the DMD method in protein folding and aggregation.

UNCORRECTED PROOF

1 Introduction

Protein folding and protein aggregation are very important problems in biology and medicine. In spite of enormous advances in experimental studies of proteins, the problem of identification of the protein native state given its amino acid sequence and the inverse problem of designing a protein with a given native state remain unsolved. Many neurological diseases including prion diseases (such as the notorious Mad Cow disease) and Alzheimer's disease, as well as various genetic disorders are related to protein missfolding and subsequent polypeptide aggregation into

S.V. Buldyrev
Department of Physics, Yeshiva University, 500 West 185th Street, New York, NY 10033 USA,
buldyrev@yu.edu

Buldyrev, S.V.: *Application of Discrete Molecular Dynamics to Protein Folding and Aggregation.*
Lect. Notes Phys. **752**, 97–131 (2008)
DOI 10.1007/978-3-540-78765-5_5 © Springer-Verlag Berlin Heidelberg 2008

insoluble fibrils [1, 2, 3]. Understanding of these processes is extremely important for prevention and treatment of these diseases. Can molecular dynamic simulations be of use in this area? All-atom molecular dynamic simulations with accurate forcefields and explicit solvent are still too slow to simulate complete folding and aggregation trajectories. Therefore, simplified coarse-grained models, which replace solvent by effective attraction or repulsion of the residues, are needed.

One such approach is discrete molecular dynamics (DMD), which replaces atoms or groups of atoms by hard spheres interacting by discontinuous stepwise potentials. DMD has been proven useful for studies of simple liquids [4, 5, 6, 7, 8, 9, 10, 11, 12], polymers [13, 14, 15, 16, 17, 18, 19], colloids [20, 21, 22], lipid membranes [23], and DNA-histone binding [24]. For a recent mini-review of the DMD applications for protein folding and aggregation, see [25, 26]. Due to its simplicity, DMD is also an ideal aid in teaching thermodynamics, physical chemistry, and polymer physics [27, 28]. Here we review recent works which use DMD in the studies of protein folding and aggregation.

2 Discrete Molecular Dynamics

DMD, also known as discontinuous molecular dynamics or event driven molecular dynamics, was introduced in 1959 by Alder and Wainwright [4] for simulations of hard spheres. Later it was used by Rapaport [13, 14, 29] for simulation of polymer chains, and finally was adopted for simulations of protein-like polymers [15].

Traditional molecular dynamics [30, 31, 32] approximately integrates Newton's equations of motion of particles interacting via continuous pair potentials (e.g., Lennard-Jones or Coulomb) by updating particles coordinates and velocities at fixed time steps of the order of a few femtoseconds. DMD [31] approximates these potentials by a discontinuous step-functions of interparticle distance r. Thus in DMD, particles move along straight lines with constant velocities until a moment of collision, i.e. a moment of time at which r becomes equal to the point of a discontinuity of the potential (Fig. 1). This discontinuity may be of an infinite height (hard-sphere, or an unbreakable chemical bond) or of a finite size (square well or shoulder, Fig. 1b). The exact time of the next collision can be obtained by finding a minimal positive solution of the correspondent quadratic equations for all pairs of particles (See Appendix A for details). Next, the velocities of the pair of colliding particles are updated using laws of energy, momentum, and angular momentum conservation. These one scalar and two vector equations are sufficient to find the six unknown components of the velocities after the collision and can be solved exactly by reduction to a single quadratic equation of energy conservation. If this equation has no roots, it means that particles do not have enough kinetic energy to jump out of the square well and they recoil back, as in hard-core collision, without change in kinetic energy. Thus, in contrast to the traditional molecular dynamics, DMD provides an exact solution of the system interacting via given discontinuous potentials with strict (subject only to rounding-off errors) conservation of energy and momentum.

UNCORRECTED PROOF

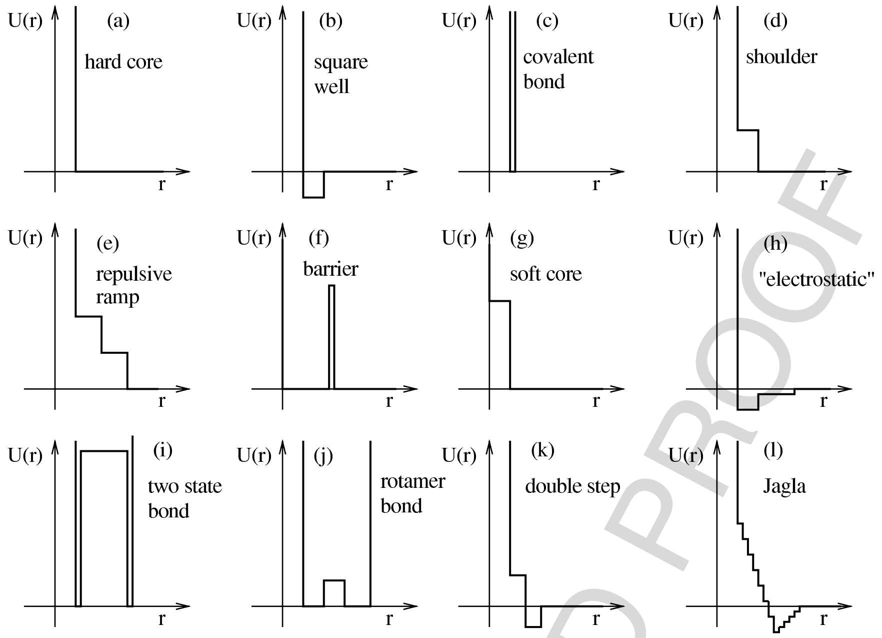


Fig. 1 A collection of DMD potentials used in various studies. (**a**) Hard spheres introduced by Alder and Wainwright [4]; (**b**) square wells and (**c**) covalent bonds were used to study polymer collapse by Rapaport [14]; (**d**) repulsive shoulder used to model hydrophilic interactions or non-native contacts [16, 51]; (**e**) repulsive ramp with two steps used in [87] for auxiliary interactions in hydrogen bond algorithm and with multiple steps in [9, 10, 11] to model liquids with negative thermal expansion coefficient; (**f**) potential barrier used in [21] for ghost particles which occupies no volume but serve as a heat bath; (**g**) soft core is used in all our studies to create an initial configuration of nonoverlapping hard spheres by running at low temperature; (**h**) long-range potential used to model electrostatic interactions in Refs. [61, 99, 106]; (**i**) two-state bond is used to create auxiliary bonds between the backbone beads if they are also linked by a covalent bond [69, 87]; (**j**) complex bond potential to simulate rotamers [61]; (**k**) double-step potential used to simulate liquid–liquid phase transitions [5, 6, 8]; (**l**) multi-step potential approximating Jagla model [38] for water [12, 19]

The distribution of atom velocities after a few collisions converges to the Maxwell distribution. Thus the chance that a pair of particles will increase its potential energy by ΔU is proportional to the flux through the rim of the square well or shoulder of the particles with the kinetic energy larger than ΔU. This flux is proportional to $\exp(-\Delta U/k_\mathrm{B}T)$, where k_B is the Boltzmann constant and T is absolute temperature. On the other hand, the probability of entering the well or descending the shoulder is always one. Hence the DMD is equivalent to the Metropolis Monte-Carlo in which the set of moves is not artificial but is equivalent to the ballistic motion of the particles. Hence DMD is suitable for finding dynamic properties of the system, such as diffusion coefficient, viscosity, and time correlation function.

A variety of different DMD potential types have been used to solve various problems in condensed matter physics (Fig. 1). As the sizes of the steps in the discontinuous approximation of a continuous potential approach zero, the DMD

UNCORRECTED PROOF

trajectory converges to a trajectory of a traditional molecular dynamics for this continuous potential. Since the coordinates and velocities are updated only at the moments of collisions, DMD is especially efficient for dilute systems interacting with very crude potentials such as a hard core plus a single repulsive shoulder or an attractive well. An example of such a system relevant to protein folding is a model of heteropolymer in a vacuum. Interestingly, such a crude approximation of a potential (after all, Lennard-Jones potential is also a crude approximation) is often sufficient to get essential physics and even chemistry right. The ballistic motion of the particles in the DMD is the main feature which allows to speed up the processes of folding and aggregation. The distant parts of the protein and different peptides approach each other ballistically instead of by slow reptation through the surrounding solvent. This helps to increase the computational speed by orders of magnitude. The disadvantage of this speed-up is that one cannot directly predict folding and aggregation rates.

AU: 'The distant ... solvent.' Text seems to be missing. Please check.

The step potential for a pair of particles of given types A and B can be encoded by a string:

$$A\ B\ r_0\ r_1\ \epsilon_1\ ...\ r_n\ \epsilon_n,$$

where r_0 is a hard core distance and $r_n > r_{n-1} > ... > r_1 > r_0$, are the distances at which the potential has a discontinuity of step ϵ_i. The values of ϵ_i are positive for repulsive shoulders and are negative for attractive wells. If the last ϵ_n is omitted, it means that the particles are linked by the permanent bond whose distance can fluctuate between r_0 and r_n.

In addition, DMD can be efficiently used to model chemical reactions since it is possible to change the type of a particle once another particle approaches it within a certain distance and forms a chemical bond. After the reaction, the members of a bonded pair may interact differently with each other and with other particles. In this way, it is possible to model formation of hydrogen bonds and take into account the maximal valence of a given atom [21, 22]. Moreover, this is an effective way to model many body interactions, since the particle type may depend on the particular configuration of its neighbors. Recently, DMD has been applied for modeling physical gels and strong glass-forming liquids using the maximal valence model [21, 22].

It is also easy to implement in DMD various macroscopic objects so long as they are planes or spheres. This can be used for modeling systems in the confined geometry. It is also possible to implement a gravitational force as well as collisions with ghost particles to model a perfect canonical thermostat [16, 21]. Very recently, DMD has been extended to non-spherical objects [33], (ellipsoids [34, 35]) and patchy surfaces which is potentially a very powerful method for modeling protein crystallization [36].

The discontinuous pair potentials are suitable for accurate modeling of dihedral angles by introduction of auxiliary bonds with a small distance between r_{n-1} and r_n connecting the next to the nearest and third nearest atoms along the chain. However, these bonds result in a lot of small interval collisions, which significantly slow down the computation. An alternative approach for efficient modeling of dihedral angles within the DMD algorithm was proposed in [37].

UNCORRECTED PROOF

As one can see, DMD is well suited for studies of simple crude models with the goal of understanding the minimal features of the system needed to reproduce a given phenomenon. The examples of successful application of the DMD are modeling fluids with several critical points [5, 6, 7, 8], water-like thermodynamic anomalies [9, 10, 11, 12], anomalous glass transitions in colloids in which the two different glasses (repulsive and attractive) can exist [20], modeling of physical gels and strong glass-formers with maximal valence model [21, 22]. Very recently, a simple model of a non-polar solvent which exhibits the decrease of hydrophobicity upon cooling has been proposed [19].

Simple geometry of the potentials used in the DMD allows us to understand the basic mechanisms of a phenomenon under study. For example, anomalous expansion of water as well as its hydrophobic effect can be reproduced by a spherically symmetric potential with a repulsive ramp (soft core) [38]. The rigid hydrogen-bond tetrahedron of the nearest neighbor water molecules corresponds in this model to a hard core, while the more flexible second shell of neighbors represents the soft core. As the temperature increases, some of the particles from the second shell may enter the first shell jumping onto the repulsive ramp. Thus the average distance between the particles reduces and the liquid may shrink upon heating. Analogous effect explains the increase of hydrophobicity upon heating. Small solute particles like alkanes can no longer find sufficiently large cages between solvent particles which become closer to each other as the temperature increases. This effect is the basis of the cold denaturation of proteins [19].

DMD can also be used for accurate prediction of physical and chemical properties of a system, such as the native state of the protein given its amino acid sequence. But in this case one faces a formidable problem of parameterization of the potentials akin to the same problem in traditional MD. Often the parameters of the DMD potential lacks any physical meaning (like auxiliary bonds) and must be introduced only to mimic the geometry of the peptide backbone or hydrogen bonds. Also, since using the explicit solvent immediately eliminates all the advantages of the DMD, one must model the hydrophobic and amphiphylic interactions by the effective attraction or repulsion between amino acids or specific atoms. Since the hydrophobic effect is produced by water molecules which form cages around hydrocarbon groups, this effect strongly depends on temperature and other neighboring groups. Therefore, one needs to introduce different potentials for the same atoms in different groups and possibly the three-body interactions by the reaction scheme above discussed. The parameters of these complex interactions can be obtained by means of statistical analysis of the protein data bases, but this requires huge effort and decades of human-years. While in the last two decades a lot of groups were involved in developing traditional MD (CHARMM [39, 40], GROMACS [41], NAMD [42], AMBER [43], LAMMPS [44]), only a handful of researches work on the development of the DMD. Nevertheless, a substantial progress has been made (e.g., the development of the PRIME model by Hall and co-workers [45]).

The bottle-neck of the DMD algorithm is the effective sorting of the collision times. The computational cost of a naive algorithm which computes all the pair

UNCORRECTED PROOF

collisions and move all the particles after each collision scales as N^3 where N is the number of particles. Several ways of solving this problem are developed [17, 31]. Our sorting algorithm is described in Appendix A. It allows to reduce computational cost to $N \ln N$.

UNCORRECTED PROOF

3 Protein Folding

A polypeptide with a uniquely folded 3d conformation (native state) is called a protein. For a long time, biophysicists were puzzled by the following questions: What makes a polypeptide a functional protein? How does an amino acid sequence define a unique 3d structure of a protein? How can one design a protein with a given native sate? How a does protein find its native state in the vast configurational space? If it would proceed by a random search, a simple estimate predicts that the folding time will be larger than the age of the Universe (Leventhal paradox) [46]. For a recent review of these problems see [47].

What is the minimal set of features of a model that ensures that a heteropolymer folds into a unique native state? This question was addressed in the 1990s with help of lattice models (see [47] and references therein). In a sense, this approach was not unlike the minimalistic studies of Picasso, who created a drawing of a bull with a minimal set of features, which however still allowed a spectator to recognize it [48]. Biologists usually do not appreciate this approach, and so in order to be helpful, we must try to move in the opposite direction, from Picasso to Velasquez.

Today it becomes clear [47] that a random heteropolymer does not have a unique native state. Its potential energy landscape has many deep minima, the difference between which are just a few $k_B T$. Thus a random heteropolymer will fold into one of these deep minima. The studies of the lattice models of heteropolymers show that in order for the protein to have a unique native state, its energy must be by several standard deviations lower than the minima of the rest of the basins. It is possible to implement an artificial mutation process based on the Metropolis algorithm which maximizes the Z-value, i.e. the ratio of the difference between the energy of the native state and the energy of a typical basin to the standard deviation of the energy distribution of the basins. Lattice heteropolymers designed in such a way fold into the native state given by a contact map, i.e. the matrix of contacts between the amino acids occupying the adjacent lattice sites in the native state. The success of lattice models suggests that the biologically active proteins are the result of natural selection, which has gradually increased the stability of the primitive pre-biotic proteins [49].

4 The One-Bead Go Model

The next step toward a more realistic picture of a protein would be to design an off-lattice model which folds to a prescribed globular native state. A natural candidate for this model would be a bead-on-a-string model which interacts via square well potentials [13, 14, 29]. The values of the potentials may be taken from an effective matrix of interactions derived from the probabilities of amino acids to be close to each other in the native state of the existing proteins [50]. This model can be very efficiently studied by the DMD. The initial globule can be created by a collapse of a homopolymer, interacting via identical attractive square wells. The initial sequence of the protein is a random sequence of twenty letters and then it is changed by mutations maximizing the energy gap between the energy of the native state and the energies of random contact maps representing missfolded states as it has been done for lattice models. Our studies have shown that this approach does not work. The bead-on-a-string model has too many degrees of freedom and too many contacts per amino acid. The number of contacts reaches ten for the beads in the central core of the bead-on-a-string model while in the lattice model it is only four. So it is impossible to design a protein-like sequence of sixty beads using only the twenty-letter code.

We have found [51] that the Go model [52, 53, 54] which uses 60×60 matrix of interactions (Fig. 2) for a sixty-bead polymer works very well. In the Go model, the beads which are within a certain distance in the native state attract to each other while those that are further away repel. Thus, the native state is by definition the ground state of the model. Moreover, its energy gap with a randomly missfolded globule is of the order of n_c, where n_c is the number of native contacts. We find that the Go model of a small globular heteropolymer always folds into a native state near the folding temperature, T_f. Moreover, this happens in a reversible way, so that the polymer folds and unfolds many times during the simulation. At $T = T_f$, the folded and unfolded conformations are equally populated and separated by a significant energy gap which is about one half of the total number of native contacts (Fig. 3). This bimodal distribution is a characteristic of the first-order phase transition in which the two phases (liquid and crystal) may coexist and the potential energy gap between the two phases is proportional to the number of molecules in the system. This is in sharp distinction to the behavior of a flexible homo-polymer near the theta point, which undergoes a second-order phase transition and has a unimodal distribution of energies.

Interestingly, the models produced by the Go algorithm from a collapsed state of a homoplymer often have intermediate states in which a tail consisting of a significant number of beads is detached from the rest of the folded globule. The partially folded states comprise an intermediate bump on the potential energy distribution. Cutting away this tail yields a perfect two-state folder, which represents the majority of the small proteins.

As the temperature of the system is reduced below the folding temperature, the distribution becomes unimodal, with the probability of being in the folded state dominating (Fig. 4). However, if the temperature of an unfolded state is reduced below a certain value, which is about 70 % of the folding temperature [51], the polymer may never find its native state and may be trapped forever in a missfolded state. This phenomenon is analogous to the glass transition in the supercooled liquids, in which the nucleation of the crystal can be avoided by fast quenching. For protein A, this phenomenon is observed at $T = 0.62$, which is about 80% of $T_f = 0.765$.

UNCORRECTED PROOF

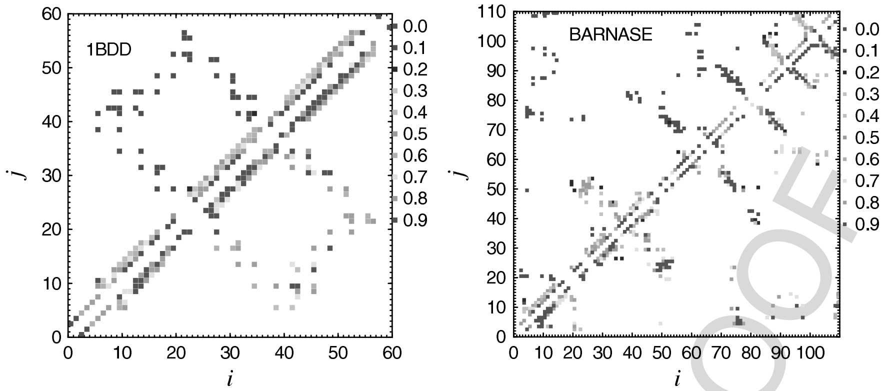


Fig. 2 A Go interaction matrix (contact map) constructed for the B domain of staphylococcal protein A (1BDD) and barnase (1BNR) using their native states taken from the Protein Data Bank. The amino acids for which C_β atoms in the native state are separated by less than 7.5 Å attract to each other via a square well interaction (Fig.1b) with $r_0 = 3.459$ Å , $r_1 = 7.5$ Å , and depth $\epsilon_1 = -\epsilon$, while other amino acids repel from each other via a square shoulder interaction (Fig. 1d) with the same r_0 and r_1 and height $\epsilon_1 = +\epsilon$. The 1BDD protein has 3 α-helices held together by 45 long-range contacts. The total number of native contacts is 160 which corresponds to the ground state potential energy of -160ϵ. A particular structure of the contact map can vary a lot. For example, SH3 domain has 3 β-hairpins. An artificial globule constructed by the homopolymer collapse has neither α-helixes nor β-sheets, but usually have about cN native contacts where N is the number of monomers and c varies between 2 and 3. The number of long-range contacts is about 0.25–0.3 of the total number of contacts. Colors indicate the probabilities of contacts in the folded (*lower triangle*) and unfolded (*upper triangle*) states at $T = T_F = 0.765$. In the unfolded state, all long-range contacts have very low probability, while the secondary structure is already partially formed. In the folded state, many long-range contacts are formed with probability larger than 0.8. These contacts form putative folding nucleus. The energy gap between the folded and unfolded state is 54ϵ. The analogous Go model for barnase has 316 native contacts which is two times larger than for protein A. This Go model also folds cooperatively into the native state but the energy gap between the folded and unfolded states is two times wider than for protein A (107ϵ). So at $T_f \approx 0.8$ barnase can undergo only few transitions between folded and unfolded states in 10^7 time units. The *upper triangle* shows contact probabilities for the unfolded state. The *lower triangle* shows the contact probabilities in the folded state. The protein consists of two α-helices and three β-hairpins. While the α-helices are well formed in the unfolded state the β-hairpins are not present. In order for this protein to fold, the β-hairpins must form cooperatively

This model seems to explain the Leventhal paradox: near the folding temperature, the secondary structure of the unfolded chain is already partially formed with about 50% of the native contacts (mainly short ranged) in place (Fig. 2). The individual elements of the secondary structure are not stable because its potential energy differs from the unstructured conformations only by few $k_B T$. However, the polymer has already lost an immense number of degrees of freedom and acts like a collection of a few secondary structural elements. Once the few critical long distance contacts (the folding nucleus) [55] are formed, the partially folded secondary structural elements come together and the polymer quickly descends into its native state. This process is similar to the formation of the critical nucleus in the first-order

UNCORRECTED PROOF

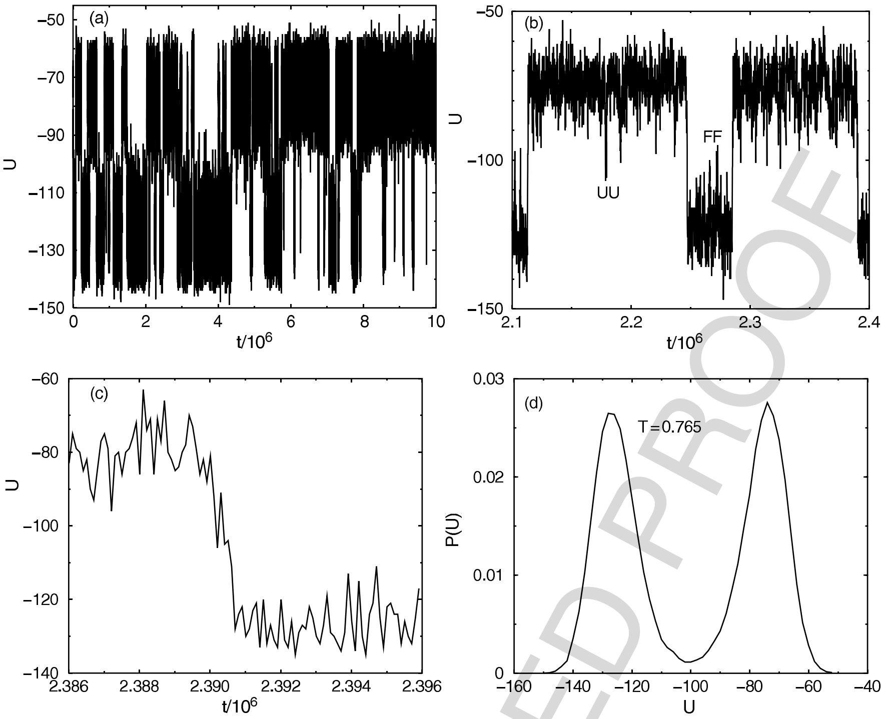


Fig. 3 (**a**) Potential energy versus time for a two-bead Go model of 1BDD protein near $T = T_F = 0.765$. The graphs for the one-bead Go models of two-state proteins look similar. About 60 folding and unfolding events are visible. (**b**) In addition, there are many unsuccessful attempts to fold (UU events) or unfold (FF events). (**c**) Each folding event takes about 600 time units which is about 300 times faster than average time spent in the folded or unfolded state. (**d**) The probability density of the potential energy is bimodal with equally populated folded and unfolded states. The average energy of the unfolded state is -73ϵ which corresponds to the existence of approximately one half of all native contacts

phase transition. During this stage the remaining 50% of the native contacts are formed [55, 56]. However, the discussed folding scenario can be an artifact of the Go model, in which the non-native contacts do not attract to each other and hence the hydrophobic collapse preceding the formation of the secondary structure cannot be observed. Adding small hydrophobic attraction between the non-native contacts to the Go interactions changes the folding scenario [57, 58]. In this case, the protein first undergoes the hydrophobic collapse into a molten globule state, in which the secondary structure is only weakly formed. If the attraction of non-native contacts is significantly weaker than the attraction of native contacts, the molten globule reorganizes itself into the native state in a first-order-like transition, but the folding process is much slower than in the case when the non-native contacts are assigned zero or even positive (repulsive) energy. It is clear that in vitro and in vivo both scenarios can take place depending on the properties of the protein and its environment.

UNCORRECTED PROOF

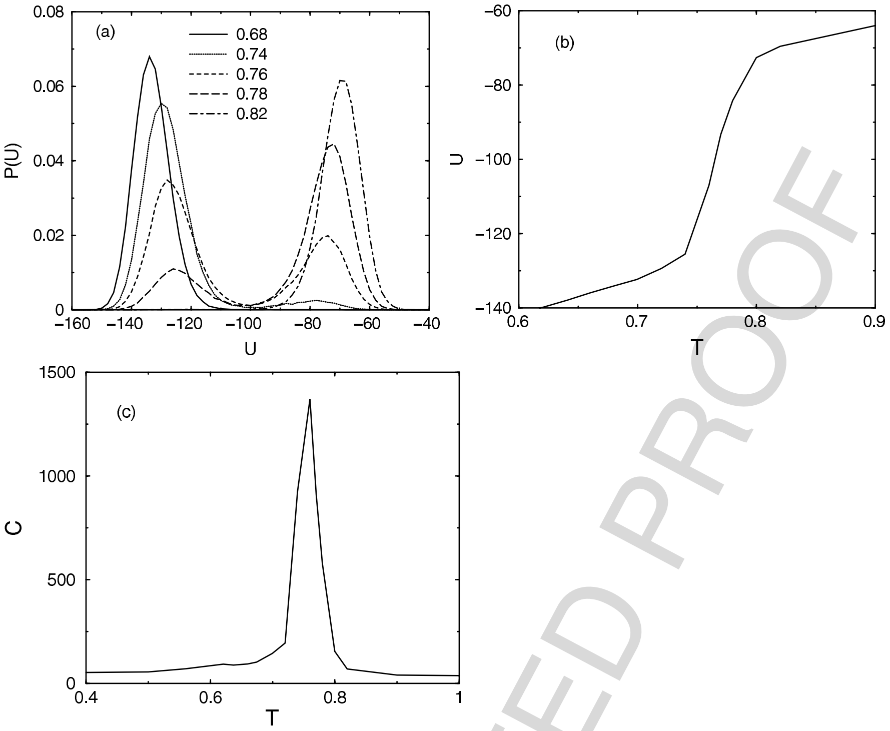


Fig. 4 (**a**) Probability density of the potential energy for different temperatures in the vicinity of the folding transition for the 1BDD protein. The populations of the folded and unfolded states change dramatically while the temperature changes from 0.7 to 0.78. Outside this temperature interval the distribution becomes unimodal. Thus the temperature interval of the folding transition corresponds to about 40 K in the physiological temperature scale. Accordingly, the average potential energy (**b**) dramatically change with temperature near folding transition and the heat capacity (**c**) has a sharp maximum. All these are features of the first-order phase transition. This behavior is typical for the one-bead and two-bead Go models

The transition time during which the trajectory descends from an unfolded to a folded state is by several orders of magnitude smaller than the average folding time, i.e. the time during which the polymer explores the vast ensemble of the unfolded states. This feature is also observed in the folding of all-atom models with explicit solvent. Actually, it is the basis of the Pande's Folding at Home project [59]. In this approach, hundreds of thousands participants world-wide run the simulations as screen-savers on their personal computers. Average folding time of a small protein is of the order of one hundred microseconds, while a typical simulation time is about 10 n. In contrast, the descent of the protein from the transition sate in which the critical nucleus is formed into the native state is within the simulations reach. Thus in one case out of ten thousand, a lucky participant may observe an actual folding. Using these results Pande and coworkers can determine the folding rate, an experimentally verifiable quantity.

UNCORRECTED PROOF

As we see, DMD helps to explain a lot of features of a real protein folding in simple terms. In summary, the DMD Go model confirms that the protein folding has many features of the first-order phase transition, in particular nucleation [60]. Of coarse, since it explicitly uses the information on the native state, it does not bring us closer to the "Holy Grail" of computational biology, which is ab initio folding of a protein given only its amino acid sequence. Can pair potentials, even with correct geometry of the peptide backbone and the side chains, but without the explicit solvent, be successful in this quest? The answer to this question is still unknown. The success of folding of the trp-cage miniprotein by the DMD [61] and by other molecular dynamics methods is probably due to a very specific sequence of this protein with several prolines which make the backbone especially rigid. We will review this study below. In reality, amino acids interact not only with each other but also with the surrounding water, which forms cages around the hydrophobic amino acids and makes hydrogen bonds with polar amino acids. Thus, solvent creates effective many-body interactions among amino acids, i.e. changes the amino acid properties by the effects of their neighbors and thus creates many more effective amino acid types than 20.

AU: 'Thus ... than 20' the sentence seems to be incomplete. Please check.

5 Transition States of Realistic Proteins

Can the Go model predict certain experimentally verifiable features of real proteins, such as their transition state ensembles, the folding pathways, and the presence of intermediates? In order to answer this question, we take [62] a well-studied protein with a known native state, Src SH3 domain, and create a Go interaction matrix (Fig. 2), assuming that if C_β atoms of the side chains are less than 7.5 Å apart in the native state, they attract with the square well of diameter 7.5 Å and potential energy $-\epsilon$, while if they are farther apart they repel from each other as hard-cores at the same distance. The later rule insures that the protein cannot form non-native contacts at all. For glycines, which do not have side chains, we use C_α instead of C_β.

Interestingly, the bead on the string model with this Go matrix and all equal bond lengths do not fold into a native state, so we have to specify the bond lengths as the distances between subsequent C_β atoms in the native state. A version of this model, in which the distances between C_α atoms in the native state specify the Go interaction matrix, does not fold into a native state cooperatively. This result indicates that the conformations of the side chains rather than those of the backbone determine the specific nature of the amino acid interactions. The model shows a much more cooperative transition than the previously studied model of an artificially constructed globule [51, 55, 56], although the SH3 domain has similar length, $N = 56$, and number of native contacts, $N_c = 160$. The distribution function of potential energy has a wide gap between the sharp maxima corresponding to the native and unfolded states, separated by a deep minimum. Similar behavior is observed for many other short proteins (Fig. 3).

UNCORRECTED PROOF

If one assumes that the potential energy (number of native contacts) is a good reaction coordinate, then the conformation corresponding to the minimum in the energy histogram must belong to the free-energy barrier separating the native and unfolded states and thus constitute the transition state ensemble (TSE), which is by definition a conformation that with 50% chance folds within the typical transition time and with 50% chance unfolds. In the SH3 domain Go model, the ratio of the transition time to the time for which protein resides in the native and unfolded states is even smaller than in the artificial globule studies. We find that the majority of the conformations near the minimum of the potential energy histogram do not correspond to actual folding or unfolding events but represent the unsuccessful attempts to fold or unfold. We call these attempts folded-folded (FF) and unfolded-unfolded (UU) events. A FF event is a part of the trajectory, which originates below the maximum of the distribution characterizing the folded states, reaches the minimum of the distribution and without ever reaching the maximum characterizing the unfolded states descends below the maximum characterizing the folded states. The UU events are defined analogously (Fig. 3b). We hypothesize that both FF and UU events have structural similarities to the TSE. We also hypothesize that the difference between the UU and FF events is that in FF events the folding nucleus is not destroyed, while in the UU events the folding nucleus does not form. Thus the contacts which represent the folding nucleus must be those with the maximum positive difference of their probabilities to be in FF and UU events. (Note that the average number of contacts in FF events is smaller than in the UU events, because we sample the UU events below the minimum of the distribution, while the FF events are sampled above the minimum of the distribution.)

We find [62] that some contacts are significantly more abundant in FF events than UU events. In the Go model of the Src SH3 domain, there are about 20 such contacts and all of them belong to the two distinct clusters of long-distance contacts in the contact map: one characterizes the contacts between the termini of the protein while another characterizes the contacts between the distal hairpin and the RT-loop. However, the absolute values of the probabilities to find contacts between the termini are not very high in the FF event, so we conclude that these contacts do not characterize the TSE and thus do not belong to the folding nucleus. On the other hand, the contacts between the RT-loop and the distal hairpin occur in more than half of FF events, so we assume that they do belong to TSE and thus form the true folding nucleus.

We also produce the P-fold analysis as an additional test of the putative folding nucleus. The P-fold analysis consists of randomly changing velocities of the amino acids in a certain conformation and then performing a simulation for a time interval which is significantly longer than a typical transition time but is significantly shorter than the time of staying in the folded or unfolded states. Then we count the fraction P of trials in which the protein ends up in the folded state. The P-fold analysis confirms our identification of the putative TSE and folding nucleus. The amino acids belonging to this putative folding nucleus form contacts between the distal hairpin and the RT loop. This conclusion is however not confirmed in the later studies in which the putative TSE is identified by the all-atom importance sampling MD [63].

UNCORRECTED PROOF

These contacts are not present in any of the 51,000 conformations sampled in the vicinity of the free-energy barrier on the two-dimensional map which uses radius of gyration and potential energy as reaction coordinates. On the other hand, the conformations with termini contacts are abundant in this putative TSE. This discrepancy questions both the ability of simple Go models to make reasonable predictions of the true folding kinetics and the validity of the all-atom modeling based on importance sampling molecular dynamics (in these studies no actual folding trajectories are obtained). Only experiment can bring the final verdict.

This discrepancy is consistent with the idea of multiple folding pathways which we found for the SH3 domain at low temperatures and some other short proteins using the one-bead Go model [62, 64]. At high temperatures, within 20% range from T_f, the protein can fold via two pathways either forming contacts between distal hairpin and RT loop or (with smaller probability) forming contacts between C-N termini. At high temperature both pathways are fast and the folding is optimal at about $0.85T_f$, which is consistent with experimental observations. However, at $T < 0.60T_F$, the protein can be trapped in the intermediate state with the contacts between the termini formed prematurely. In order to proceed to the folded state, the protein must first break these contacts which requires certain activation energy. That is why, the time spent in the trap diverges at low T following the Arrhenius law.

In addition, we study unfolding of nine other proteins which are known to have folding intermediates. In general, our results agree with the experimental findings: namely for the two-state folding protein like SH3 and Im9 domains, the Go model predict cooperative folding with no intermediates, while for the proteins with intermediates, including Im7 which is homologous to Im9 but is known to have intermediates, the Go model also predicts intermediates. This remarkable success of Go models suggests that it is the topology of the native state rather than the amino acid sequence that determines the kinetics and thermodynamics of the globular proteins.

A different variant of the one-bead model with Go interactions has also been used for DMD simulations of various folding pathways in α-helical [57, 58] and β-stranded proteins [65]. These simulations use a pseudo dihedral angle potential which creates a chirality bias toward right-handed α-helices [16]. In these models all the contacts, both native and non-native, can form but the interaction energy of the native contacts is larger than that of the non-native ones. By varying this energy gap between native and non-native contacts, various folding scenarios are observed. For the large energy gap, the protein can quickly not only fold into a native state, but can also be trapped in the partially missfolded conformations. For the low energy gap, the protein first collapses into a compact disordered structure similar to a molten globule [66] and then slowly folds into a native state. Thus, these studies show that both scenarios are possible for protein folding. In the first scenario, parts of secondary structural elements form in the unfolded state [67], which then fold into the native state. During this process missfolded conformations can form which must unfold for the successful folding into the native state. In the second scenario, proteins collapse into a disordered globule, which later fold into the native state either in a cooperative transition or via non-obligatory intermediates.

AU: 'For ... can quickly fold into ...' has been changed to 'For ... can quickly not only fold into ...' Is this ok?

UNCORRECTED PROOF

However, a third scenario in which the hydrophobic collapse happens simultaneously and cooperatively with the formation of the secondary structure is probably the most likely one. Modern theory of hydrophobic interactions [68] predicts complete dewetting of large polymer globules and formation of the water phase boundary around globules exceeding 1 nm in diameter. The formation of such a globule is a two-state, first-order-like phase transition even for a flexible homopolymer. Recently, this transition has been demonstrated by the DMD simulations of a hard sphere polymer in a water-like Jagla solvent [19]. The formation of the relatively weak hydrogen bonds between the backbone carbonyl and amides which are crucial for formation of the secondary structure is unlikely if the backbone is surrounded by water molecules which can form much stronger hydrogen bonds with the backbone groups. Expulsion of water by the hydrophobic dewetting caused by the hydrophobic side chains enhances the formation of the backbone hydrogen bonds which in its turn serves as positive feedback for the collapse. Unfortunately, this scenario is impossible to simulate replacing the effect of solvent by effective pair potentials between protein atoms.

6 The Two-Bead Go Model

We see that the one-bead Go model is too flexible and needs adjustments of the distances between C_β atoms in order to take into account the geometry of the backbone and the side chains. Thus, we take a next step "toward Velasquez" and explicitly add the side chains (each represented by a single bead C_β) to the backbone, which is still represented by the chain of C_α [69]. In order to reduce the flexibility of the backbone, we add auxiliary bond linking next to the nearest C_α atoms along the backbone. These bonds make rigid isosceles triangles $C_{\alpha,i-1}C_{\alpha,i}C_{\alpha,i+1}$ with the angle at $C_{\alpha,i}$ of approximately 96^o and the distance between $C_{\alpha,i-1}$ and $C_{\alpha,i}$ of approximately 3.8Å. The $C_{\beta,i}$ is attached to the $C_{\alpha,i}$ bead by a covalent bond of 1.53Å and by two auxiliary bonds linking it with $C_{\alpha,i-1}$ and $C_{\alpha,i+1}$ in order to keep the covalent bond approximately orthogonal to the plane of the triangle $C_{\alpha,i-1}C_{\alpha,i}C_{\alpha,i+1}$. Thus, the entire protein consists of imperfect tetrahedra connected by their edges in such a way that $C_{\beta,i}$ and $C_{\beta,i+1}$ point in opposite directions. The methods for determining the Go interaction matrix is the same as in the one-bead model (Fig. 2). Note that glycines lack C_β and we compute their native contact map using their C_α coordinates. Both covalent and auxiliary bonds can fluctuate by a few percent within a square well. This flexibility intends to mimic the actual statistics of the backbone geometry. Thus the model can form approximate β sheet conformations as well as α-helixes.

We use the two-bead model to study the SH3 domain TSE [69] performing P-fold analysis [70] and FF–UU event analysis as in [62]. As in the one-bead model, the SH3 domain folds in a highly cooperative two-state transition. The transition state appears to be the same in both one-bead and two-bead models. In addition, we simulate Φ values using a virtual screening method which we develop for the

UNCORRECTED PROOF

Go model and compare them with the experimental Φ values. The virtual screening method computes the shifts in the Gibbs potentials ΔG_{U}, ΔG_{T}, and ΔG_{F}, of the unfolded, transitional, and folded states of the protein, produced by a "mutation" of an amino acid which turns off the interaction energies of this amino acid with its neighbors in the native state. It is assumed that such point mutations usually do not alternate the topology of the native and transition states. The change in the Gibbs potential of a certain state X due to mutation is defined as $\Delta G_X = -k_{\mathrm{B}}T \ln\langle \exp(-\Delta E_X/k_{\mathrm{B}}T)\rangle$, where ΔE_X is the change in potential energy due to the mutation, and $\langle ...\rangle$ determines the average over all the observed conformations in the considered state. Finally, we compute the Φ value of the amino acid as $\Phi = (\Delta G_{\mathrm{T}} - \Delta G_{\mathrm{U}})/(\Delta G_{\mathrm{F}} - \Delta G_{\mathrm{U}})$, which compares the increase in the height of the free energy barrier to the decrease in the protein stability [71].

Note that no additional simulations are made in the virtual screening method but we use the conformations obtained in the original Go-model. Thus, the Φ values are fully determined by the contact map distributions of the transition state. Amino acids which have the same number of contacts in the transition state as in the native state have $\Phi \approx 1$. The amino acids which do not form native contacts in the transition state have $\Phi \approx 0$. Note that in the Go model the Φ values can be only within the range $[0,1]$ while in experiments they can be sometimes larger than one or even negative; thus, one cannot expect high correlation between the simulated and the experimental values. Indeed, we found the correlation coefficient $r = 0.58$. Particularly disturbing is the high probability of the contact between amino acids L24 and G54 from the RT loop and distal hairpin in our simulations and their low experimental Φ values. On the other hand, these are the most conserved amino acids in the SH3 family, so it seems that they should be important for the protein stability. In fact, the mutation of G54 in experiments destabilizes the native state but reduces the transition state barrier, thus it has a negative experimental Φ which may indicate its participation in the transition state. This may be explained by the presence of the backbone hydrogen bond between G54 and L24, which is not destroyed in mutations.

To further test the role of this particular contact in the Go model, we replace it with a permanent bond. The crosslinking of these amino acids significantly increases the population of the transitional state, while the crosslinking of the termini does not change it significantly. Nevertheless, the question of the relative importance of the two folding pathways in SH3 domain remains open. A recent article of Lam et al. [72] shows that by changing the relative strength of the Go-interactions in the different segments of the two-bead model of the SH3 domain, one can dramatically increase the probability of the folding pathway via formation of the contacts between the termini without the change in the stability of the native state and cooperativity of the folding transition.

The secondary structure of the SH3 domain consists of only β-hairpins and does not have α-helices. As an example, we test the two-bead model by folding two small proteins, the B domain of staphylococcal protein A (PDB access code 1BDD) which has three α-helices and barnase (PDB access code 1BNR) which has both α-helices and β-hairpins (Fig. 2). It appears that 1BDD folds cooperatively in a two-state

UNCORRECTED PROOF

transition the same way as the SH3 domain (Figs. 3 and 4). At the folding temperature, $T_{\mathrm{f}} = 0.765$, both folded and unfolded states are equally populated and the protein undergoes rapid folding and unfolding transition. In the unfolded state, the α-helices are already well formed but the long-range contacts do not present. Thus during folding, the α-helices come together and form the tertiary structure. The folding nucleus belongs to the set of long-range contacts.

Similar situation is observed in barnase, which cooperatively folds at $T_{\mathrm{f}} \approx 0.8$ in a two-state process. This result coincides with the conclusions of [64] in which the absence of the intermediates for barnase has been reported based on a few unfolding trajectories of the one-bead Go model. We study the structure of folded and unfolded states at $T = 0.8$. In the unfolded states, the α-helices are well formed but some of the β-hairpins are not present. This is clear because the α-helices are formed by the short-range contacts and thus their formation costs much smaller entropy loss than the formation of β-hairpins for which the contacts have longer range. Accordingly, the protein folds simultaneously with the two central β-hairpins which are likely to be a part of the TSE. There is an experimental evidence that barnase has an on-pathway folding intermediate [73]. The fact that this intermediate aggregates may indicate that it has exposed β-strands. This is consistent with our simulation results.

However, in reality, the formation of the β-hairpins may precede the α-helices. The two-bead Go model does not take into account the entropy of the side chains. Accordingly, β-hairpins may have larger entropy than α-helices and may form at higher temperatures. By changing the energy [72] or the range of the Go-interactions, one can change the folding pathway and reverse the order of formation of α-helices and β-hairpins.

The two-bead Go model is now publicly available for the P-fold analysis of the arbitrary protein structures [70]. It can be used for studies of large-scale conformational dynamics of long proteins consisting of thousands of residues and their binding [74]. It has been employed in the study of the conformations of the denaturated proteins [67]. The nature of the Go interactions leads to the high abundance of the native secondary structural elements in the denaturated states. However, this result may depend on the relative strength of the native and non-native contacts as discussed in [57, 58]. Very recently the two-bead Go model has been used for simulation of histones binding to DNA [24]. Each DNA nucleotide has been modeled by three effective beads.

7 The Two-Bead Model with Hydrogen Bonds: Studies of Protein Aggregation

It is well known that many genetic neurological diseases are caused by aggregation of proteins into insoluble fibrils formed by the β-sheets crosslinked by hydrogen bonds [1, 2, 3]. While an isolated protein can still fold into its native state, the proteins in concentrated solutions can attach to each other by the exposed β-strands, and then find a new deeper free-energy minimum corresponding to the in-

UNCORRECTED PROOF

soluble fibrils with a regular structure. Thus, the fibril formation in many aspects is similar to crystallization except that in protein crystallization the proteins remain in the native state and are attached together by relatively weak side chain interactions. The mechanism of this phenomena is of crucial importance in developing drugs which would prevent protein aggregation and thus stop the development of such devastating neurological syndromes as Alzheimer's disease, Parkinson's disease, Huntington's disease, and prion diseases [1, 2, 3, 75, 76].

As a first step in modeling aggregation, we introduce hydrogen bonds into our two-bead model in addition to the Go interactions [77, 78]. The amino acid in each peptide is identified by its index i, which is its number starting from the N terminus. We assume that if a certain pair of amino acids i and j within the same peptide interacts via an interaction potential, determined by the native state, it interacts with the same potential even if its members i and j belong to different peptides. Also, we assume that the hydrogen bonds can form between the amino acids in the same peptide and between the different peptides, except that the hydrogen bonds between amino acids i and j from the same peptide are forbidden if $|i-j| < 3$. This rule is derived from the extensive studies of the structures in the Protein Data Bank (PDB).

In reality, the backbone hydrogen bonds are formed between nitrogens and oxygens from the carbonyl groups of the backbone. Thus each amino acid can form at most two backbone hydrogen bonds one by donating a hydrogen by the nitrogen and another by accepting a hydrogen by the carbonyl. The geometry of the peptide backbone is such that these two bonds must be approximately parallel. In the two-bead model we do not have carbonyls and nitrogens so we introduce the effective bonds between C_α beads. Each C_α is allowed to have two hydrogen bonds. So, each C_α keeps track of the number of hydrogen bonds and the amino acids linked by these bonds. If two beads $C_{\alpha,1}$ and $C_{\alpha,2}$ have no hydrogen bonds then they will always form a new hydrogen bond as soon as they come to a distance of 5 Å. If $C_{\alpha,2}$ already has a hydrogen bond with $C_{\alpha,1}$ and comes to within 5 Å of the bead $C_{\alpha,3}$, the hydrogen bond between $C_{\alpha,2}$ and $C_{\alpha,3}$ can be formed only if the distance between $C_{\alpha,1}$ and $C_{\alpha,3}$ is between 8.7 Å and 10 Å. In addition to this hydrogen bond, an auxiliary bond is formed between $C_{\alpha,1}$ and $C_{\alpha,3}$ which is modeled by the infinite square well of this width (between 8.7 Å and 10 Å). This auxiliary bond keeps the angle between the two hydrogen bonds close to 180° during the time of the existence of these hydrogen bonds. If the beads linked by a hydrogen bond approach 5 Å distance, the hydrogen bond breaks according to the normal rules of the DMD and the auxiliary bonds which may exist between them and their partners break simultaneously without any energy loss. We take the energy of the hydrogen bond $\epsilon_{\mathrm{HB}} = 3\epsilon_{\mathrm{Go}}$, because the hydrogen bond interactions are much stronger than the hydrophobic interactions between the side chains. Of course, the role of water, which may also form the hydrogen bonds with the backbone, is totally neglected. Surprisingly, this highly simplified model, while keeping the native structure of a single protein, produces perfect β-sheets between different peptides.

AU: 'Without any energy cost' has been changed to 'without any energy loss'. Is this ok?

We decided to simulate the aggregation of the SH3 domain [77] because we had already studied its folding and because aggregation does not depend on the amino acid–specific interactions and thus can in principle involve any protein. We placed

UNCORRECTED PROOF

eight identical proteins in the simulation box and raised the temperature, to completely unfold them. Then we ran long equilibrium simulations at different temperatures. The proteins moved ballistically around the box, so the chance that they can soon come in the vicinity of each-other was very high. At first they formed disordered oligomers which later transformed into β-sheet fibrils. The fastest fibril formation happened near folding temperature T_F, when the individual proteins frequently unfolded but retained significant amount of the secondary structural elements in the unfolded state. Accordingly, the SH3 domain proteins formed parallel and antiparallel β-sheets between the RT loops which produce a sort of a barrel with an axis parallel to the direction of the hydrogen bonds. The terminal regions of these proteins remained highly disorganized. The fact that this model can correctly predict the domain swapping between two proteins indicate that even this very simple model can correctly capture some features of protein aggregation [77, 78].

AU: There has been changed in the para 'we decided ... aggregation'. Please Check if it is ok?

Similar results have been achieved in the simulations of β-amyloid peptides whose native state has been modeled by the α-helices [79]. Aggregation of the amyloid-β peptide is believed to be the leading cause of the Alzheimer disease. Interestingly, the peptides first form unstructured oligomers with high degree of α-helical structure still present, and only later do they organize themselves in the perfect multi-layered β-sheets. This finding is in accordance with the present hypothesis that the death of neurons in the Alzheimer disease is caused not by the ordered amyloid fibrils accumulated in the plaques, but by short-lived intermediate oligomers which are precursors for fibrillization [80, 81]. These oligomers are highly mobile and can attach themselves to the cell membranes and probably puncture them. Similar β-fibrils have been obtained by Hall and coworkers in the aggregation of polyalanine [82, 83] and polyglutamine [84] by an intermediate resolution protein model similar to the four-bead model discussed next. The polyglutamine fibril formation is the molecular basis of Huntington disease.

8 The Four-Bead Model: Studies of the α-Helix-to-β-Hairpin Transition

The next step toward a realistic protein model that can fold into a native state without explicitly specifying its topology by the Go interaction matrix is to create a model which can reproduce spontaneous formation of the secondary structure, i.e. α-helices and β-sheets. The role of secondary structure formation is crucial in understanding the protein folding and aggregation. In α-helices, all the backbone hydrogen bonds are used, thus they cannot aggregate into fibrils, while in β-sheets only half of the hydrogen bonds are used and thus they can easily aggregate. In order for the peptide with a significant amount of α-helices to aggregate, the α-helices must spontaneously transform into β-sheets [85].

Go models predict that about 50% of the secondary structure is formed in the unfolded sate before the cooperative folding takes place. The secondary structure in the unfolded state is not stable because in the presence of water, backbone

UNCORRECTED PROOF

can form hydrogen bonds with surrounding water molecules which are stronger than the interpeptide hydrogen bonds. The hydrogen bond C=O···H–O–H between a carbonyl group and a water molecule is about 21 kJ/mol, while the carbonyl–nitrogen hydrogen bond –C=O···H–N– is only 8 kJ/mol. Once the folding nucleus is formed and cooperative folding takes place due to hydrophobic collapse, the water molecules are expelled from the interior of the protein and the protein backbone has no other choice but to form hydrogen bonds between carbonyls and nitrogens. Thus, the secondary structure becomes stable in the folded state. This argument suggests that it is impossible to create an accurate protein model without taking into account the local environment of hydrogen bonds. More advanced models of hydrogen bonds, which can be turned on and off depending on the conformation of the protein in the vicinity of the carbonyl and amide groups, are currently being developed by several research groups.

The four-bead model that we have developed is similar to the PRIME model of Carol Hall and co-workers [86]. This model correctly reproduces the backbone geometry. It replaces each amino acid by a rigid tetrahedron of four beads: N, C_α, CO, and C_β (Fig. 5). In glycines C_β is absent. The rigidity is achieved by three covalent bonds, N–C_α =1.56 Å, C_α–CO=1.51 Å, and C_α–C_β =1.53, and three auxiliary bonds, N–CO=2.44 Å, C_β–N=2.44 Å, and C_β–CO=2.49 Å. The tetrahedra representing amino acids i and $i+1$ are linked together by the rigid planar quadrilateral (plate) $C_{\alpha,i}$–CO_i–N_{i+1}–$C_{\alpha,i+1}$ formed by the peptide bond CO_i–N_{i+1} =1.33 Å linking two amino acids together, two covalent bonds $C_{\alpha,i}$–CO_i, and N_{i+1}–$C_{\alpha,i+1}$ which participate also in the correspondent tetrahedra, auxiliary bonds $C_{\alpha,i}$–N_{i+1} =2.41 Å, CO_i–$C_{\alpha,i+1}$ =2.43 Å, and the diagonal $C_{\alpha,i}$–C_{α_i+1} =3.78 Å, which maintains the quadrilateral rigidity. Note that $C_{\beta,i}$ and $C_{\beta,i+1}$ atoms in the adjacent amino acids point in the opposite directions. The beads N, C_α, CO, and C_β are modeled by hard spheres of radii 1.69, 1.76, 1.75, and 1.54 respectively.

The amino acid tetrahedra can freely rotate (like doors around hinges) around N–C_α and C_α–CO bonds forming the two Ramachandran dihedral angles Φ and Ψ,

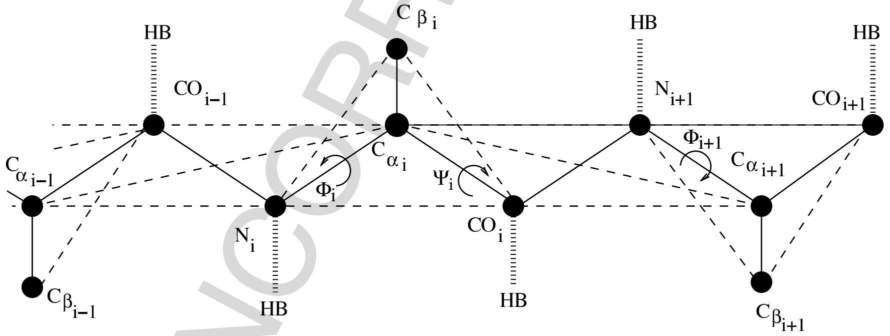


Fig. 5 The four-bead model of the protein backbone. *Bold lines* show covalent bonds, *dashed lines* show auxiliary bonds which helps maintain correct backbone geometry, and *thick broken lines* show possible hydrogen bonds

UNCORRECTED PROOF

respectively, between the plane $N_iC_{\alpha,i}CO_i$ and the adjacent quadrilaterals $C_{\alpha,i-1}CO_{i-1}N_iC_{\alpha,i}$ and $C_{\alpha,i}CO_iN_{i+1}C_{\alpha,i+1}$. The sequence of Ramachandran angles $\Psi_1, \Phi_2, \Psi_2, ...\Psi_{n-1}, \Phi_n$ completely determines the backbone conformation. The absolutely planar β-strand conformation corresponds to all $\Psi_i = 0$, $\Phi_i = 0$. Mathematically, the Ramachandran angles are determined by the sequence of the four successive backbone bond vectors, $\boldsymbol{a}_1$=$CO_{i-1}N_i$, $\boldsymbol{a}_2$=$N_iC_{\alpha,i}$, $\boldsymbol{a}_3$=$C_{\alpha,i}CO_i$, and $\boldsymbol{a}_4$=CO_iN_{i+1}:

$$\Phi = \pm \mathrm{acos}\left(\frac{[\boldsymbol{a}_2 \times \boldsymbol{a}_1] \cdot [\boldsymbol{a}_2 \times \boldsymbol{a}_3]}{|\boldsymbol{a}_2 \times \boldsymbol{a}_1||\boldsymbol{a}_2 \times \boldsymbol{a}_3|}\right),$$

where the sign coincides with the sign of $([\boldsymbol{a}_1 \times \boldsymbol{a}_3] \cdot \boldsymbol{a}_2)$ and

$$\Psi = \pm \mathrm{acos}\left(\frac{[\boldsymbol{a}_3 \times \boldsymbol{a}_2] \cdot [\boldsymbol{a}_3 \times \boldsymbol{a}_4]}{|\boldsymbol{a}_3 \times \boldsymbol{a}_2||\boldsymbol{a}_3 \times \boldsymbol{a}_4|}\right),$$

where the sign coincides with the sign of $([\boldsymbol{a}_2 \times \boldsymbol{a}_4] \cdot \boldsymbol{a}_3)$.

The hydrogen bonds between N_i and CO_j are modeled as a thin square well interaction of maximal distance $b_{\max} = 4.2$ Å and minimal distance $b_{\min} = 4.0$ Å with negative potential energy $-\epsilon_{\mathrm{HB}}$. To maintain a correct orientation of the hydrogen bond, we introduce four auxiliary bonds which appear and disappear together with the hydrogen bond (Fig. 6a). These bonds are created between the reacting beads and their neighbors in the opposite backbone: $N_iC_{\alpha,j} = r_1$, $N_iN_{j+1} = r_2$, $CO_jC_{\alpha,i} = r_3$, and $CO_jCO_{i-1} = r_4$. The energy of these bonds is determined as a step function:

$$U(r_k) = \begin{cases} \infty & r_k < d_{\mathrm{min},k} \\ \epsilon_{\mathrm{HB}} & d_{\mathrm{min},k} < r_k < d_{0,k}, \\ \epsilon_{\mathrm{HB}}/2 & d_{0,k} < r_k < d_{1,k}, \\ 0 & d_{1,k} < r_k < d_{\mathrm{max},k}, \\ \infty & r_k > d_{\mathrm{max},k}. \end{cases} \tag{1}$$

The values of $d_{\mathrm{min},k}$, $d_{0,k}$, $d_{1,k}$, and $d_{\mathrm{max},k}$ are within the range of 4.4 Å–5.6 Å and their tables are presented in [87] The total potential energy change when N_i and CO_j come to a distance $b_{\max}$ is thus $\Delta U = -\epsilon_{\mathrm{HB}} + \sum_{k=1}^{4} U(r_k)$. Obviously, $-\epsilon_{\mathrm{HB}} < \Delta U < 3\epsilon_{\mathrm{HB}}$. If one of the reacting beads already has a hydrogen bond, or if $|i-j| < 4$, or if the kinetic energy of N_i and CO_j is not sufficient to overcome the potential barrier ΔU, the hydrogen bond does not form and the beads collide as hard spheres. Otherwise, the hydrogen bond forms and the kinetic energy of the two beads is changed by $-\Delta U$. After the hydrogen bond has formed the molecular dynamics proceeds according to the general rules taking into account the discontinuities of the auxiliary bond potential until the beads N_i and CO_j again come at the distance $b_{\max}$. At this point, the change in potential energy $\Delta U' = \epsilon_{\mathrm{HB}} - \sum_{k=1}^{4} U(r_k)$ is computed and the hydrogen bond breaks if the kinetic energy of N_i and CO_j is sufficient to overcome the barrier $\Delta U'$. Thus during hydrogen-bond formation and breaking the total energy and momentum are strictly conserved. This algorithm results in a rather flexible hydrogen bond which can form if one of the reacting beads (e.g., CO) comes at any point on the surface of spherical segments of radius $b_{\max}$ surrounding another

UNCORRECTED PROOF

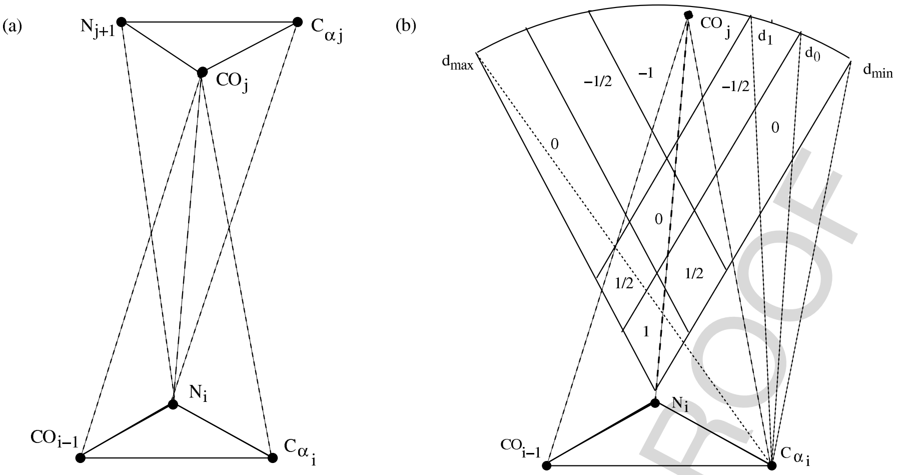
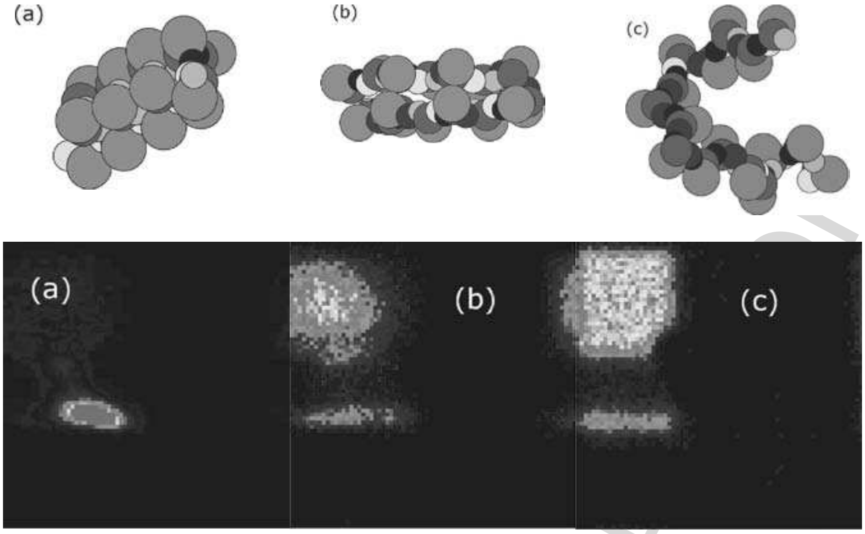


Fig. 6 (**a**) A hydrogen bond in our four-bead model. *Thick dashed lines* indicate hydrogen bond, *thin dashed lines* indicate auxiliary bonds helping to maintain hydrogen bond orientation, and *bold lines* show the covalent bonds of the backbone. See explanation in the text. (**b**) A projection of the potential energy landscape of the hydrogen bond onto the peptide plate. The *arc* indicates the sphere of radius $b_{\max}$. Whenever a CO_j bead touches this sphere within spherical segments indicated by *thin bold lines* a hydrogen bond may form. The borders of these segments appear as *straight lines* on the projection because they are produced by the intersections of the hydrogen bond sphere and the auxiliary bond spheres of various radii ($d_{\min} < d_0 < d_1 < d_{\max}$, dotted lines) with the centers at the neighboring backbone beads CO_{i-1} and $\mathrm{C}_{\alpha,i}$ which all lie in the projection plane. The numbers show the potential energy change according to (1) upon possible formation of this bond provided that the backbone of CO_j bead (not shown) has an optimal orientation. The drawing approximately reproduces the geometry for the values of $d_{\min}, d_0, d_1, andd_{\max}$ given in [87]. The similar construction must be done for the contribution of the other two auxiliary bonds

UNCORRECTED PROOF

bead (e.g., N) as indicated in Fig. 6b. These spherical segments are formed by the intersection of the sphere of radius $b_{\max}$ with the center at CO and the spheres of radii $d_{\min}$, d_0, d_1, and $d_{\max}$ surrounding the neighboring beads (e.g., CO and C_α). Figure 6b shows the projection of these segments on the plane of beads C_αCON, thus the boarders of these segments appear as straight lines. The numbers on the figure indicate the change in potential energy without taking into account the other two auxiliary bonds which are not shown $[\Delta U - U(r_1) - U(r_2)]$. This construction which belongs to Feng Ding [87] is the main difference between our four-bead model and the PRIME model of Carol Hall and co-workers [86].

We model a polyalanine of 16 amino-acids. The graphs of the Ramachandran angles for α-helix, β-hairpin, and random coil are in good agreement with the experimental ones (Fig. 7). The ground state of this model is the α-helix with potential energy $U_\alpha = -12\epsilon_{\mathrm{HB}}$, which is significantly below the energy $U_\beta = -6\epsilon_{\mathrm{HB}}$ of the β-hairpin; however the entropy S_α of the α-helix is very small comparatively to the entropy U_β of the β-hairpin which is visually apparent from the spread of the points on the Ramachandran plots. An accurate method of finding the entropy of the



AU: Pls differentiate colors in black & white.

Fig. 7 *Upper panel*: typical conformations of the α-helix (**a**), β-hairpin (**b**), and random coil (**c**). The large *gray* beads represent C_β. The *gray* scale code indicates from black to white: N, CO, C_α, C_β, hydrogen bonded and terminal N, hydrogen bonded and terminal CO. *Lower panel*: Ramachandran plots for the α-helix (**a**), β-hairpin (**b**), and random coil (**c**) states of the 4-bead model. The scales range from 0 to 360° for both Φ (horizontal axis) and Ψ (vertical axis). Each cell corresponds to a bin of 5° in Φ and Ψ. Colors in rainbow order indicate the probabilities for Φ and Ψ to belong to a certain bin from *red* (*high*) to *blue* (*low*). *Black* cells indicate zero probability

4-bead model based on the root mean square deviation of the beads from a representative conformation is presented in our original publication [87]. The potential energy of the random coil is $U_c = 0$, but its entropy S_c is even higher. We found that there is a window of temperatures when the free energies $F_x = U_x - TS_x$ of all the three states are approximately equal. In fact, at $T = 0.12\epsilon_{\text{HB}}/k_{\text{B}}$, the free energies of β-hairpin and α-helix are populated with equal probability and the model can spontaneously undergo a reversible α-helix to β-hairpin transition (Fig. 8). We found that the only pathway from an α-helix to a β-hairpin leads through a completely unfolded conformation. The frequency of these transitions are proportional to $\exp((F_\alpha - F_c)/k_{\text{B}}T)$ and $\exp((F_\beta - F_c)/k_{\text{B}}T)$, which are highly dependent on temperature. At $T = 0.13\epsilon_{\text{HB}}/k_{\text{B}}$, the β-hairpin is at equilibrium with the random coil and the α-helix is almost never observed. In contrast, at $T = 0.11\epsilon_{\text{HB}}/k_{\text{B}}$, the spontaneous transition between an α-helix and a β-hairpin is never observed and the peptide can be trapped in a metastable β-hairpin conformation. If we estimate $\epsilon_{\text{HB}} = 21$ kJ/mol, [88] this temperature range corresponds to 276–328 K, i.e. to physiological conditions. Note that the transitions between all the three states resemble the first-order phase transitions. On the other hand, the β-hairpin can be regarded as a high temperature intermediate in the folding transition to the α-helix.

UNCORRECTED PROOF

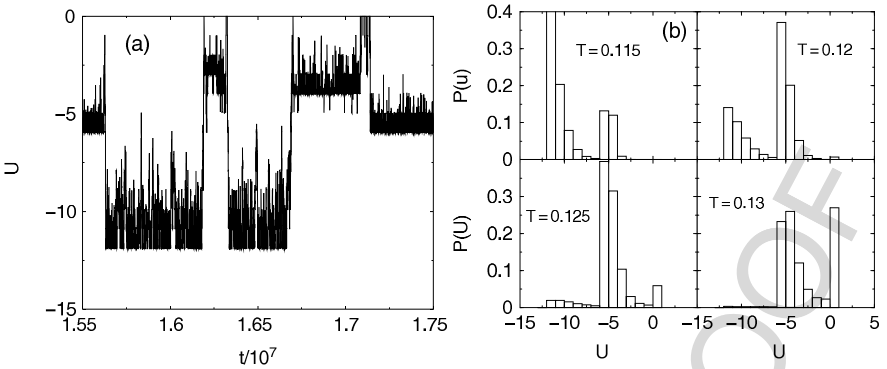


Fig. 8 (**a**) Potential energy versus time for the four-bead model of the 16-poly-alaninine peptide at $T = 0.125$ which is close to the equilibrium temperature of the β-hairpin ($U = 6\epsilon_{\mathrm{HB}}$) to α-helix transition ($U = 12\epsilon_{\mathrm{HB}}$). Several spontaneous transitions from a β-hairpin to an α-helix and back are observed. The transitions occur through complete unfolding to a random coil conformation ($U = 0$). The temperature window for which this behavior is possible is narrow. (**b**) Potential energy histograms for four different temperatures close to the glass transition. For $T = 0.115$, the random coil conformations are almost never observed, and the spontaneous transition becomes impossible in the simulation time scale window for $T < 0.11$. For $T = 0.13$ the peak of the histogram corresponding to the α-helix becomes almost invisible and for higher temperatures the transition to an α-helix never occurs

In order to study if the cooperative folding to the α-helix can be achieved by introducing side chain interactions, we include the hydrophobic interactions (H) for some of the C_β beads as square well potentials with depth ϵ_{HP} and range of attraction 6.5 Å. We model the polar side chains (P) by the hard spheres of the original diameter. We study the sequence PPHPPHHPPHPPHHPP, which was experimentally designed [86, 89, 90] to fold into the native α-helix. Indeed, when the relative strength of the hydrophobic interactions reaches $\epsilon_{\mathrm{HP}} = 0.25\epsilon_{\mathrm{HB}}$ we achieve the cooperative folding from a random coil to an α-helix, via a structureless molten globule state [66]. The folding transition is characterized by a very sharp maximum in the specific heat over the range of temperatures from 0.122 to 0.134. Thus, our model provides one of the first successful simulation of folding of a short peptide.

A missing link in understanding the amyloidogenesis of α-helix-rich proteins to β-sheet-rich fibrils is the possible presence of a metastable β-hairpin intermediate state, prone to aggregation [85]. Our results suggest a generic framework that explains why this β-hairpin intermediate is favorable in terms of free energy. Although the potential energy of a β-hairpin is higher than that of an α-helix, its entropy is also higher, thus the β-hairpin can appear as a high temperature intermediate. Also our simulation is consistent with the recent experimental results which show that the changes in the solvent can induce conformational changes in the protein. Indeed, the solvent can strongly affect the relative strength of hydrophobic interactions and hydrogen bonds. In recent years, a major progress in the DMD simulations of polyalanine and polyglutamine has been achieved by Hall and co-workers [84]. They also successfully simulated the effect of side chain interactions.

UNCORRECTED PROOF

In [91] the four-bead model has been applied to the studies of the β-amyloid aggregation. The side chains of the amyloid peptide with 40(42) residues DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV(IA) have been represented by hard spheres located at C_β, except for six glycines (G), which have been presented by only three beads. At low $T = 0.1$ the conformations are mostly α-helical, while for larger T, mostly β-stranded conformations have been observed in agreement with the simulation of polyalanine [87]. When two identical peptide are placed in the simulation box they form various β-stranded dimers linked together by parallel or antiparallel β-strands. These conformations serve as initial conformations for the all-atom MD simulations which employ the Sigma program [92] for calculation of the free energy [93] difference to determine the stability of various dimer conformations and to compare the stability of Aβ-42 and Aβ-40 dimers. The majority of the peptide conformations produced by DMD passed the stability test by the all-atom MD. This fact suggests that the four-bead model generates realistic protein conformations. However, the observed conformations do not correspond to the structure of the Aβ conformations in the Aβ fibrils [94] and do not show significant differences between the stabilities of Aβ-42 and Aβ-40 dimers. This is not strange since the four-bead model used in these studies does not take into account the amino acid–specific interactions.

Later, the four-bead model has been used to investigate the aggregation of microglobulin [95], which is the molecular basis of the complications in patients undergoing long-term hemodialysis. These studies have addressed an important practical question of the role of the disulfide bond in the aggregation pathways. The side chain interactions have been modeled by the Go interaction matrix as in [79]. The resulting trimers produced by DMD have been further studied by all-atom AMBER-8 simulations [43].

9 Simulations of Amino Acid–Specific Interactions

The example provided by the four-bead model shows that we indeed have some hope to simulate ab initio folding of short proteins and make verifiable predictions of the protein aggregation and missfolding. Also, this example shows an extreme complexity of parameterization of the DMD model, with lots of parameters which do not have much physical meaning like auxiliary interactions used in the hydrogen-bond algorithm. Another problem is that very few research groups are working on the development of the DMD code and there is no a single publicly available, well-documented source code for the DMD simulations of biomolecules. Obviously, development of such a code needs huge funding, while the granting agencies are focused on immediate biomedical applications and are reluctant to invest money in untraditional methods like DMD.

Nevertheless, some success has been achieved in this direction. In 2005, Ding et al. successfully simulate a folding transition of a trp-cage miniprotein NLYIQWLKDGGPSSGRPPPS to a 1.5 Å resolution [61]. They have developed a more detailed "$5++$"-bead model of protein with explicit carbonyl oxygens and heavy

UNCORRECTED PROOF

side chains with C_γ, C_δ beads, bifurcated side chains with two β beads, and with an additional covalent bond between the side chain and the backbone for proline. The explicit carbonyl oxygens lead to a much more realistic and simple model of the backbone hydrogen bonds than in the four-bead model. Different type of interactions have been included for salt bridges, aromatic, and aromatic–proline interactions. They have found a folding transition with an intermediate of 3.5 Å from the native state and the folded state of 1.5 Å from the NMR-resolved native state. However, as they decrease the temperature, the intermediate does not completely disappear and the distance from the NMR native state increases. This discrepancy clearly indicates that the parameters of the model are still imperfect although all the interactions have been defined from the extensive statistical analysis of the protein structures in the PDB. Another concern is that trp-cage miniprotein is a very specific one with especially rigid backbone due to four prolines. Thus, the relative success in its folding may not be transferable to other proteins.

More recently, similar model [96] has been applied for the studies of polyglutamine aggregation, which is the cause of nine human diseases including Huntington's disease. Glutamine has been modeled by six beads: four beads of the backbone C_α, C', O, N and two beads of the side chain, one bead representing methylene groups (-CH_2-CH_2-) and another representing the carboxylamine ($CONH_2$) group. The authors show that the propensity to form β-sheets increases with the length of the polyglutamine repeat. This may explain why if the repeat length exceeds a critical value of 35–40 glutamines the disease starts to develop and becomes more severe as the repeat length increases in a lifetime of a patient. The same model have been also employed to study the α-helix-to-β-sheet transition with a subsequent aggregation of a 17-residue peptide named ccβ, SIRELEARIRELELRIG [97]. Conformational changes of this small protein-like peptide can serve as a model for prion diseases.

Recent DMD studies [98, 99, 100] have been done in close collaboration with the experimental group of David Teplow and are aimed to model specific pathways of amyloid aggregation in Alzheimer's disease [101, 102]. Urbanc et al. use DMD simulations of β-amyloid aggregation using the original four-bead model with various strength of hydrophobic interactions on the C_β beads, which are intended to model the amino acid–specific interactions. The hydrophobic strength has been taken from standard hydrophobicity tables [103]. These simulations have been done with 32 peptides and show that they aggregate into micelle-like disordered oligomers of various sizes with highly hydrophobic amino acids in the center and hydrophilic amino acids forming a shell around the hydrophobic core which prevents further aggregation. Interestingly, the Aβ-42 which is genetically linked with the Alzheimer disease phenotype forms larger oligomers than Aβ-40 which lacks a highly hydrophobic isoleucine at the C-terminus. This finding is constant with the experimental results of [81, 104]. These studies also reveal a statistically predominant turn near the N-terminus of the aggregated peptides. This turn is due to glycines Gly37 and Gly38 and is stabilized by the strong hydrophobic interaction between two valines Val36 and Val39. While in Aβ40, no other nearby amino acid is involved in this turn, in Aβ42, methionine Met35 strongly interacts with isoleucine Ile41 and valines Val39 and Val40 and thus stabilizes this turn even more. This is especially significant,

UNCORRECTED PROOF

since in vitro oxidation of methionine in Aβ42 by Bitan et al. [105] reduces the aggregation propensity of Aβ42 and makes it equal to that of Aβ-40.

An attempt to build a united atom DMD model has been recently made by Borreguero et al. [106]. This model is a further development of Feng et al. [61] $5++$-bead model and takes into account all atoms except hydrogens. In collaboration with Teplow's group, they studied the conformational statistics of the central amyloid segment Aβ(21-31) which is believed to form a folded structure in the oligomers. The results of simulations predict a loop structure with a turn caused by the hydrophobic interactions between valine and a long hydrophobic part of lysine. The charged tail of lysine is competing to form a salt bridge with the aspartic acid and the glutamic acid. This salt bridge stabilizes the loop. By varying the strength of electrostatic interactions, it is possible to shift the most predominant electrostatic interaction from Glu22-Lys28 to Asp23-Lys28. This may indicate a special role which Glu22 has in familial mutations which increase the risk of Alzheimer disease. This example shows what level of molecular details can be achieved by DMD. Of course, all these predictions may not be correct because the behavior of the model highly depends on the hundreds of parameters describing the interactions. These parameters are obtained by Borreguero from the extensive studies of the PDB but still a huge effort on optimizing the interactions is needed before DMD will achieve predictive power. At present, the results of DMD can provide a useful food for thought for the experimentalists but one has to always take them with reservations. Nevertheless, I believe that with the increase of computation power and with enormous effort of devoted graduate students, DMD will soon become a routinely used predictive tool in molecular biology.

Acknowledgements I am very grateful to all the coauthors of all the joint works on DMD in which I have participated. I am obliged to A. Yu. Grosberg and A. R. Khokhlov who invited me in 1995 to make a CD for the popular book [27] *Giant Molecules*. This project led to the development of the fast and universal DMD code suitable for simulations of protein folding. I am grateful to E.I. Shakhnovich without whose expertise this work on protein folding and aggregation would be impossible. A crucial role in application of DMD particularly to protein folding belongs to the former BU graduate students N. V. Dokholyan, F. Ding, and J. Borreguero. They developed the two-bead, four-bead, and united atom models for DMD. In particular, F. Ding made significant additions to the DMD code for the hydrogen bonding. I thank Brigita Urbanc and Simcha Rimler for critical comments on the manuscript. Last but not least, I want to emphasize the role of H. E. Stanley whose interest in interdisciplinary studies made this collaboration possible. I acknowledge financial support of NSF, NIH, PRF, and anonymous private foundations. I also gratefully acknowledge the partial support of this research through the Dr. Bernard W. Gamson Computational Laboratory at Yeshiva College.

Appendix A Details of the DMD Algorithm

The structure of the DMD algorithm is the following:

(1) Find collision times of all pairs of neighboring particles and record them into collision tables
(2) Find the next collision time.

UNCORRECTED PROOF

(3) Clean up the tables from the data involving colliding particles.
(4) Move the colliding pair to the time of collision and find the new velocities after collision.
(5) Find the new collision times of the collided particles with their neighbors.
(6) Compute time averages of the properties of interest and save the data if needed.
(7) Go to Step (2)

UNCORRECTED PROOF

A.1 Find the Next Collision

Between collisions, particles move along straight lines with constant velocities. When the distance between the particles, r, becomes equal to r_k at which $U(r)$ has a discontinuity, the velocities of the interacting particles instantaneously change. The interaction time t_{ij} for two particles with coordinates $\boldsymbol{r}_i$, $\boldsymbol{r}_j$ and velocities $\boldsymbol{v}_i$, $\boldsymbol{v}_j$ satisfies the quadratic equation

$$(\boldsymbol{r}_{ij} + t_{ij}\boldsymbol{v}_{ij})^2 = R_{ij}^2,$$

where $R_{ij} = r_k$ and k depends on the initial distance between particles $\boldsymbol{r}_{ij} = \boldsymbol{r}_i - \boldsymbol{r}_j$ and their relative velocity $\boldsymbol{v}_{ij} = \boldsymbol{v}_i - \boldsymbol{v}_j$. This quadratic equation may have two positive roots, two negative roots, two roots of different signs, or no roots at all. The roots are determined by the formula

$$t_{ij} = \frac{-(\boldsymbol{v}_{ij} \cdot \boldsymbol{r}_{ij}) \pm \sqrt{(\boldsymbol{v}_{ij} \cdot \boldsymbol{r}_{ij})^2 + v_{ij}^2 \left(R_{ij}^2 - r_{ij}^2\right)}}{v_{ij}^2},$$

where the actual collision time corresponds to sign “plus” if roots have different signs or “minus” otherwise. The value of k in $R_{ij} = r_k$ is selected to minimize $t_{ij} > 0$. If there are no positive roots, it means that the particles will not interact and $t_{ij} = \infty$.

A.2 Move the Colliding Particles Forward Until a Collision Occurs

We find the next collision time

$$\delta t = \min_{i<j} t_{ij}$$

for all possible pairs of particles and propagate the system to time

$$t' = t + \delta t$$

so that

$$\boldsymbol{r'}_i = \boldsymbol{r}_i + \delta t \boldsymbol{v}_i.$$

At this moment, the distance between the centers of colliding particle-pairs becomes equal to r_k. The minimization of the t_{ij} is optimized by dividing the system into small cubic cells whose size is equal to the largest interaction distance. The collision times are computed only for particles in the nearest neighboring cells and are stored in the collision lists of the atoms belonging to this cell. After two particles collide, their future collision times with other atoms become invalid and must be removed from their collision lists and from the collision lists of the atoms in whose collision lists these particles occur. Only these affected atoms are moved to the next collision time, and the new collision times of these atoms with the collided particles are computed. The rest of the atoms in the system are not affected and stay at their positions.

In order to keep track of the atom position in space-time, each atom structure stores (besides the atom type, current coordinates, and velocity components) the update time, (i.e., the time at which the atom coordinates were last updated) and the time of leaving the cell. The collision times larger than this value are not kept. After an atom enters a new cell (which is treated as an event equivalent to the collision with other particles) its collision list is empty, and it is filled again with collisions with atoms in the new neighboring cells. After collision tables are updated, the new nearest collision time is found for each cell containing the affected atoms and then those cells participate in the binary tree sorting procedure similar to the World Cup schedule. This algorithm reduces the computational costs to $N \ln N$, where $\ln N$ comes from the tree sorting and for all practical purposes can be neglected. However, this algorithm becomes impractical when the largest interaction distance becomes greater than $1/4$ of the system box. Further improvements can be achieved computing a list of all atoms within certain distance to a given atom [33]. The program spends most of the CPU time in the calculation of the next collision times, many of which will never occur, because an atom will collide with somebody else sooner. The calculations of the future collision times during the update of the collision tables can be parallelized. But the problem of effective scalable parallelization for the DMD (to the best of my knowledge) has not yet been solved.

A.3 Implement Collision Dynamics of the Colliding Pair

Finally, we find the new velocities $\boldsymbol{v}'_i$ and $\boldsymbol{v}'_j$ after the collision. These velocities must satisfy the momentum conservation law

$$m_i\boldsymbol{v}_i + m_j\boldsymbol{v}_j = m_i\boldsymbol{v}'_i + m_j\boldsymbol{v}'_j,$$

the angular momentum conservation law

$$m_i[\boldsymbol{r}'_i \times \boldsymbol{v}_i] + m_j[\boldsymbol{r}'_j \times \boldsymbol{v}_j] = m_i[\boldsymbol{r}'_i \times \boldsymbol{v}'_i] + m_j[\boldsymbol{r}'_j \times \boldsymbol{v}'_j],$$

and the energy conservation law

UNCORRECTED PROOF

$$\frac{m_i v_i^2}{2} + \frac{m_j v_j^2}{2} + U_{ij} = \frac{m_i v_i'^2}{2} + \frac{m_j v_j'^2}{2} + U'_{ij},$$

where U_{ij} and U'_{ij} are the values of the pair potential before and after the collision, equal to $U(R_{ij} \pm \epsilon)$, depending on the direction of the initial relative velocity $\boldsymbol{v}_{ij}$, initial distance $\boldsymbol{r}_{ij}$, and the value of R_{ij}. These equations are equivalent to six scalar equations, which are sufficient to determine the six unknown components of the velocities $\boldsymbol{v'}_i$ and $\boldsymbol{v'}_j$. By introducing a new coordinate system with the origin at the center of the particle j, and the x-axis collinear with the vector $\boldsymbol{r'}_{ij}$, we construct the expressions for the velocities that satisfy the momentum and the angular momentum conservation laws:

$$\begin{aligned} \boldsymbol{v}'_i &= \boldsymbol{v}_i + A\boldsymbol{r}'_{ij} m_j, \\ \boldsymbol{v}'_j &= \boldsymbol{v}_j - A\boldsymbol{r}'_{ij} m_i, \end{aligned} \tag{2}$$

where constant A is determined from the energy conservation law:

$$A = a\frac{\pm\sqrt{1 + 2(U_{ij} - U'_{ij})(m_i + m_j)/(R_{ij}^2 a^2 m_i m_j)} - 1}{m_i + m_j} \tag{3}$$

and $a = (\boldsymbol{v}_{ij}, \boldsymbol{r}_{ij})/R_{ij}^2$. The sign “plus” in the expression for A corresponds to the motion after the collision in the same direction as before the collision, i.e. the particles penetrate into the attractive well or the soft core if they move toward each other before the collision, or leave them if they move away from each other. Note that this may happen only if the expression under the square root is positive, i.e. if there is enough kinetic energy to overcome the potential barrier:

$$\frac{R_{ij}^2 a^2 m_i m_j}{2(m_i + m_j)} \geq U'_{ij} - U_{ij}.$$

Otherwise, the reflection happens, the particles do not change their state: $U'_{ij} = U_{ij}$, and the sign in the expression for A must be “minus”.

Appendix B Calculation of Energy and Temperature

The total energy of our system is defined as:

$$E = K + U, \tag{4}$$

where K and U are the kinetic and potential energy, respectively. The kinetic energy is a sum of contributions from the individual particles

$$K = \sum_{i=1}^{N} \frac{m_i v_i^2}{2}, \tag{5}$$

UNCORRECTED PROOF

while the evaluation of the potential energy contribution involves summing over all pairs of interacting particles

$$U = \sum_{i<j} U_{ij} \ . \tag{6}$$

The temperature of the system T is calculated according to the equipartition theorem. For a d-dimensional system the instantaneous temperature can be defined as

$$T = 2K/(N\mathrm{d}k_{\mathrm{B}}) \tag{7}$$

B.1 Temperature Rescaling

In the DMD algorithm, the energy strictly conserves, so it corresponds to the microcanonical ensemble. To maintain the temperature constant or to slowly cool the system down, we use the Berendsen method [107] of velocity rescaling, multiplying all the velocities by a factor $\sqrt{T'/T}$, which is determined by

$$T' = T\,(1 - \kappa\Delta t) + \kappa\Delta t T_o \ , \tag{8}$$

where Δt is an approximately constant interval of time between two successive rescalings, T_o is the temperature of the heat bath, T is the instantaneous temperature before rescaling, T' is the instantaneous temperature after rescaling, and κ is the heat exchange coefficient. Usually, we select Δt as a time during which N collisions occur. In order to keep the old collision tables after rescaling, we actually rescale the energies of interactions and take this into account by keeping track of the ratio of the actual physical velocities and the unrescaled velocities in the computer. The inverse correction factor is applied to time. Interestingly, this correction factor exponentially inflates and may soon reach astronomical values. Once the correction factor becomes too large (e.g., 10) or too small (e.g. 0.1), we rescale velocities and time, return the interaction energies to the original value, and recalculate the collision tables from scratch.

Appendix C Calculation of Pressure

For ergodic systems, a thermodynamic average of a quantity f can be achieved in MD by averaging over a sufficiently large time Δt,

$$\langle f \rangle_{\Delta t} \equiv \frac{1}{\Delta t} \int_{t}^{t+\Delta t} f(t)\mathrm{d}t \ . \tag{9}$$

The calculation of pressure in MD has another difficulty because due to the periodic boundaries the system does not have walls which create external pressure.

UNCORRECTED PROOF

Nevertheless, the average pressure P over a long enough period of time can be effectively computed using the virial theorem:

$$P = \frac{2}{Vd}\left\langle \sum_{i=1}^{N} \frac{m_i v_i^2}{2} \right\rangle_{\Delta t} - \frac{1}{Vd}\left\langle \sum_{i=1}^{N} \boldsymbol{f}_i \cdot \boldsymbol{r}_i \right\rangle_{\Delta t} \tag{10}$$

where $\boldsymbol{f}_i$ is the force acting on particle i from all other particles. Note that $\sum_{i=1}^{N} m_i v_i^2/2$ is by definition (7) equal to $\mathrm{d}k_{\mathrm{B}}NT/2$ and

$$P = \frac{Nk_{\mathrm{B}}}{V}\langle T\rangle_{\Delta t} - \frac{1}{Vd}\left\langle \sum_{i=1}^{N} \boldsymbol{f}_i \cdot \boldsymbol{r}_i \right\rangle_{\Delta t} . \tag{11}$$

When the system has walls, this equation gives the value of the pressure acting from the walls to the system. In the absence of walls, it gives the value of the internal pressure in the system. Thus, this equation provides the basis for the computation of pressure in molecular dynamics simulations.

In discrete molecular dynamics, the force $\boldsymbol{f}_i$ is equal to zero except at the moments of collision with other particles, when it is equal to infinity. We count all the collisions of a given particle i with a given particle j that occur in the time interval from t to $t+\Delta t$, using index $K_{ij} = 1, 2, 3, \ldots$ We denote the times of these collisions $t_{K_{ij}}$ and the change in momentum of particle i at the moments $t_{K_{ij}}$ as

$$\Delta \boldsymbol{p}_{K_{ij}} = m_i[\boldsymbol{v}_i(t_{K_{ij}} + \epsilon) - \boldsymbol{v}_i(t_{K_{ij}} - \epsilon)] , \tag{12}$$

where ϵ is an infinitesimally small value. Since the force acting on the particle i is the derivative of momentum with respect to time,

$$\boldsymbol{f}_i = \sum_{j=1}^{N} \sum_{K_{ij}} \Delta \boldsymbol{p}_{K_{ij}} \delta(t - t_{K_{ij}}) , \tag{13}$$

where $\delta(t - t_{K_{ij}})$ is a Dirac δ-function and the sum over K_{ij} is taken over all collisions between particle i and j during time interval $(t, t+\Delta t)$.

Integration involved in the averaging over time [see (9)] eliminates δ-functions and we obtain

$$P = \frac{Nk_{\mathrm{B}}}{V}\langle T\rangle_{\Delta t} - \frac{1}{\Delta t V d} \sum_{i=1}^{N} \sum_{K_{ij}} \sum_{j=1}^{N} (\Delta \boldsymbol{p}_{K_{ij}} \cdot \boldsymbol{r}_i) . \tag{14}$$

Finally, we can count all the collisions that occur in interval $(t, t+\Delta t)$ by index ℓ. Each collision is specified by the particles $i(\ell)$ and $j(\ell)$ involved in the collision $(i < j)$ and is counted twice in the sum of (14)—the first time when i is from the first sum and the second when i is from the second sum. According to momentum conservation, $\Delta \boldsymbol{p}_{i(\ell)} = -\Delta \boldsymbol{p}_{j(\ell)}$. Thus we rewrite (14) as

$$P = \frac{Nk_{\mathrm{B}}}{V}\langle T\rangle_{\Delta t} - \frac{1}{\Delta t V d} \sum_{\ell} \{\Delta \boldsymbol{p}_{i(\ell)}(t_\ell) \cdot [\boldsymbol{r}_i(t_\ell) - \boldsymbol{r}_j(t_\ell)]\} , \tag{15}$$

UNCORRECTED PROOF

where the sum is taken over all collisions ℓ that occur at moments t_ℓ during the time interval Δt. Finally, taking into account Eqs. (2) and (3),

$$P = \frac{Nk_{\mathrm{B}}}{V}\langle T\rangle_{\Delta t} - \frac{1}{\Delta t V d}\sum_{\ell} m_{i(\ell)} m_{j(\ell)} R^2_{ij(\ell)} A_\ell, \tag{16}$$

where A_ℓ is given by expression (3) for $i = i(\ell)$, $j = j(\ell)$.

The DMD algorithm allows also the constant pressure simulations. The easiest way to do it is to apply Berendsen barostat, analogous to Berendsen thermostat, with the difference that now all the coordinates are rescaled periodically by a factor $(1+\eta)$, where $\eta < \epsilon$ is a small quantity, proportional to the difference between the average pressure over this period of time and the desired pressure P_0 of the barostat. The problem is that after the rescaling some pairs of particles may occur in the zone of the infinite potential, since after rescaling they may become closer than their hardcore interaction distance or outside the range of the permanent bond. To solve this problem, we add an inner hard-core which constitute $1-\epsilon$ of the true hard core, and the outer bond distance which is larger than actual bond distance by factor of $1+\epsilon$. The particles that appear to be within the gap between the actual and the inner hardcore cannot go inside the inner hardcore but can freely move through the outer hardcore. The analogous algorithm works for the bonds. No new rescaling takes place before all the particles become outside the actual hardcore. Of course, after rescaling, all the collision tables must be reconstructed from scratch. So rescaling should not be done too often, otherwise the simulation will significantly slow down.

References

1. A. Smith: Nature **426**, 883 (2003)
2. C. M. Dobson: Nature **426**, 884 (2003)
3. D. J. Selkoe: Nature **426**, 900 (2003)
4. B. J. Alder, T. E. Wainwright: J. Chem. Phys. **31**, 459 (1959)
5. M. R. Sadr-Lahijany, A. Scala, S. V. Buldyrev, H. E. Stanley: Phys. Rev. Lett. **81**, 4895 (1998)
6. G. Franzese, G. Malescio, A. Skibinsky, S. V. Buldyrev, H. E. Stanley: Nature **409**, 692 (2001)
7. S. V. Buldyrev, H. E. Stanley: Physica A **330**, 124 (2003)
8. A. Skibinsky, S. V. Buldyrev, G. Franzese, G. Malescio, H. E. Stanley: Phys. Rev. E **69**, 61206 (2004)
9. P. Kumar, S. V. Buldyrev, F. Sciortino, E. Zaccarelli, H. E. Stanley: Phys. Rev. E **72**, 021501 (2005)
10. Z. Yan, S. V. Buldyrev, N. Giovambattista, H. E. Stanley: Phys. Rev. Lett. **95**, 130604 (2005)
11. P. A. Netz, S. V. Buldyrev, M. C. Barbosa, H. E. Stanley: Phys. Rev. E **73**, 061504 (2006)
12. L. Xu, S. V. Buldyrev, C. A. Angell, H. E. Stanley: Phys. Rev. E **74**, 031108 (2006)
13. D. C. Rapaport: J. Phys. A **11**, L213 (1978)
14. D. C. Rapaport: J. Chem. Phys. **71**, 3299 (1979)
15. Y. Zhou, C. K. Hall, M. Karplus: Phys. Rev. Lett. **77**, 2822 (1996)
16. Y. Zhou, M. Karplus, J. M. Wichert, C. K. Hall: J. Chem. Phys. **107**, 10691 (1997).
17. S. W. Smith, C. K. Hall, B. D. Freeman: J. Comp. Phys. **134**, 16 (1997)
18. N. V. Dokholyan, E. Pitard. S. V. Buldyrev, H. E. Stanley: Phys. Rev. E **65**, R030801 (2002)

UNCORRECTED PROOF

19. S. V. Buldyrev, P. Kumar, P. G. Debenedetti, P. J. Rossky, and H. E. Stanley, Proc. Natl. Acad. Sci. USA **101**, 21077 (2007).
20. G. Foffi, K. A. Dawson, S. V. Buldyrev, F. Sciortino, E. Zaccarelli, P. Tartaglia: Phys. Rev. E **65**, 050802 (2002)
21. E. Zaccarelli, S. V. Buldyrev, E. La Nave, A. J. Moreno, I. Saika-Voivod, F. Sciortino, P. Tartaglia: Phys. Rev. Lett. **94**, 218301 (2005)
22. A. J. Moreno, S. V. Buldyrev, E. La Nave, I. Saika-Voivod, F. Sciortino, P. Tartaglia, E. Zaccarelli: Phys. Rev. Lett. **95**, 157802 (2005)
23. C. Davis, H. Nie, N. V. Dokholyan: Phys. Rev. E, **75**, in press (2007)
24. S. Sharma, F. Ding, N. V. Dokholyan: Biophysical Journal, **92**, 1457 (2007)
25. F. Ding, N. V. Dokholyan: Trends Biotech. **23**, 450 (2005).
26. B. Urbanc, J. M. Borreguero, L. Cruz, H. E. Stanley: Methods in Enzymology **412**, 314 (2006).
27. A. Yu. Grosberg, A. R. Khokhlov: *Giant Molecules* (Academic Press, 1997)
28. Center for Polymer Studies, Boston University: Virtual Molecular Dynamics Laboratory, http://cps.bu.edu/education/vmdl/ (2007)
29. D. C. Rapaport: J. Comput. Phys. **34**, 184 (1980)
30. M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, New York, 1989)
31. D. C. Rapaport: *The Art of Molecular Dynamics Simulation* (Cambridge University Press: Cambridge, 1997)
32. D. Frenkel, B. Smit: *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, San Diego, 1996)
33. A. Donev, S. Torquato, F. H. Stillinger: J. Comp. Phys. **202**, 737 (2005)
34. A. Donev, S. Torquato, F. H. Stillinger: J. Comp. Phys. **202**, 765 (2005)
35. C. De Michele, A. Scala, R. Schilling, F. Sciortino: J. Chem. Phys. **124**, 104509, (2006)
36. C. De Michele, S. Gabrielli, P. Tartaglia, F. Sciortino: J. Phys. Chem. B **110**, 8064 (2006)
37. Y. Zhou, M. Karplus: Proc. Natl. Acad. Sci. USA **94**, 14429 (1997)
38. E. A. Jagla: Phys. Rev. E **58**, 1478 (1998)
39. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus: J. Comp. Chem. **4**, 187 (1983)
40. A. D. MacKerell, Jr., B. Brooks, C. L. Brooks, III, L. Nilsson, B. Roux, Y. Won, M. Karplus: CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. In: *The Encyclopedia of Computational Chemistry*, vol. 1, ed. P. v. R. Schleyer et al. (John Wiley & Sons: Chichester, 1998) pp. 271–277
41. D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. C. Berendsen: J. Comp. Chem. **26**, 1701 (2005)
42. L. Kale, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, K. Schulten: J. Comp. Phys. **151**, 283312 (1999)
43. D. A. Case, T. E. Cheatham, III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, R. Woods: J. Computat. Chem. **26**, 1668 (2005).
44. S. J. Plimpton: J. Comp. Phys. **117**, 1 (1995)
45. H. D. Nguyen, C. K. Hall: Proc. Natl. Acad. Sci. USA **101**, 16180 (2004)
46. M. Karplus: Fold. Des. **2**, 569 (1997)
47. E. I. Shakhnovich: Chem. Rev. **106**, 1559 (2006)
48. R. R. Shearer: Art Journal, **55**, 64 (1996).
49. K. B. Zeldovich, P. Chen, B. E. Shakhnovich, E. I. Shakhnovich: PLoS Comp. Bio. in press (2007)
50. S. Miyazawa, R. L. Jernigan: J. Mol. Biol. **256**, 623 (1996).
51. N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, E. I. Shakhnovich: Folding & Design **3**, 577 (1998).
52. H. Taketomi, Y. Ueda, N. Go: Int. J. Peptide Protein Res. **7**, 445 (1975)
53. N. Go, H. Abe: Biopolymers **20**, 991 (1981)
54. H. Abe, N. Go: Biopolymers **20**, 1013 (1981)

AU: Please update Ref. [23].

AU: Please update Ref. [49].

UNCORRECTED PROOF

55. N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, E. I. Shakhnovich: J. Mol. Biol. **296**, 1183 (2000)
56. A. Scala, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, E. I. Shakhnovich: Phys. Rev. E **63** 032901 (2001)
57. Y. Zhou, M. Karplus: Nature **401**, 400 (1999)
58. Y, Zhou, M. Karplus: J. Mol. Biol. **293**, 917 (1999)
59. V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E.J. Sorin, B. Zagrovic: Biopolymers **68**, 91 (2003)
60. A. R. Fersht: Curr. Opin. Struc. Biol. **7**, 3 (1997)
61. F. Ding, S. V. Buldyrev, N. V. Dokholyan: Biophys. J. **88**, 147 (2005)
62. J. M. Borreguero, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, E. I. Shakhnovich: J. Mol. Biol. **318**, 863 (2002)
63. F. Ding, W. Guo, N. V. Dokholyan, E. I. Shakhnovich, J.-E. Shea: J. Mol. Biol. **350**, 1035 (2005)
64. J. M. Borreguero, F. Ding, S. V. Buldyrev, H. E. Stanley, N. V. Dokholyan: Biophys. J. **87**, 521 (2004).
65. H. Jang, C. K. Hall, Y. Zhou: Biophys. J. **83**, 819 (2002)
66. O. B. Ptitsyn: Adv. Protein Chem. **47**, 83 (1995)
67. F. Ding, R. K. Jha, N. V. Dokholyan, Structure **13**, 1047 (2005)
68. K. Lum, D. Chandler, J. Weeks: J. Phys. Chem. B **103**, 4570 (1999).
69. F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, E. I. Shakhnovich: Biophys. J. **83**, 3525 (2002)
70. S. Sharma, F. Ding, H. Nie, D. Watson, A. Unnithan, J. Lopp, D. Pozefsky, N. V. Dokholyan: Bioinformatics **22**, 2693 (2006)
71. C. Clementi, H. Nymeyer, J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000)
72. A. R. Lam, J. M. Borreguero, F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, E. Shakhnovich: J. Mol. Biol. **373**, 1348 (2007).
73. A. R. Fersht: Proc. Natl. Acad. Sci. USA. **97**, 14121 (2000)
74. Y. Chen, N. V. Dokholyan, J. Biol. Chem. **281**, 29148 (2006)
75. J. W. Kelly: Curr. Opin. Struct. Biol. **8**, 101 (1998)
76. S. Y. Tan, M. B. Pepys: Histopathology **25**, 403 (1994)
77. F. Ding, N. V. Dokholyan, S.V. Buldyrev, H. E. Stanley, E. I. Shakhnovich: J. Mol. Biol. **324**, 851 (2002)
78. F. Ding, K. C. Prutzman, S. L. Campbell, N. V. Dokholyan: Structure, **14**, 5 (2006).
79. S. Peng, F. Ding, B. Urbanc, S. V. Buldyrev, L. Cruz, H. E. Stanley, N. V. Dokholyan: Phys. Rev. E **69**, 041908 (2004)
80. D. M. Walsh, I. Klyubin, J. V. Fadeeva, W. K. Cullen, R. Anwyl, M. S. Wolfe, M. J. Rowan, D. J. Selkoe: Nature **416**, 535 (2002)
81. G. Bitan, M. D. Kirkitadze, A. Lomakin, S. S. Vollers, G. B. Benedek, D. B. Teplow: Proc. Natl. Acad. Sci. USA **100**, 330 (2003)
82. H. D. Nguyen, C. K. Hall: Biophys. J. **87**, 4122 (2004)
83. H. D. Nguyen, C. K. Hall: J. Biol. Chem. **280**, 9074 (2005)
84. A. J. Marchut, C. K. Hall: Biophys. J. **90**, 4574 (2006)
85. J. S. Richardson, D. C. Richardson: Proc. Natl. Acad. Sci. USA **99**, 2754 (2002)
86. A. V. Smith, C. K. Hall: Proteins Struct. Funct. Genet. **44**, 344 (2001)
87. F. Ding, J. M. Borreguero, S. V. Buldyrev, H. E. Stanley, N. V. Dokholyan: Proteins Struct. Funct. Genet. **53** 220 (2003)
88. B. Honig, A. S. Yang: Adv. Protein Chem. **46**, 2758 (1995).
89. S. P. Ho, W. F. DeGrado: J. Am. Chem. Soc. **109**, 6751 (1987)
90. Z. Guo, D. Thirumalai: J. Mol. Biol. **263**, 323 (1996)
91. B. Urbanc, L. Cruz, F. Ding, D. Sammond, S. Khare, S. V. Buldyrev, H. E. Stanley, N. V. Dokholyan: Biophys. J. **87** 2310 (2004)
92. J. Hermans, R. H. Yun, J. Leech, D. Cavanaugh: Sigma documentation, University of North Carolina (1994). http://hekto.med.unc.edu:8080/HERMANS/software/SIGMA/index.html

UNCORRECTED PROOF

93. Y. N. Vorobjev, J. Hermans: Biophys. Chem. **78**, 195 (1999).
94. A. T. Petkova, Y. Ishii, J. J. Balbach, O. N. Antzutkin, R. D. Leapman, F. Delaglio, R. Tycko: Proc. Natl. Acad. Sci. USA **99**, 16742 (2002).
95. Y. Chen, N. V. Dokholyan: J. Mol. Biol. **354**, 473 (2005).
96. S. D. Khare, F. Ding, K. N. Gwanmesia, N. V. Dokholyan, PLoS Comp. Biol. **1**, e30 (2005).
97. F. Ding, J. J. LaRocque , N. V. Dokholyan, J. Biol. Chem. **280**, 40235 (2005).
98. B. Urbanc, L. Cruz, S. Yun, S. V. Buldyrev G. Bitan, D. B. Teplow, H. E. Stanley, "In Silico Study of Amyloid Beta Protein Folding and Oligomerization," Proc. Natl. Acad. Sci. **101**, 17345–17350 (2004).
99. S. Yun, B. Urbanc, L. Cruz, G. Bitan, D. B. Teplow, H. E. Stanley: Biophys. J. **92**, 4064 (2007).
100. A. Lam, B. Urbanc, J. M. Borreguero, N. D. Lazo, D. B. Teplow, H. E. Stanley: Discrete Molecular Dynamics Study of Alzheimer Amyloid β-protein, *Proceedings of The 2006 International Conference on Bioinformatics & Computational Biology*, CSREA Press, Las Vegas, Nevada, 322–328 (2006).
101. B. Urbanc, L. Cruz, D. B. Teplow, H. E. Stanley, Current Alzheimer Research **3**, 493 (2006).
102. D. B. Teplow, N. D. Lazo, G. Bitan, S. Bernstein, T. Wyttenbach, M. T. Bowers, A. Baumketner, J.-E. Shea, B. Urbanc, L. Cruz, J. Borreguero, H. E. Stanley: Account of Chemical Research **39**, 635 (2006).
103. J. Kyte, R. F. Doolittle: J. Mol. Biol. **157**, 105 (1982)
104. G. Bitan, A. Lomakin, D. B. Teplow: J. Biol. Chem. **276**, 35176 (2001)
105. G. Bitan, B. Tarus, S. S. Vollers, H. A. Lashuel, M. M. Condron, J. E. Straub, D. B. Teplow: J. Am. Chem. Soc. **125**, 15359 (2003)
106. J. M. Borreguero, B. Urbanc, N. D. Lazo, S. V. Buldyrev, D. B. Teplow, H. E. Stanley, Folding events in the 21-30 region of amyloid-beta-protein (A beta) studied in silico,6020 Proc. Natl. Acad. Sci. **102**, 6015 (2005)
107. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. Di Nola, J. R. Haak: J. Chem. Phys. **81**, 3684 (1984)

UNCORRECTED PROOF