Effect of single-point sequence alterations on the aggregation propensity of a model protein

Dusan Bratko[1,2*], Troy Cellmer[2], John M. Prausnitz[2,3] and Harvey W. Blanch[2]

[1]Department of Chemistry, Virginia Commonwealth University, Richmond, VA 23284
[2]Department of Chemical Engineering, University of California, Berkeley, CA 94720
[3]Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

September 16, 2005

Sequences of contemporary proteins are believed to have evolved through process that optimized their overall fitness including their resistance to deleterious aggregation. Biotechnological processing may expose therapeutic proteins to conditions that are much more conducive to aggregation than those encountered in a cellular environment. An important task of protein engineering is to identify alternative sequences that would protect proteins when processed at high concentrations without altering their native structure associated with specific biological function. Our computational studies exploit parallel tempering simulations of coarse-grained model proteins to demonstrate that isolated amino-acid residue substitutions can result in significant changes in the aggregation resistance of the protein in a crowded environment while retaining protein structure in isolation. A thermodynamic analysis of protein clusters subject to competing processes of folding and association shows that moderate mutations can produce effects similar to those caused by changes in system conditions, including temperature, concentration, and solvent composition that affect the aggregation propensity. The range of conditions where a protein can resist aggregation can therefore be tuned by sequence alterations although the protein generally may retain its generic ability for aggregation.

*dnb@berkeley.edu

I. INTRODUCTION

Abnormal protein aggregation, resulting in insoluble, biologically inactive agglomerates, represents a serious problem in production, formulation, processing and storage of protein drugs[1]. *In vivo*, the occurrence of ordered fibrillar aggregates is associated with several debilitating diseases including Alzheimer's and Parkinson's disease, Bovine Spongiform Encephalopathy (BSE) (mad cow disease), Creutzfeldt-Jacob's disease, and amyotrophic lateral sclerosis (ALS)[2]. Biotechnological concerns and efforts toward prevention or cure of neurodegenerative diseases continue to motivate extensive experimental and computational research aimed at identifying system properties that can be tuneded to suppress or control aggregation. These properties include solvent composition[3], presence of molecular chaperones or aggregation inhibitors[4], as well as mutations[5] affecting protein ability to aggregate[6,7]. Two classes of mutations are of potential interest. In biotechnology and biomedical research, emphasis is on minimally intrusive substitutions that can prevent or slow aggregation while preserving the function of the protein. In materials science, on the other hand, there is emerging interest in designing novel nanomaterials comprising, or templated by, fibrillar protein aggregates[8-11]. Laboratory screening of a large number of protein variants is, however, very expensive and time consuming. Computer-assisted screening of potential mutations may significantly reduce the number of experimental sequences to be examined.

While state-of-the-art modeling techniques do not suffice for full-atom simulations of multi-chain protein systems on a practically relevant time scale, coarse-grained protein models have provided useful insights into aggregation mechanisms and can suggest guidelines to control the aggregation propensity of a protein[12-38]. Studies of aggregating proteins with Go residue potentials highlighted strong correlations between binding states of adjacent protein residues[23].

According to this previous work, a change in the binding affinity of a selected residue can noticeably modulate the probabilities of intra-molecular and inter-protein bonds involving *neighboring* residues, with the effect propagating several residue lengths along the chain contour in both directions[23]. In accord with experimental observations, complemented by single-chain modeling[16], restricted mutations in the protein sequence should therefore be able to cause a pronounced effect on both protein folding and its aggregation propensity.

To test these ideas by probing concrete sequence substitutions, however, requires models extending beyond the Go representation wherein the monomers lack association with specific amino-acid residues. In the present work, we exploit a variation[39] of the Miyazawa-Jernigan (MJ) interaction matrix originally derived[40] from a statistical analysis of contact residue-residue probabilities in a large number of protein structures available in the Protein Data Bank. A two-parameter modification of the Miyazawa-Jernigan matrix proposed by Leonhard *et al.* yields an improved protein-like behavior of model polypeptides through a systematic renormalization of residue-solvent interactions[39,41,42]. In particular, the proposed optimization, applicable to a broad spectrum of studied sequences, improves protein folding rates, folding cooperativities, and aggregation resistance, bringing the behavior of model proteins closer to that typically observed in experimental systems[39].

We consider a 64-residue sequence, previously optimized with respect to folding in isolation, and modifications introduced through single-point changes in the protein's primary structure. Amino-acid residue substitutions, expected to produce significant changes in aggregation propensity are chosen based on residue contact maps and calculated correlations between the residue binding state and the overall proximity to protein native structure. A detailed description of our criteria for rational selection of attempted mutations is given in Section III.1.

Only mutations that preserve the original native structure are considered. Single residue substitutions that have a relatively small effect on the protein in isolation are found to strongly enhance aggregation at moderate concentrations. Because we sample only sequences with identical folded conformations, it is obvious that sequence engineering may produce notable improvements in aggregation resistance when the starting primary structure is not optimized; such improvement can be achieved without compromising the function of the protein in question. In view of highly specific recognition mechanisms implicated in protein folding as opposed to relatively less selective inter-protein binding, the observed strong effect of mutations on the competition between folding and association can be predominantly attributed to subtle stabilization or destabilization of the protein native state. This view is supported by the results of a thermodynamic analysis of multiple-chain folding landscapes for the original sequence and its variants presented below.

II. MODEL AND METHODS

II.1 Model

To reduce the considerable computational costs of multi-chain protein simulations, we describe the aqueous protein solution using a lattice-model with renormalized[39] Mijazawa-Jernigan residue-residue interactions. This representation secures the essential resolution of the amino-acid alphabet, and is sufficiently efficient to produce qualitatively meaningful insights within an accessible simulation time[39,41,42]. Protein molecules are described as self-avoiding amino-acid residue chains with isotropic inter-residue potentials extracted from a statistical analysis of known protein structures collected in the Brookhaven Protein Data Bank. Presuming strong screening of coulombic interactions[43-45], the quasi-chemical model used in the derivation

of the MJ potential limits the interactions between the beads to pairs occupying neighboring lattice sites. According to Leonhard and coauthors[39], the performance of the MJ model can be improved by a two-parameter renormalization of pair potentials. The proposed modification, originally formulated within the context of the Ising model with explicit solute-solute (*ij*), solute-solvent (*i0*), and solvent-solvent (*00*) interactions[39,41,42], considers only changes in solute-solvent (*i0*) terms. As we have shown recently[33], these changes can be incorporated into the equivalent implicit model that absorbs all solvent effects into effective (solvent-averaged) solute-solute potentials. This procedure replaces the original MacMillan-Mayer potentials e_{ij} among amino-acid residues i and j (collected in Table V of Miyazawa and Jernigan[40]) by the renormalized interactions $_{ij}$ given by:

$$
\begin{aligned}
v_{ij} = e_{ij} \quad & \frac{1}{2}(1 \quad C_s)(e_{ii} + e_{jj}) \quad 2 \quad \frac{C_s}{n}^{n} e_{ii} \quad \text{if} \quad |(r_{i,m} \quad r_{j,m})| = a \\
\text{and} \quad & v_{ij}(r_{i,m} \quad r_{j,m}) = 0 \quad \text{if} \quad |(r_{i,m} \quad r_{j,m})| > a \quad \text{or} \quad |i \quad j| = 1 \\
& v_{ij}(r_{i,m} \quad r_{j,m}) = \quad \text{if} \quad |(r_{i,m} \quad r_{j,m})| = 0,
\end{aligned}
\qquad (1)
$$

Here, $0 \quad C_s \quad 1$, and are the two adjustable model parameters affecting the residue selectivity and nonspecific attraction, respectively[39]. The summation over i is performed over all twenty amino-acid residue types. The choice $C_s = 1$ and $= \frac{C_s}{2n}^{n} e_{ii}$ preserves the original MJ potentials. Deviations of C_s from unity reduce apparent differences in residue hydrophobicities, while $- \frac{C_s}{2n}^{n} e_{ii}$ represents a baseline shift of effective potentials for all residue pairs. A detailed analysis is available elsewhere [33,39,41,42].

In our present calculation, we consider the 64-mer sequence *KEKSTAGRVASGVLDSVACGVLGDIDTLQGSPIAKLKTFYGNKFNDVEASQAHMIPNYTLPE* [41,42] or its variations obtained by single-residue substitutions. We use the following values of the two adjustable parameters, C_s =0.2 and $= 0.16$ because they were shown[41,42] to optimize the observed protein-like behavior of the selected model polypeptide. The lowest temperature we use, $T=T_0$, is set to 0.375, which corresponds to 0.858 T_m, where T_m is the melting (thermal unfolding) temperature of the isolated protein. Our implicit solvent simulations give melting temperature T_m=0.437, in good agreement with T_m=0.43 obtained using the explicit-solvent model[39]. Fig. 1 of ref.[33] compares residue hydrophobicities (or hydrophilicities) in the rescaled model quantified in terms of pair potentials between identical residues, $v_{ii}(r_{ii} = a)$. In analogy with our previous work[19,23,33], we express the energy function of M interacting chains, each containing N monomer units as a sum of intra-molecular interactions, V_m, and inter-molecular interactions, V_{mn}:

$$V(r_{MN}) = \sum_{m=1}^{M} V_m(r_{N;m}) + \sum_{m=1}^{M-1} \sum_{n=m+1}^{M} V_{mn}(r_{N;m}, r_{N;n})$$

$$V_m(r_{N;m}) = \sum_{i=1}^{N-2} \sum_{j=i+1}^{N} v_{ij}(r_{i,m} - r_{j,m}) \qquad (2)$$

$$V_{mn}(r_{N;m}, r_{N;n}) = \sum_{i=1}^{N} \sum_{j=1}^{N} v_{ij}(r_{i,m} - r_{j,n})$$

Above, lattice step a equals the monomer length, and $\boldsymbol{r}_{i,m}$ is the position of the i-th bead of chain m. Therefore, the 3(MN-1) dimensional vector $\boldsymbol{r}_{MN}$ completely describes the configuration of the system.

II.2 Simulation

Folding behavior of individual chains is considered using canonical (*N,V,T*) Monte Carlo simulations. Protein concentration, measured in terms of volume fraction $_V=NM/L^3$, is controlled through number, *M*, and length, *N*, of the chain, and the size of the simulation box. Boundary effects are taken into account through minimum-image periodic conditions[46]. Simulation moves[47] include displacements end beads, corner flips, and crankshaft moves of bead pairs located at the bottom of a U turn. Further, we allow slithering-snake reptation moves[48] and translations of chains or groups of chains. Details of our (*N,V,T*) simulation are given in ref.[19].

To generate multi-temperature results needed in the weighted histogram analysis (see below), and to mitigate local trapping on the rugged free energy landscape of aggregating proteins, some of the single-chain and *all* multi-chain simulations were performed using the parallel tempering technique described earlier[23]. This technique facilitates barrier crossings by sampling several replicas of a given model system at slightly different temperatures. During simulation, swaps between adjacent temperature levels are attempted periodically with probabilities that preserve canonical (Boltzmann) statistics. For systems trapped in local minima, escape is facilitated during the time spent at an elevated temperature. In our simulations, temperature swaps were attempted after each cycle of *MN* attempted displacements (pass). The attempted swap of systems *i* and *j* between temperatures T_m and T_n, was accepted with the probability[49]:

$$p_s = \min\{1, \quad \exp[\ (\frac{1}{k_B T_n} \quad \frac{1}{k_B T_m})(V_i \quad V_j)]\} \tag{3}$$

Based on empirical considerations[23], we typically used six replicas at (reduced) temperature levels ranging from 1 to 1.3 and swapping acceptances between 10-30%. Similar acceptances have been reported in recent replica simulations of peptide aggregation in a continuum representation[50]. In our reduced units, temperature is expressed in units T_0, energy in units $k_B T_0$, and distance in lattice step length, a.

II.3 Weighted Histogram Analysis

The Weighted Histogram Analysis Method (WHAM)[51] was used to analyze simulation data. WHAM minimizes the error in the density-of-states function and facilitates calculation of free-energy surfaces. In each simulation, six quantities were monitored: the total system potential energy V, the total number of intra-protein contacts N_{intra}, the total number of inter-protein contacts N_{inter}, the contribution to the overall potential from interactions between beads on the same chain V_{intra}, the contribution to the overall potential from interactions between beads on different chains V_{inter}, and the average radius of gyration R_g^2. For the remainder of the report R_g^2, is referred to as R_g.

Calculating the density-of-states function ( ) for six quantities is computationally impractical. Thus, it was necessary to calculate several , each a function of different thermodynamic parameters. A generic description of is

$$(V, \ _1, \ _2) = \frac{\sum_{j=1}^{k} N_k(V, \ _1, \ _2)}{\sum_{j=1}^{k} n_j \exp(\ f_j \ \ _j V)} \tag{4}$$

where $_1$ and $_2$ are any of the parameters mentioned above, N_k is the number of occurrences for samples with (V, $_1$, $_2$), f_j is the free energy of simulation j, is $1/k_bT$, k is the number of simulations, and n_j is the number of samples from simulation j. Free energies were calculated by solving the following two equations self-consistently

$$P(V, \ _1, \ _2) = \frac{\sum_{i=1}^{k} N_i(V, \ _1, \ _2)\exp(\ V)}{\sum_{j=1}^{k} n_j \exp(\ f_j \ \ _j V)} \tag{5}$$

$$\exp(\ f_k) = \sum_{V, \ _1, \ _2} P_k(V, \ _1, \ _2) \tag{6}$$

where P is the probability of observing a state with (V, $_1$, $_2$). Thermodynamic averages are then calculated from

$$\left\langle \ _1 \right\rangle = \frac{\sum_{V, \ _1} (\ _1) * \ (V, \ _1, \ _2)\exp(\ V)}{\sum_{V, \ _1, \ _2} (V, \ _1, \ _2)\exp(\ V)} \tag{7}$$

using $_1$ as an example. Free energies are given by

$$F(\ _1, \ _2) == \ k_bT \ln\left\{P(\ _1, \ _2)\right\}$$

$$P(\ _1,\ _2) = \sum_V P(V,\ _1,\ _2) \quad . \tag{8}$$

III. RESULTS AND DISCUSSION

III.1 Sequence selection

Below we present a comparison between the folding and aggregation behavior of three sequences: the 'wild type' sequence (WT) characterized in our earlier studies, and two variations obtained by a single amino-acid substitution on the WT sequence. The original (WT) sequence (specified above) has been designed[41,42] using a sequence-annealing procedure described earlier[39], and the substitutions were chosen subject to the condition that they preserve the original native-state conformation. Operating under this constraint, the balance between the folded and aggregated states can be adjusted by residue substitutions that affect the stability of the native state, the aggregates, or both. Thermodynamics of aggregated states is not as sequence-sensitive as that of the native structure. A major effect of the substitutions on the competition between folding and aggregation is therefore associated with changes in the stability of the native conformation. Targeting residues whose bonding states are most strongly correlated with the proximity to the native state maximizes this effect. For every residue, the magnitude of this correlation can be quantified in terms of function *C(i)*

$$C(i) = \sum_{j\ i} < h(i)\ h(j) > \ - < h(i) > \sum_{i\ j} < h(j) > \tag{9}$$

Following the formalism introduced elsewhere[23], we define the residue bonding state, *h(i)*, as the fraction of realized native inter-residue bonds involving the specified residue *i*. Fig. 1 compares correlation functions *C(i)* for all 64 residues in the WT sequence at T=1.0. The three-dimensional

structure of the WT native state is illustrated in Fig. 1 of ref.[34]. Because of their high coordination, the bonding states of eight core residues (i = 22, 25, 28, 33, 36, 39, 54 and 55)

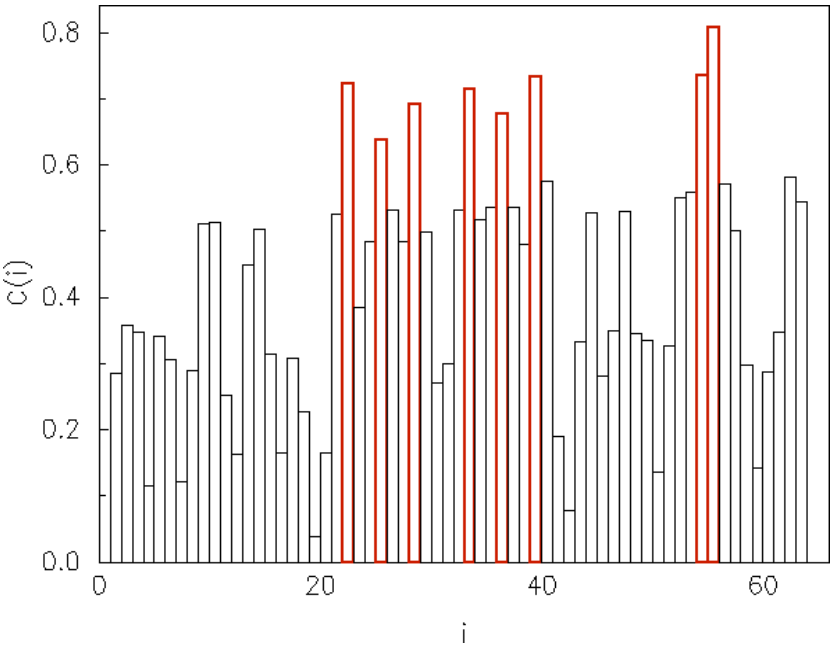

Fig. 1 Relative magnitudes of the correlation $C(i)$ between the bonding state of residue i, $h(i)$, with the proximity to the folded state measured by the order parameter n_n .

feature strongest correlations with the overall proximity to the native state. Among those, residues 54, 55 and 36 are also most likely to participate in inter-protein contacts in aggregated states (c.f. Table 2 of ref.[34]). Using the contact maps for the WT sequence determined in the preceding work[33], we choose two of the eight possible sites, 36 and 54, as representatives of two very different local topologies. Residue 36 sits in the middle of the 11-residue string (31-41) bound anti-parallel to the string of residues 20-30 (the longest anti-parallel segments), and is coordinated only by relatively close contour neighbors. Residues 54 and 55, on the other hand, are coordinated by residues distant on the chain contour. Of these two residues, methionine (M) residue 54 is chosen as the second site of attempted mutation because, within a given model, its

interactions are closer to those of the leucine residue at the other mutation point, 36. Such a choice makes it easier to relate eventual differences in the effects of substitutions to local topologies around targeted residues. Because both *M* and *L* are strongly hydrophobic, substitutions by polar residues prove disruptive to the protein native structure. Substitutions with a residue of intermediate hydrophobicity, like threonine, *T*, however, result only in a mild destabilization of the native state while its structure remains identical to that of the original (WT) sequence. Substitutions *L*36*T* and *M*54*T* were therefore selected for comparison of model protein aggregation behavior with that of the unperturbed sequence.

III.2 Comparison of native-state stabilities for isolated chains with the original and modified sequences

In a lattice-model representation, a useful measure of the proximity to the folded state of a model protein is the number of native contacts, n_n. For a fully folded 64-mer protein with given protein sequence, n_n =81. In Fig. 2 we compare probability distributions $P(n_n)$ for the three considered sequences. The function $-k_BT \ln P(n_n)$ represents a measure of the potential of mean force associated with given value of n_n, i.e. the free energy of the protein subject to a constrained number of native contacts, n_n. The temperature $T=T_0$ corresponds to approximately 0.86 T_m, where T_m is the melting temperature for the WT sequence, i.e. T_m=1.16 T_0. Melting temperatures for the two mutants are about 10% lower, ~1.045 and 1.025 ± 0.02 T_0 for sequences *L*36*T* and *M*54*T*, respectively. While both mutant sequences are somewhat destabilized in comparison to the WT sequence, all three sequences are characterized by similar probability distributions for the number of native contacts with the most probable states lying within the near-folded basin with $n_n=75 \pm 4$. At $T=T_0$, all three sequences have significant native state populations and are

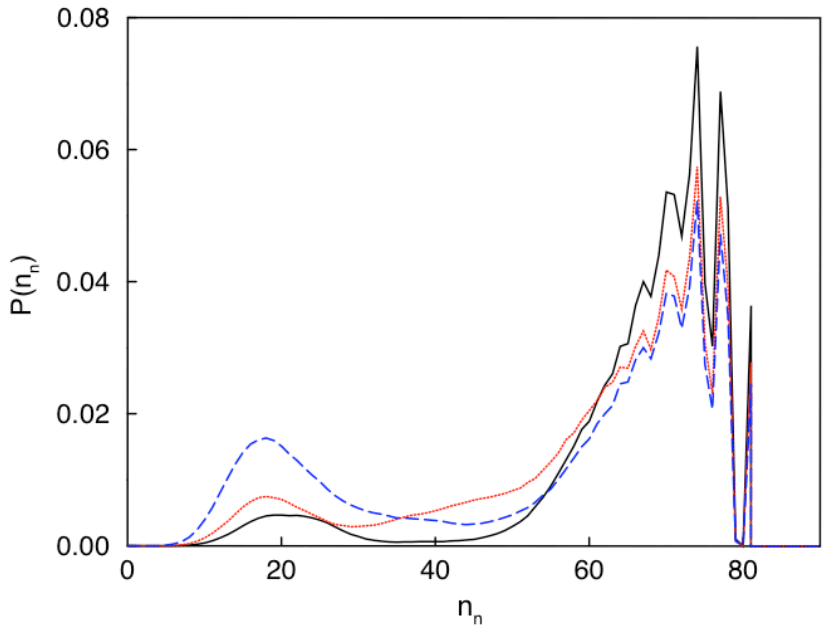

Fig. 2 Probabilities of states with n_n native contacts for the WT protein (black solid) and mutants *L*36*T* (red) and *M*54*T* (blue dashed).

therefore regarded as model representatives of a family of proteins capable of performing the same biological function. As shown below, however, the relatively small stability differences can translate to quite different behaviors under the influence of additional destabilizing factors such as an elevated temperature or increased protein concentration leading to competition between folding and aggregation. Similar comments have been made in the context of single-chain simulations of the aggregation-prone *E*22*Q* mutant of the 10-35 segment of the Alzheimer's -amyloid peptide whose structural fluctuations are noticeably stronger than those for the WT sequence[16]. Further, experiments probing the aggregation of protein G and several of its mutants

showed that their ability to form amyloid fibrils under destabilizing conditions was strongly correlated with their individual stabilities under such conditions[52].

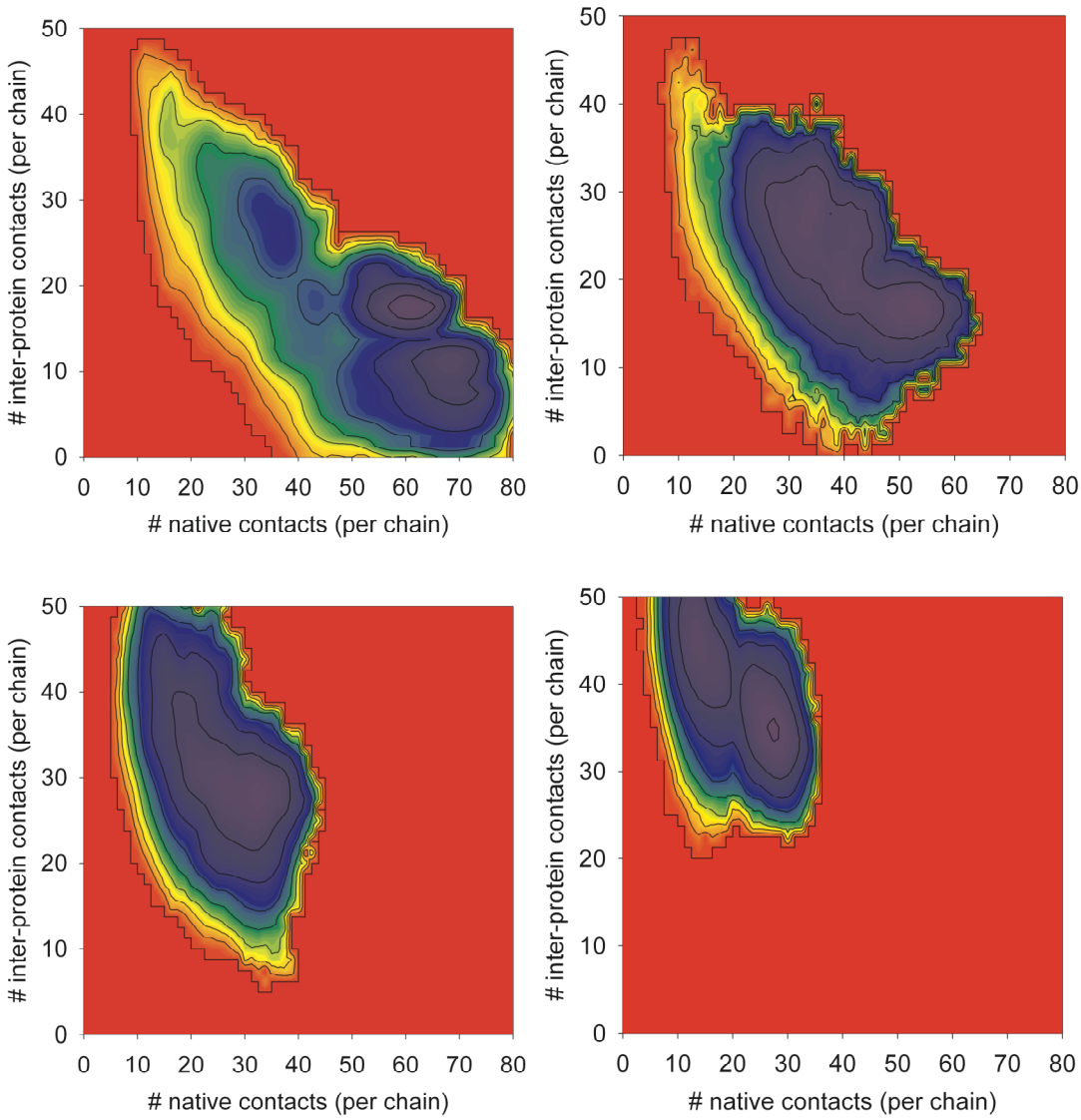

Fig. 3 Free-energy landscapes of 4-chain WT protein systems at packing fractions $_v$=6, 15, 26 and 50%. Structures are characterized in terms of two order parameters: the number of native and inter-protein contacts per chain, n_n and n_i, respectively.

III.3 Structural behavior at elevated protein concentration

In previous work, we demonstrated that the folding landscape of a protein undergoes significant changes when surrounded by adjacent proteins[33,34]. An example of such changes is presented in Fig. 3 showing free-energy landscapes of a system of four WT protein chains at volume fractions $_v$=6, 15, 26 and 50%. The order parameters used to describe both the degrees of folding and inter-protein association are the number of native bonds per chain, 0 n_n 81, and the number of inter-protein contacts per chain, n_i. Free-energy wells on the (n_n, n_i) plane correspond to highly populated states that contribute most to the overall solution behavior. In the relatively dilute system, $_v$=6%, the low free-energy region comprises nearly folded structures with few inter-protein contacts. The two adjacent minima correspond to comparably stable states with slightly different numbers of inter-protein bonds. Aggregated states with low n_n and high n_i are not favored at this concentration. As the concentration rises, the positions of the free-energy minima shift to misfolded (low n_n), strongly associated states with many inter-protein contacts (high n_i). A snapshot showing a typical aggregate of misfolded chains is given in Fig. 4. Despite some inter-protein bonding and concomitant destabilization of the folded states, the dilute system may be viewed as one where protein molecules retain the biological function associated with the native form.

Temporal fluctuations in the fraction of realized native bonds for the dilute WT system are illustrated in the top two graphs in Fig. 5. Here, the number of attempted simulation moves (in passes) is used as the time variable to enable a qualitative comparison between the different sequences. The graph on the left side illustrates the folding process in a run where the initial state in the simulation corresponds to random, unfolded conformation of all four 64-mer chains in the system. Once folded, the system undergoes short-lived structural fluctuations but consistently

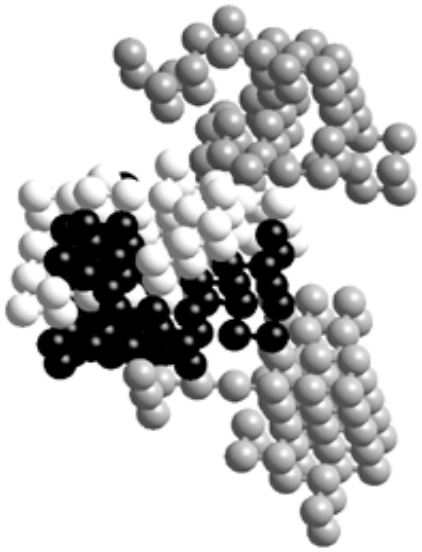

Fig. 4 A snapshot of an aggregate comprising four model protein chains at protein volume fraction ϕ_v=15%.

returns to the folded state; no long-lived unfolded structures are observed when starting from the folded conformation.

While the two mutants resemble the WT protein in isolation (Fig. 2), their behavior becomes markedly different in multi-chain cases, even at a packing fraction as low as ϕ_v~7%, used in the examples in Fig. 5. The second and third pairs of graphs in Fig. 5 correspond to four-chain mutant *L*36*T* and *M*54*T* systems, respectively. The left side shows trajectories originated from unfolded random states. The chains never refold regardless of the time of simulation. The graphs on the right side describe the time dependence of the structures of initially folded multi-chain mutant systems. While pre-folded systems exhibit certain resistance against aggregation, they eventually unfold and never return to the predominantly folded state. As shown in previous studies, [19,23,32,33], the drop in the number of native contacts upon aggregation is accompanied by

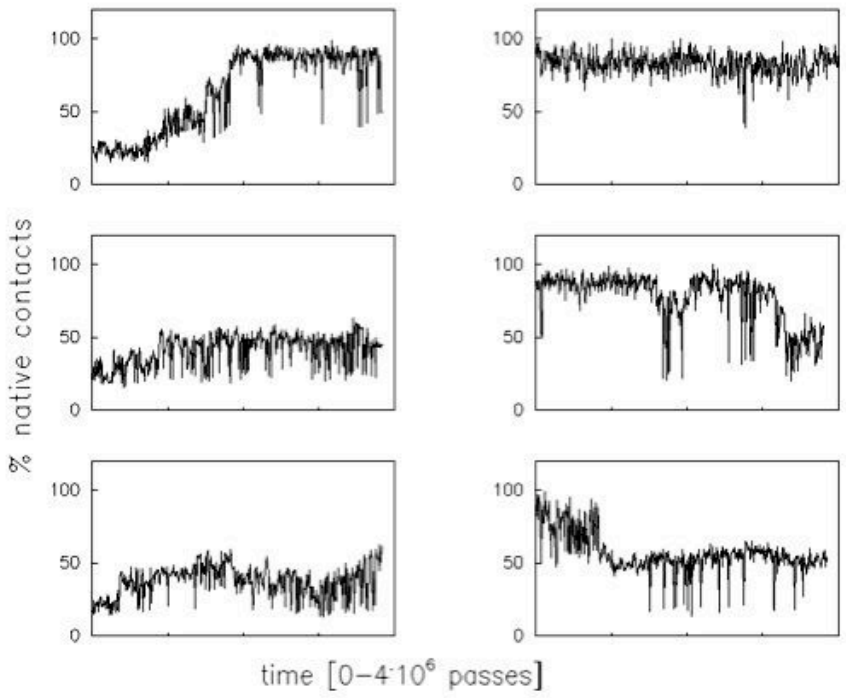

Fig. 5 Simulation-time dependence of the fraction of native bonds in a 7 volume % solution of WT protein (top row) or its mutants *L*36*T* (2nd row) and *M*54*T* (3rd row). Graphs on the left are initiated in unfolded state and those on the right start from a pre-folded configuration.

simultaneous increase in the number of inter-chain bonds such that the energy of stable aggregates remains close to that of isolated native proteins. The above behavior was reproduced consistently in many additional runs (not shown), clearly confirming that, under appropriate conditions, the competition between refolding and aggregation can be strongly influenced even by single-point mutations. Although performed on a simplified model, our calculations suggest possibilities for elucidation of mutation effects in hereditary diseases involving protein aggregation. Similarly, they provide a context for computer-assisted protein engineering leading to protein drugs with improved aggregation resistance while maintaining desired structure and biological activity.

III.4 Thermodynamic characterization of sequence effects in associating systems

While the ability to aggregate is considered a generic feature of proteins[5,9,10,53,54], the substitutions discussed here merely modulate the aggregation propensity at the specified conditions. In what follows, we present results of a thermodynamic analysis, which shows that the effects of (moderate) residue substitutions may be comparable to those of changing external variables like temperature or protein concentration. For this purpose, we rely on a minimalist system subject to competition between protein folding and association: we consider a two-chain 64-mer system of WT or mutant proteins in a simulation cell of size L=12, $_v$~7%. Here, the term association is preferred over aggregation as we consider relatively small protein clusters. Ref.[33] provides a thorough analysis of single and multi-chain WT systems. Fig. 6 shows the temperature dependencies of the system heat capacity, C_v, (Fig. 6a) and the two characteristic order parameters that measure the extents of folding and association, the numbers of native and inter-protein contacts, n_n (Fig. 6b) and n_i (Fig. 6c). The peak in the heat capacity is associated with the breakup of native bonds as the protein unfolds. In multi-chain systems, the melting transition also involves the formation of new inter-protein bonds. The trade-off between native and inter-chain bonds facilitates unfolding of the WT protein at appreciable concentration. As a result, the melting point T_m *shifts* from 1.16 in isolation to 1.12 in the two-chain system. The melting points in mutant systems are lower than those for the WT both in isolation and at finite concentration. However, transition temperatures T_m in the two chain systems, ~1.07 for *L*36*T* and 1.03 for the *M*54*T* sequences, are close to and even slightly above the respective T_m ~1.045 and 1.025 for the same sequences in isolation. The presence of inter-chain bonds, therefore, reduces the differences between stabilities of compact states in the WT and the mutant systems. This

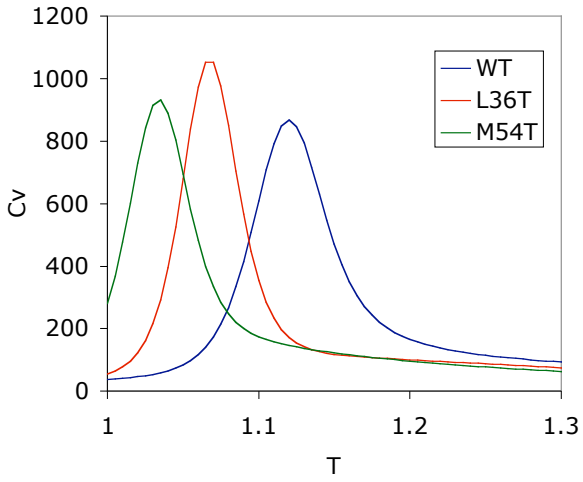

Fig. 6a

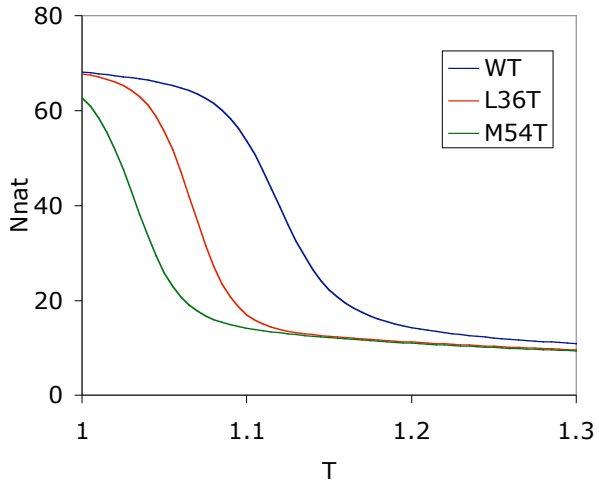

Fig. 6b

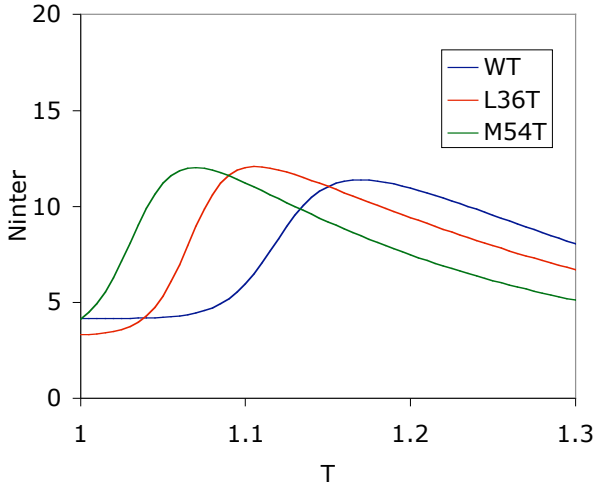

Fig. 6c

Fig. 6 Temperature dependencies of the heat capacity (a), and the numbers of native (b) and inter-protein contacts (c) in a two-chain system containing the WT protein or its one-point mutants *L*36*T* and *M*54*T* at the packing fraction $_v$~7%.

shows that, in the present examples, inter-protein interactions are less specific than the native ones, as the interactions with neighboring chains make up for some of the energetic disadvantage of the native state of the mutant relative to the WT protein.

Fig. 7 shows three-dimensional free-energy landscapes for the WT sequence (top row), mutant *L*36*T* (2nd row), and mutant *M*54*T* (3rd row) at two reduced temperatures, *T*=1.0 (left), and *T*=1.075 (right). Both temperatures are below the melting point T_m of the WT protein. The higher of the two trial temperatures, *T*=1.075, however, slightly exceeds the melting temperatures of the mutants. At T=1.0, the landscape of the WT protein *in the two-chain system* retains the funnel-like shape conducive to folding. Performing the single-point mutations introduces auxiliary free-energy minima corresponding to misfolded conformations. These structures are stabilized by multiple inter-chain bonds. Our data for a two-chain associate illustrate an onset of the concentration-induced transition to the aggregated form observed in the four-chain mutant systems, as illustrated in Fig. 5. As the temperature rises to 1.075, the misfolded associated structures gradually take over in mutant systems. At this temperature, the WT free-energy landscape also begins developing a local minimum corresponding to misfolded associates similar to the mutant *L*36*T* landscape at the lower temperature *T*=1.0. The two-chain *L*36*T* system at *T*=1.0 is almost as far from the system's melting temperature as is the WT two-chain system at

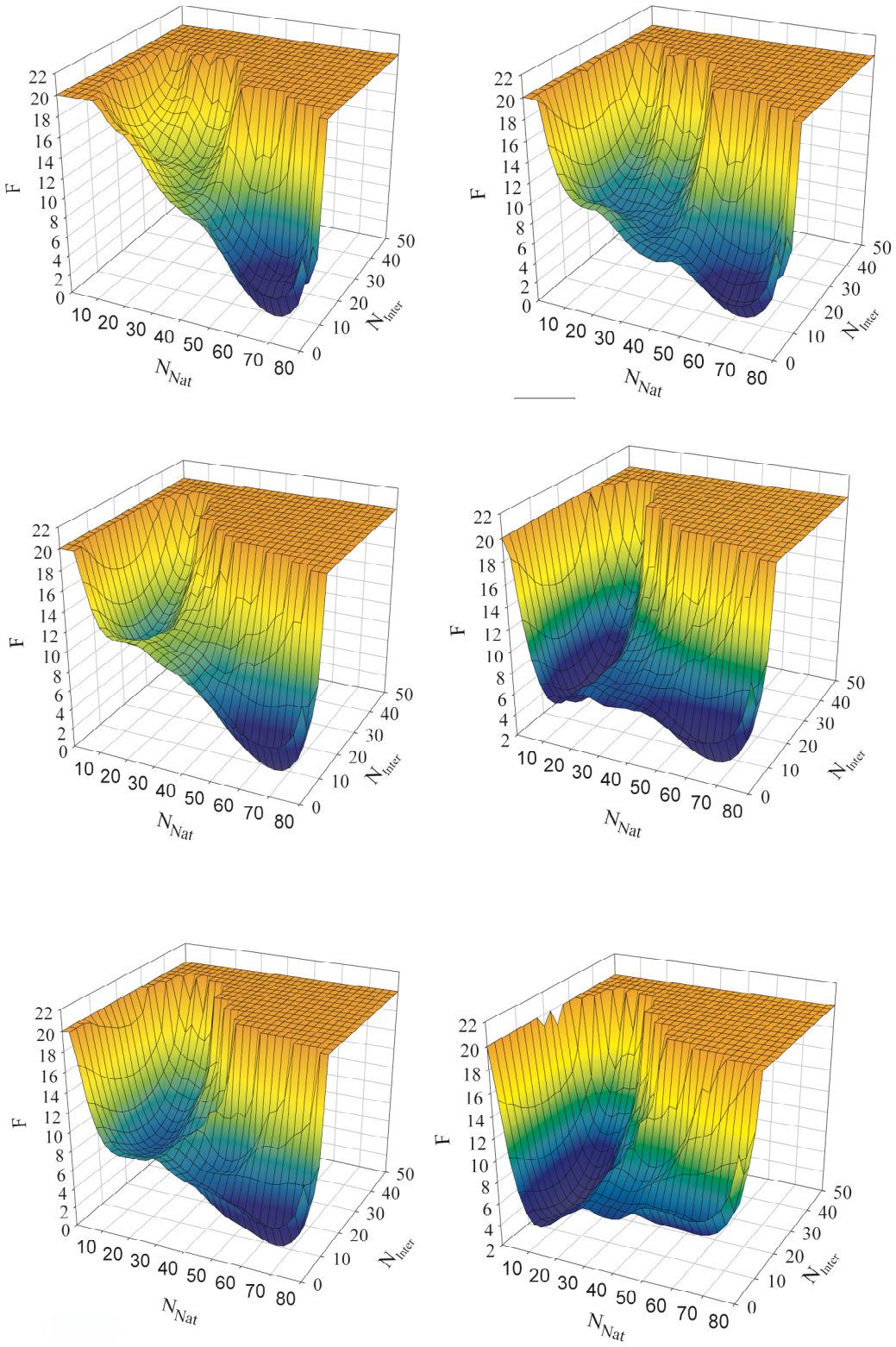

Fig. 7 Free energy landscapes of two-chain WT (top row), mutant *L*36*T* (2nd row) and *M*54*T* (3rd row) systems at temperatures T=1.0 (left) and T=1.075 (right) at volume fraction ~7%.

T=1.077. Based on the temperature dependencies shown in Figures 6b and 6c, the free-energy landscapes of the WT and the mutants respond to an increase in the temperature in an analogous manner, albeit at different transition temperatures, T_m. The shift in T_m is somewhat bigger with the mutant *M*54*T* than that for the *L*36*T* case. It appears commensurate to the difference between the stabilities of the native states associated with the three different sequences (see e.g. Fig. 2).

The generality of our observations is limited by a small number of trial sequences; further, our findings are based on studies of small, disordered oligomers that can be regarded as precursors of large and more orderly aggregates. Nevertheless, the above examples lend support to the notion that appropriate sequence substitutions can modulate the thermodynamics of folding and association in a manner analogous to a change in external conditions. This is to say that we can use subtle mutations to shift the coexistence lines in a multi-dimensional phase diagram of a protein solution without modifying the phase behavior qualitatively. While only temperature and concentration have been used as control system variables in the present calculations, generalizations of this notion may apply to other properties such as the change in solvent composition, including the addition of denaturant. This is to say that, because of strong correlations between the binding states of the specified monomer and the proximity to the folded state of the protein as a whole[23], strengthening or weakening the bonds of a selected residue can have an effect similar to modulating the overall strength of intra-protein bonds for the whole chain.

Our observations can be extrapolated to multi-chain *disordered* oligomers. Less can be said about the effect of sequence alterations on the transition to ordered multi-chain aggregates such as those observed in amyloid fibrils. These structures are held together primarily by nonspecific (hydrogen bond) interactions among backbone groups augmented by attraction

between hydrophobic side chains. Substitutions we choose here weaken inter-chain hydrophobic forces and are not likely to stabilize the aggregated form. Their effect on the competition between the native and aggregated states can therefore be attributed to reduced stability of native monomers. Mutation effects analogous to those observed with low oligomers can also be expected in systems comprising higher and more ordered aggregates. Recent discontinuous molecular dynamics studies have shown that stable ordered fibrillar structures eventually replace amorphous aggregates when grown beyond the critical nucleus size typically comprising about 10^2 or more peptide chains[28]. In view of computational costs, such calculations have so far been performed only for systems of relatively short oligopeptides whose unfavorable surface-to-volume ratio does not support stable native structures representative of folded proteins. In future work, we plan to exploit the axial periodicity of ordered fibrillar phases to model large ordered aggregates by means of periodic boundary conditions. Within a two-box system open to monomer transfer analogous to the Gibbs ensemble simulation[46], this type of calculation is expected to enable a thermodynamic characterization for the competition between the two stable phases (the native state and ordered aggregates), complementing the studies of oligomeric intermediates described in the present work.

IV. CONCLUDING REMARKS

Most proteins can exist both in stable folded states and as aggregates, with the prevailing form determined by external conditions. Simple model systems presented here provide examples where single-point sequence alterations can shift the coexistence boundary between the different regimes without affecting the native-state conformation of the protein. This way, they suggest the possibility of modulating protein-aggregation propensity when external conditions are

imposed by system requirements, e.g. in living cellular environments or in the processing and storage of protein drugs. Without ruling out the possibility of more specific mechanisms, which could be captured using a molecular-level description of associating proteins, our coarse-grained calculations suggest that adequate effects can generally be obtained by moderate sequence substitutions leading to mild destabilization or stabilization of the protein native state. While the former cases can be associated with the genetic propensity to certain protein-aggregation diseases, the latter provide a basis for envisaged engineering of comparatively aggregation-resistant proteins. With expected development of more realistic, yet computationally tractable protein models, molecular simulations hold promise to become an important tool in computer-assisted design of novel proteins with optimized aggregation behavior, leading to improved refolding yield of inclusion bodies, or to enhanced and controlled aggregates that serve as nanomaterials.

Acknowledgment

For financial support, the authors are grateful to the National Science Foundation under award BES-0432625 and to the Office for Basic Sciences of the U.S. Department of Energy.

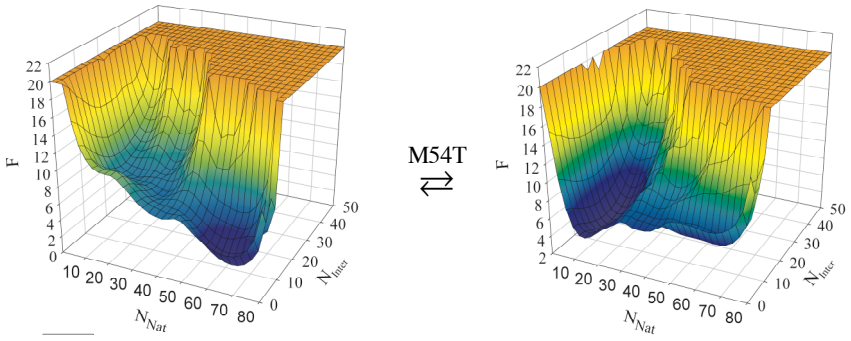

Figure for T.O.C.

(1) Fink, A. L. *Folding & Design* **1998**, *3*, R9.
(2) Dobson, C. *Phil. Trans. R. Soc. London B* **2001**, *356*, 133.
(3) Fernandez, A.; Boland, M. D. L. *Febs Letters* **2002**, *529*, 298.
(4) Lanckriet, H.; Middelberg, A. P. J. *Biotechnology Progress* **2004**, *20*, 1861.
(5) Villegas, V.; Zurdo, J.; Filimonov, V. V.; Aviles, F. X.; Dobson, C. M.; Serrano, L. *Protein Science* **2000**, *9*, 1700.
(6) Clark, E. D. *Curr Opin Biotechnol* **2001**, *12*, 202.
(7) Fowler, S. B.; Poon, S.; Muff, R.; Chiti, F.; Dobson, C. M.; Zurdo, J. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **2005**, *102*, 10105.
(8) Aggeli, A.; Bell, M.; Boden, N.; Keen, J. N.; McLeish, T. C. B.; Nyrkova, I.; Radford, S. E.; Semenov, A. *Journal of Materials Chemistry* **1997**, *7*, 1135.
(9) Koga, T.; Taguchi, K.; Kobuke, Y.; Kinoshita, T.; Higuchi, M. *Chemistry-a European Journal* **2003**, *9*, 1146.
(10) MacPhee, C. E.; Dobson, C. M. *Journal of the American Chemical Society* **2000**, *122*, 12707.
(11) Hamada, D.; Yanagihara, I.; Tsumoto, K. *Trends in Biotechnology* **2004**, *22*, 93.
(12) Gupta, P.; Hall, C. K.; Voegler, A. C. *Protein Science* **1998**, *7*, 2642.
(13) Istrail, S.; Schwartz, R.; King, J. *Journal of Computational Biology* **1999**, *6*, 143.
(14) Harrison, P. M.; Chan, H. S.; Prusiner, S. B.; Cohen, F. E. *Journal of Molecular Biology* **1999**, *286*, 593.
(15) Harrison, P. M.; Chan, H. S.; Prusiner, S. B.; Cohen, F. E. *Protein Science* **2001**, *10*, 819.
(16) Massi, F.; Straub, J. E. *Biophys J* **2001**, *81*, 697.
(17) Massi, F.; Peng, J. W.; Lee, J. P.; Straub, J. E. *Biophys J* **2001**, *80*, 31.
(18) Massi, F.; Straub, J. E. *J Comput Chem* **2003**, *24*, 143.
(19) Bratko, D.; Blanch, H. W. *Journal of Chemical Physics* **2001**, *114*, 561.
(20) Blanch, H. W.; Prausnitz, J. M.; Curtis, R. A.; Bratko, D. *Fluid Phase Equilibria* **2002**, *194*, 31.
(21) Ma, B. Y.; Nussinov, R. *Protein Science* **2002**, *11*, 2335.
(22) Jang, H.; Hall, C. K.; Zhou, Y. Q. *Biophysical Journal* **2002**, *82*, 646.
(23) Bratko, D.; Blanch, H. W. *Journal of Chemical Physics* **2003**, *118*, 5185.
(24) Zanuy, D.; Nussinov, R. *Journal Of Molecular Biology* **2003**, *329*, 565.
(25) Cellmer, T.; Bratko, D.; Blanch, H. *Biophysical Journal* **2003**, *84*, 41A.
(26) Jang, H. B.; Hall, C. K.; Zhou, Y. Q. *Protein Science* **2004**, *13*, 40.
(27) Hall, C. K.; Nguyen, H. D. *Neurobiology Of Aging* **2004**, *25*, S171.
(28) Nguyen, H. D.; Hall, C. K. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **2004**, *101*, 16180.
(29) Nguyen, H. D.; Hall, C. K. *Biophysical Journal* **2004**, *87*, 4122.
(30) Peng, S.; Ding, F.; Urbanc, B.; Buldyrev, S. V.; Cruz, L.; Stanley, H. E.; Dokholyan, N. V. *Physical Review E* **2004**, *69*.
(31) Urbanc, B.; Cruz, L.; Yun, S.; Buldyrev, S. V.; Bitan, G.; Teplow, D. B.; Stanley, H. E. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **2004**, *101*, 17345.
(32) Cellmer, T.; Bratko, D.; Prausnitz, J. M.; Blanch, H. *Biotechnology And Bioengineering* **2005**, *89*, 78.

(33) Cellmer, T.; Bratko, D.; Prausnitz, J. M.; Blanch, H. *Journal Of Chemical Physics* **2005**, *122*.
(34) Cellmer, T.; Bratko, D.; Prausnitz, J. M.; Blanch, H. W. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 11692.
(35) Clark, L. A. *Protein Science* **2005**, *14*, 653.
(36) Tarus, B.; Straub, J. E.; Thirumalai, D. *Journal Of Molecular Biology* **2005**, *345*, 1141.
(37) Nguyen, H. D.; Hall, C. K. *Journal Of Biological Chemistry* **2005**, *280*, 9074.
(38) Fawzi, N. L.; Chubukov, V.; Clark, L. A.; Brown, S.; Head-Gordon, T. *Protein Science* **2005**, *14*, 993.
(39) Leonhard, K.; Prausnitz, J. M.; Radke, C. J. *Protein Science* **2004**, *13*, 358.
(40) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534.
(41) Leonhard, K.; Prausnitz, J. M.; Radke, C. J. *Physical Chemistry Chemical Physics* **2003**, *5*, 5291.
(42) Leonhard, K.; Prausnitz, J. M.; Radke, C. J. *Biophysical Chemistry* **2003**, *106*, 81.
(43) Carlsson, F.; Malmsten, M.; Linse, P. *Journal Of Physical Chemistry B* **2001**, *105*, 12189.
(44) Wu, J. Z.; Bratko, D.; Prausnitz, J. M. *Proceedings of the National Academy of Sciences of the United States of America* **1998**, *95*, 15169.
(45) Striolo, A.; Bratko, D.; Wu, J. Z.; Elvassore, N.; Blanch, H. W.; Prausnitz, J. M. *Journal of Chemical Physics* **2002**, *116*, 7733.
(46) Frenkel, D.; Smit, B. *Understanding molecular simulation*; Academic Press: New York, 1996.
(47) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248.
(48) Wall, F. T.; Mandel, F. *Journal of Chemical Physics* **1975**, *63*, 4592.
(49) Gront, D.; Kolinski, A.; Skolnick, J. *Journal of Chemical Physics* **2000**, *113*, 5065.
(50) Cecchini, M.; Rao, F.; Seeber, M.; Caflisch, A. *Journal Of Chemical Physics* **2004**, *121*, 10748.
(51) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *Journal of Computational Chemistry* **1992**, *13*, 1011.
(52) Ramirez-Alvarado, M.; Merkel, J. S.; Regan, L. *Proceedings of the National Academy of Sciences of the United States of America* **2000**, *97*, 8979.
(53) Zurdo, J.; Guijarro, J. I.; Jimenez, J. L.; Saibil, H. R.; Dobson, C. M. *Journal of Molecular Biology* **2001**, *311*, 325.
(54) Chiti, F.; Webster, P.; Taddei, N.; Clark, A.; Stefani, M.; Ramponi, G.; Dobson, C. M. *Proceedings of the National Academy of Sciences of the United States of America* **1999**, *96*, 3590.