Simulations of Kinetically Irreversible Protein Aggregate Structure

Sugunakar Y. Patro and Todd M. Przybycien
Bioseparations Research Center, Howard P. Isermann Department of Chemical Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180-3590 USA

ABSTRACT We have simulated the structure of kinetically irreversible protein aggregates in two-dimensional space using a lattice-based Monte-Carlo routine. Our model specifically accounts for the intermolecular interactions between hydrophobic and hydrophilic protein surfaces and a polar solvent. The simulations provide information about the aggregate density, the types of inter-monomer contacts and solvent content within the aggregates, the type and extent of solvent exposed perimeter, and the short- and long-range order all as a function of (i) the extent of monomer hydrophobic surface area and its distribution on the model protein surface and (ii) the magnitude of the hydrophobic-hydrophobic contact energy. An increase in the extent of monomer hydrophobic surface area resulted in increased aggregate densities with concomitant decreased system free energies. These effects are accompanied by increases in the number of hydrophobic-hydrophobic contacts and decreases in the solvent-exposed hydrophobic surface area of the aggregates. Grouping monomer hydrophobic surfaces in a single contiguous stretch resulted in lower aggregate densities and lower short range order. More favorable hydrophobic-hydrophobic contact energies produced structures with higher densities but the number of unfavorable protein-protein contacts was also observed to increase; greater configurational entropy produced the opposite effect. Properties predicted by our model are in good qualitative agreement with available experimental observations.

INTRODUCTION

Insoluble protein aggregates are encountered in a wide variety of biological systems and bioprocessing operations including inclusion bodies in recombinant bacteria, plaques associated with neurological disorders, as well as aggregates produced in protein refolding, precipitation, crystallization, formulation, and storage procedures. Aggregates of protein pharmaceuticals often exhibit reduced bioactivity (Becker et al., 1987) and have been implicated in the generation of immunogenic reactions (Lewis et al., 1969a, b; Moore and Leppert, 1980). In order to redissolve precipitates in a process stream, suspend aggregates in an excipient, or block aggregate formation in vivo or in vitro in a rational fashion, it is necessary to understand the structure and properties of protein aggregates.

Protein and colloid aggregation *processes* have been studied and modeled extensively; however, protein aggregate *properties* have not been studied in great detail. Protein aggregation has been investigated mainly as a competitive process for protein folding (Bowden and Georgiou, 1990; Kiefhaber et al., 1991). The majority of unstructured colloid aggregation models are confined to random ballistic deposition simulations (Jullien and Meakin, 1989). The protein aggregation process has been studied through random sequential adsorption (Feder, 1980) and stochastic cellular automaton (Stenberg and Nygren, 1991) models. However, these models typically treat protein monomers as uniform discs of negligible thickness, ignoring the structure and organization of the protein surface (Feder, 1980). Oosawa et al. (1972) proposed and Higo et al. (1992) extended a "poker-chip" model with distinguishable contacting sites to simulate the effects of intermolecular interactions and temperature on solid state phase transitions in a two-dimensional protein crystal. Again, although the order in simulated crystals was analyzed, no assessment of the physical properties of the crystals were made. Dill and collaborators (1989) have developed a three-dimensional statistical mechanical lattice model for protein folding equilibria which they later extended to aggregation equilibria (Fields et al., 1992). They modeled aggregation as a balance between two opposing forces, namely the hydrophobic interaction which favors aggregation and the configurational and translational entropies which favor disaggregation. In this structured model, a protein is abstracted as a copolymer of hydrophobic and hydrophilic monomers. They predicted phase diagrams for thermally induced and denaturant induced protein aggregates as a function of chain length and monomer hydrophobic fraction.

With the possible exception of the self-association of denatured protein, the existence of completely random protein aggregation processes and protein aggregates may be a myth. Brems and co-workers (1992) have recently reported that the self-association of human insulin can be drastically altered by substitutions at one or two key sites on the B chain. Przybycien and Bailey (1989) observed a decrease of two to three orders of magnitude in the apparent Smoluchowski aggregation rate constant for α-chymotrypsin precipitation by KSCN when a specific dimerization site was blocked. Brems (1988) observed that precipitation of bovine growth hormone (bGH) was inhibited by addition of fragments 96–133 or 109–133 derived from bGH. It was suggested that these fragments bind specifically to the intermediate participating in the aggregate formation. The effectiveness of several bGH-derived fragments to inhibit precipitation was found to vary. Casal et al. (1988) studied the denaturation of

Received for publication 26 August 1993 and in final form 27 December 1993.

Address reprint requests to Todd M. Przybycien, Bioseparations Research Center, The Howard P. Isermann Department of Chemical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590. Tel.: 518-276-6377; Fax: 518-276-4030.

0006-3495/94/05/1274/16 $2.00

β-lactoglobulin B through infrared spectroscopy. They were able to identify distinct mechanisms for thermal denaturation, alkaline denaturation, and denaturation due to pH variation. They identified specific regions of the protein tertiary structure to be operative in each type of aggregation. These studies indicate that the intermolecular interactions of proteins within aggregates clearly impact the aggregation process and, hence, the properties of the resulting aggregate.

We have developed a two-dimensional model to simulate the final structure of kinetically irreversible protein aggregates. The protein monomer surface was assumed to consist only of hydrophobic and hydrophilic patches and both the monomers and aggregates were suspended in a polar solvent. Aggregate structural properties were assessed as a function of (i) the number and orientation of hydrophobic and hydrophilic patches on the monomer surface and (ii) the relative magnitude of the hydrophobic-hydrophobic interaction free energy. The structural properties we addressed include the aggregate density, the distribution of intermolecular contacts within the aggregate, the nature and extent of solvent-exposed aggregate surface area, the pore size distribution within the aggregate, and the extent of short- and long-range order. Protein aggregation may be thermodynamically (De Young et al., 1993a) or kinetically (Zettlmeissl et al., 1979; Mulkerrin and Wetzel, 1989; Orsini and Goldberg, 1978) controlled. In this work, we are interested in assessing the properties of kinetically irreversible aggregates, typified by salt-induced precipitates (Przybycien and Bailey, 1989) and inclusion bodies (Mitraki et al., 1991). The aggregation process itself, is not the focus of our study. However, we expect our simulation process to result in final aggregate states comparable with those obtained by physically realistic and, hence, more computationally intensive aggregation processes. Reversible aggregate structures will be treated separately.

THEORY

Model development

To generate simulated protein aggregates for structural characterization, we start with a closed, constant "volume" system consisting of a dilute "solution" of two-dimensional, hexagonal protein monomers in a polar solvent to which a precipitating agent, e.g., salt, has been added. This system is maintained in thermal equilibrium with the surroundings. The precipitating agent causes the net interaction between a protein monomer and the solvent to be unfavorable. The system is subsequently subdivided into a *solution* phase and a *solvated precipitate* phase. The solution phase is unstructured and is assumed dilute enough that protein-protein interactions are negligible. Indeed, experimental work with protein aggregates has indicated that the solution phase is largely devoid of nonspecific aggregates (Przybycien and Bailey, 1989). The solvated precipitate phase, or simply the precipitate phase, comprises a hexagonal lattice which is initially occupied only by solvent. This system is pictured in Fig. 1. The final state of the precipitate phase is akin to a cross-sectional view of an aggregate particle.

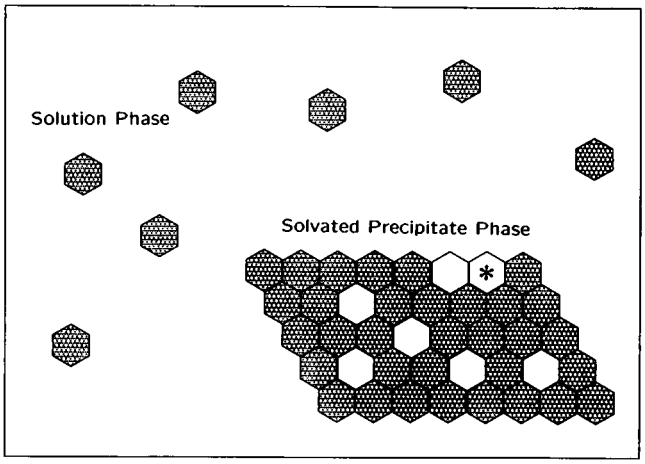

FIGURE 1 The *solution* and *precipitate* phases of the system. Dispersed protein monomers are shown to indicate the *solution* phase and the *solvated precipitate* phase is shown in the form of a lattice. The filled hexagons in the lattice represent protein monomers within the aggregate and the unfilled hexagons represent the entrapped solvent. A description of interactions disrupted and formed during the process of transferring a monomer from solution phase to the '*' site in the precipitate phase is given in the text.

The only change that is allowed for the system is the transfer of monomers, one-by-one, from the solution phase to the precipitate phase accompanied by the simultaneous back-transfer of an equivalent volume of solvent from the precipitate phase to the solution phase. In a broad sense, the driving force for the transfer is the eventual replacement of unfavorable protein-solvent interactions in the solution phase with more favorable protein-protein interactions in the precipitate phase. Each transfer increases the extent of aggregation, ξ, and is governed by a Monte-Carlo algorithm explained below. As our aim is to simulate a kinetically irreversible aggregate, the transfer process continues vectorially in the direction of increasing extent of aggregation until the total entropy of the system plus the surroundings, S_{irr}, is maximized (Prigogine, 1967). Physically this irreversibility corresponds to a negligible resolubilization rate in the face of subsequent monomer addition; high salt concentrations are believed to lead to such kinetically irreversible situations due to the presence of very deep primary free energy well (De Young et al., 1993b). If pressure effects are neglected, we may maximize the total entropy by minimizing the Gibb's energy of the system, G_{sys}; for a closed system at constant T and P, where only changes in ξ are allowed, irreversible thermodynamics gives (de Donder and Rysselberghe, 1936; Denbigh, 1950)

$$\left(\frac{dS_{irr}}{d\xi}\right)_{T,P} = -\frac{1}{T}\left(\frac{dG_{sys}}{d\xi}\right)_{T,P.} \quad (1)$$

Transfers continue, until $(dS_{irr}/d\xi) = (dG_{sys}/d\xi) = 0$. The structure of the resulting precipitate phase is then analyzed. In essence, we are treating the phase transfer process by the chemical reaction:

$$nP + P_m \rightarrow (n-1)P + P_{m+1}, \quad (2)$$

where P is soluble, monomeric protein and P_m is an aggregate composed of m monomers. The driving force for this irreversible process is the affinity (de Donder and Rysselberghe, 1936) of the above reaction.

As the protein molecules are represented by regular hexagons, sides 1 through 6 of an individual hexagon represent six different "surfaces" of the two-dimensional projection of a protein molecule. When a hexagon is replaced by another hexagon, it represents a substitution of the volumes in the three-dimensional sense. Thus, we can account for protein-protein (p, p), protein-solvent (p, s) and solvent-solvent (s, s) interactions depending upon the nature of adjacent hexagon sides. The formation of favorable (p, p) and (s, s) interactions in the precipitate and solution phases respectively, and the disruption of unfavorable (p, s) interactions in both phases, constitute an energetic gain; however, the confinement of protein to the lattice in the precipitate phase results in a loss in entropy. The overall free energy change of the system associated with a single transfer, ΔG_{sys}, can be interpreted as a combination of the energetic gain and the entropic loss. For a transfer to be favorable, the energetic gain must compensate for the entropic loss.

For the transfer of a protein molecule from the solution phase into an isolated or dilute region of the precipitate phase, with no adjacent protein monomers, six (s, s) interactions are replaced by six (p, s) interactions in the precipitate phase. Simultaneously, six (p, s) interactions in the solution phase are disrupted and six (s, s) interactions are formed due to the equivalent solvent back-transfer into the solution phase. In other words, there is no net contribution from the interaction energies for this type of transfer as the increase in the free energy of the precipitate phase is compensated for by an equal decrease in the free energy of the solution phase. Only the decrease in configurational entropy contributes to the system free energy change for such a transfer. The criterion for acceptance of protein molecules into the aggregate phase based on the net ΔG_{sys} is discussed in the next section.

For a more general transfer of a single monomer from the solution phase to any solvent occupied position in the lattice surrounded by n protein monomers,

$$\begin{aligned}\Delta G_{sys} &= \sum_{i=1}^{n} \Delta\Delta G(p,p)_i + \Delta G_{config} \\ &= \Delta G_{ppt} + \Delta G_{sol} + \Delta G_{config},\end{aligned} \tag{3}$$

where $\Delta\Delta G(p,p)_i$ is the free energy change associated with the formation of the ith (p, p) interaction *in solution* and ΔG_{config} is the free energy penalty resulting from the reduction in entropy due to the decreased number of accessible configurational arrangements, in terms of monomer translational and rotational states, in the lattice phase. We have further parsed $\sum_{i=1}^{n} \Delta\Delta G(p,p)_i$ into ΔG_{ppt} and ΔG_{sol}, the changes in the free energy of the precipitate and solution phases, respectively. This formalism has enabled us to consider aggregate formation from monomers with net unfavorable solvent interactions and to estimate the free energy trajectories for the precipitate and solution phases. The free energy changes of both phases are expanded in terms of an explicit accounting of the number and type of intermolecular interactions that are formed and disrupted during the transfer process.

For the precipitate phase,

$$\begin{aligned}\Delta G_{ppt} = &\sum_{i=1}^{n} \Delta G(p,p)_i + \sum_{i=n+1}^{6} \Delta G(p,s)_i \\ &- \left(\sum_{i=1}^{n} \Delta G(p,s)_i + (6-n)\Delta G(s,s) \right),\end{aligned} \tag{4}$$

where $\Delta G(x, y)$ is the free energy change accompanying the formation of an (x, y) interaction *in vacuo*. The index i refers to the specific protein surfaces involved for (p, p) and (p, s) interactions. The first two terms on the right hand side of Eq. 4 account for the formation of n (p, p) contacts and $6 - n$ (p, s) contacts in the precipitate phase accompanying the insertion of the monomer at that lattice site. The next two terms describe the corresponding disruption of n (p, s) and $6 - n$ (s, s) interactions resulting from the removal of an equivalent volume of solvent from that lattice site. For example, for a transfer to the lattice point marked by the asterisk in Fig. 1, three (s, s) contacts and three (p, s) contacts will be replaced by three (p, s) and three (p, p) contacts in the precipitate phase. The indexing is necessary as the protein surfaces are distinguishable in our system; the interaction energy of each (p, s) and (p, p) contact may be unique. Thus ΔG_{ppt} will depend on ξ as changes from one transfer to the next, reflecting the number and orientation of protein monomers surrounding the selected lattice site as well as the orientation of the inserted monomer, occur as the lattice evolves.

The free energy change of the solution phase is given by

$$\Delta G_{sol} = 6\Delta G(s,s) - \sum_{i=1}^{6} \Delta G(p,s)_i. \tag{5}$$

The first term on the right hand side of Eq. 5 accounts for solvent backtransfer to the solution phase and the second term accounts for the desolvation of the monomer upon transfer. Note that the solution phase is assumed to be dilute; no (p, s) contacts are formed by the backtransfer of solvent. Therefore, ΔG_{sol} is independent of ξ. In view of Eqs. 3–5, we have formally expanded $\Delta\Delta G(p,p)$ in terms of its constituent interactions as

$$\begin{aligned}&\Delta\Delta G(p,p') \\ &= \Delta G(p,p') + \Delta G(s,s) - [\Delta G(p,s) + \Delta G(p',s)].\end{aligned} \tag{6}$$

The change in configurational entropy, ΔS_{config}, is given by (Hill, 1960):

$$\Delta S_{config} = k_B(\ln \Omega_{ppt} - \ln \Omega_{sol}), \tag{7}$$

where k_B is Boltzmann's constant and Ω_{ppt} and Ω_{sol} are the number of accessible translational and rotational states for the protein molecule in the precipitate phase and solution phase, respectively. Ω_{ppt} is simply the product of the number

of solvent-occupied lattice sites and the number of unique monomer rotational orientations. Using Hill's order of magnitude estimate of $\Omega_{sol} \sim e^{N_A}$, where N_A is the Avogadro's number, $\Omega_{ppt} \ll \Omega_{sol}$ at any level of occupancy for a reasonably sized lattice. Hence,

$$\Delta S_{config} \sim -k_B N_A = -R, \tag{8}$$

where R is the universal gas constant. Although the configurational entropy is strictly a function of protein concentration (Fields et al., 1992; Flory, 1953), we assume an extensive dilute solution phase with negligible concentration changes. A value of R kcal/mol for the configurational entropy was also estimated as a first approximation by Flory (1953).

The free energy penalty accompanying a single transfer is given by

$$\Delta G_{config} = -T\Delta S_{config} \sim RT, \tag{9}$$

and is assumed to be the same for all transfers and for all types of protein monomers. Since ΔG_{config} increases the system free energy, there is a bias towards transfers into positions with net favorable (p, p) contacts that can offset ΔG_{config}. The net result is that the placement of monomers in the precipitate phase is biased toward locations adjacent to occupied lattice sites rather than to isolated lattice regions. These sites are analogous to nuclei and serve to limit the artificial randomness that can arise from transfers to lattice sites that are out-of-register with one another.

Model execution

The simulation is initiated by choosing a particular site on the lattice at random and attempting to place a protein molecule at that site with a random rotational orientation. If the selected site is occupied by solvent (Feder, 1980; Jullien and Meakin, 1989) and if the ΔG_{sys} associated with the transfer is nonpositive, the placement is immediately accepted. The ΔG_{sys} associated with each transfer is calculated as given by Eqs. 3–5. If $\Delta G_{sys} > 0$ for the transfer, Metropolis sampling is employed (Pattou et al., 1991). This assumes that the probability of an energetically unfavorable move being accepted is finite and is governed by a Boltzmann distribution; the probability of acceptance is proportional to $e^{(-\Delta G_{sys}/RT)}$. This accounts for fluctuations proportional to $e^{(-\Delta S_{sys}/R)}$ that occur along irreversible paths (Katchalsky and Curran, 1965). When the lattice is dilute, $\Delta G_{sys} \sim \Delta G_{config}$ and Metropolis sampling will permit only occasional transfers. This is analogous to a nucleation *process*. If the move is unacceptable even by Metropolis sampling, that site is rejected and the protein molecule is returned to the solution phase. This is termed a "reject." Following a reject, another placement attempt is made in accordance with the prescription above. To compensate for the finite nature of the lattice, the model employs periodic boundary conditions. The absence of microscopic reversibility ensures kinetic irreversibility; proteins that are transferred to the precipitate phase are not allowed to return to the solution phase. This process continues until any further transfers become energetically unfavorable, a condition usually termed the "jamming limit" (Feder, 1980; Jullien and Meakin, 1989).

The jamming limit, ξ_{JL}, is identified as the extent of aggregation which minimizes the cumulative or total system free energy change, $\Delta G'_{sys}(\xi)$ given by

$$\Delta G'_{sys}(\xi) = \sum_{\xi=0}^{\xi} \Delta G_{sys}(\xi). \tag{10}$$

Since the initial system free energy is an arbitrary constant,

$$\left.\frac{d\Delta G'_{sys}(\xi)}{d\xi}\right|_{\xi=\xi_{JL}} = \left.\frac{dG_{sys}}{d\xi}\right|_{\xi=\xi_{JL}} = 0, \tag{11}$$

and ξ_{JL} is the extent of aggregation that minimizes G_{sys} and maximizes S_{irr}.

As the progress in $\Delta G'_{sys}$ is deterministic, the minimum along a particular path is recognized only in hindsight. In order to locate ξ_{JL}, the simulation continues transferring monomers until the lattice is saturated at the 99% confidence level. In other words, we allow a sufficient number of rejects between two successful placements such that on a statistical basis, virtually all attempts to fill any empty lattice site with any orientation of protein molecule have failed. The number of allowed rejects between two successful placements is calculated as:

$$0.99 = 1 - \Omega_{ppt}\left(\frac{\Omega_{ppt} - 1}{\Omega_{ppt}}\right)^m, \tag{12}$$

where m is the maximum allowable number of rejects at any point of aggregation to reach the saturation limit with 99% confidence. At this point in the simulation, it was empirically verified that $\Delta G'_{sys}$ had passed through the minimum for that path. We then followed the simulation backwards to find the state of the lattice corresponding to the minimum free energy. It should be noted that this is the minimum along a given kinetically irreversible path and does not necessarily represent the global minimum along all possible paths.

The structure obtained at $\xi = \xi_{JL}$ represents a cross-sectional view of the structure of a possible kinetically irreversible protein aggregate for that set of conditions. At the jamming limit, the lattice is typically partly filled with protein molecules and remainder with solvent. The solvent regions in the lattice are termed pores. It is not surprising to observe pores in the final aggregate structure as both protein crystals (Matthews, 1968) and aggregates (Shih et al., 1992) have significant solvent content. For each simulation, $\Delta G'_{ppt}$, $\Delta G'_{sys}$, $\Delta G'_{sol}$, the pore-size distribution, the distribution of (p, p) contacts that are present, and the nature of solvent accessible surface area in the final aggregate are determined at the jamming limit.

Interaction free energies and monomer configurations

We considered two types of protein surfaces: hydrophobic ($H\Phi$) and hydrophilic (HF); no electrostatic interactions were included. Thus, three types of (p, p) interactions are possible: $(H\Phi, H\Phi)$, $(H\Phi, HF)$ and (HF, HF). This is along

the lines of Dill's work on the folding and aggregation of "HP chains" composed of only hydrophobic and polar moieties (Shortle et al., 1992; Yue and Dill, 1992). Each side of a monomer hexagon corresponds to a single average hydrophobic or hydrophilic residue. As such, the hexagons do not correspond to the projected surface area of any particular protein. However, by scaling the interaction energies, we may simulate any extent of surface area; the sides of the hexagon serve to orient the $H\Phi$ and HF surfaces with respect to one another rather than to limit their extent. In addition, ($H\Phi$, s), (HF, s), and (s, s) interactions are explicitly accounted for as described before.

The free energy changes accompanying the formation of each of the six interactions *in vacuo* are needed to carry out the simulations; the intermolecular interaction energies used are reported in Table 1.

Estimates for the $\Delta\Delta G(p,p)$ are available in the literature. These values were, in turn, used to estimate the $\Delta G(x, y)$ based on the following assumptions:

1) $\Delta G(s, s)$ was assumed close to the average free energy of hydrogen bond formation in water (Voet and Voet, 1990; Weiner et al., 1984);

$$\Delta G(s, s) \approx \Delta G(H\text{—bond}). \tag{13}$$

2) The (HF, s) interaction occurs via hydrogen bonding and should be similar to the (s, s) interaction. In addition, Creighton (1990) argues that water-water and protein-water hydrogen bonds are more transient than protein-protein hydrogen bonds. Extending that argument, we chose a slightly more negative value for (HF, s) interaction than for (s, s) interaction,

$$\Delta G(HF, s) \lesssim \Delta G(s, s). \tag{14}$$

3) The ($H\Phi$, s) interaction free energy is unfavorable and is assumed to be about the same as $\Delta\Delta G$ of ($H\Phi$, HF) interaction.

$$\Delta G(H\Phi, s) \approx \Delta\Delta G(H\Phi, HF). \tag{15}$$

A similar assumption was made by Fields et al. (1992) in their equilibrium aggregation work.

4) The formation of ($H\Phi$, $H\Phi$) interactions in solution is thought to be the major driving force for protein folding (Creighton, 1990; Kauzmann, 1959) and protein aggregation (Fields et al., 1992) and, therefore, should have a favorable contribution

$$\Delta\Delta G(H\Phi, H\Phi) = \Delta G(H\Phi, H\Phi) + \Delta G(s, s) - 2\Delta G(H\Phi, s) < 0. \tag{16}$$

The intense interest in the hydrophobic interaction has motivated many efforts to quantify $\Delta\Delta G(H\Phi, H\Phi)$ and as a result, there is wide variation in the reported values (Ben-Naim, 1990, 1991; Cantor and Schimmel, 1980; Chothia, 1974; Cornette et al., 1987; Crosio et al., 1990). Several of these efforts are based on the free energy of amino acid transfer from water to organic solvents (see Cornette et al. (1987); Miller et al. (1987); Wolfenden et al. (1981) and references therein) and by analysis of heat capacity data (Yang et al., 1992). These values are only approximate as it is unclear how closely any given solvent represents the hydrophobic core of a globular protein (Prevost et al., 1991) and moreover, as Ben-Naim (1990) pointed out conditional solvation effects are neglected in these calculations. We have used an estimate of −4.5 kcal/mol (a suitable average from (Chothia, 1974) and (Crosio et al., 1990)), for $\Delta\Delta G(H\Phi, H\Phi)$ as shown in Table 1 and later varied this from −1 kcal/mol (close to the value computed by Ben-Naim (1991)) to −5 kcal/mol in steps of −1 kcal/mol in a parametric study of the impact of $\Delta\Delta G(H\Phi, H\Phi)$ on aggregate structure. $\Delta G(H\Phi, H\Phi)$ can be calculated from $\Delta\Delta G(H\Phi, H\Phi)$.

5) The formation of (HF, HF) interactions in solution is assumed unfavorable as HF groups lose their solvation free energy (Ben-Naim, 1991; Yang et al., 1992)

$$\Delta\Delta G(HF, HF) = \Delta G(HF, HF) + \Delta G(s, s) - 2\Delta G(HF, s) > 0. \tag{17}$$

Ben-Naim (1991) reported a value of +6.4 kcal/mol for $\Delta\Delta G(HF, HF)$ for complete loss of solvation upon contact formation. $\Delta G(HF, HF)$ was computed from this value.

6) Mismatched protein-protein interactions are energetically unfavorable,

$$\Delta\Delta G(H\Phi, HF) = \Delta G(H\Phi, HF) + \Delta G(s, s) - \Delta G(HF, s) - \Delta G(H\Phi, s) > 0. \tag{18}$$

$\Delta\Delta G(H\Phi, HF)$ was reported to be +7.0 kcal/mol (Ben-Naim, 1991) and this was used to compute $\Delta G(H\Phi, HF)$.

It should be noted that the detailed accounting of individual interactions in terms of $\Delta G(x, y)$ values was based on available $\Delta\Delta G(p,p)$ values and is distinct from the hydrogen bond inventory argument, a method that is believed to yield questionable $\Delta\Delta G$ values (Ben-Naim, 1991). The set of $\Delta G(x, y)$ values was chosen such that their relative values satisfy the generally accepted notions expressed in assumptions 1–6. Small variations in the individual interaction energies do not alter the nature of our results qualitatively as they merely impact the *parsing* of the net $\Delta\Delta G(p, p)$ in Eq. 3 rather than the value itself.

TABLE 1 Interaction free energies in solution and in vacuo

	$\Delta\Delta G(p, p)$*		$\Delta G(x, y)$‡		
	$H\Phi$	HF	$H\Phi$	HF	S
	kcal/mol				
$H\Phi$	−4.5§		+15.5		
HF	+7.0	+6.4	+13.5	−0.6	
S			+7.0	−6.5	−6.0

* $\Delta\Delta G(p, p)$ corresponds to (p, p) contact formation in solution.

‡ $\Delta G(x, y)$ corresponds to (x, y) contact formation in vacuo.

§ $\Delta\Delta G(H\Phi, H\Phi)$ was also varied from −1 to −5 kcal/mol in steps of −1 kcal/mol for type 1 monomers, leading to corresponding changes in the $\Delta G(x, y)$ used.

Seven different $H\Phi$, HF surface configurations named as type 1 through type 7 have been used as model protein monomers and are shown in Fig. 2. These monomers represent the structure of protein that is competant for aggregation, and were held invariant during the course of simulation; no further configurational changes upon aggregation were considered. Simulations were run with each of the seven monomer configurations to explore the impact of the spatial distribution of interaction sites on aggregate structure. Monomers with fewer than three $H\Phi$ surfaces were not examined as the values of intermolecular interactions considered will not always form precipitates, on a statistical basis.

Solution, aggregate, and system energetics

Fig. 3 is a schematic showing the typical progress in ΔG^t_{sys} with respect to ξ. The inset is a plot of the typical variation of change in free energy for the solution and precipitate phases of the system as the lattice is filled. The data reported is averaged over 500 runs for a 64 × 64 lattice. The free energy of the precipitate phase always increases as favorable solvent-solvent interactions are replaced by relatively less favorable protein-solvent and protein-protein interactions. But, the increase is not linear because the number of protein-protein interactions increases as the lattice fills. The free energy of the solution phase decreases *linearly*, as a constant number of solvent-solvent interactions are substituted for a fixed set of protein-solvent interactions with each transfer. The slope of the ΔG^t_{sol} curve varies only with the ratio of hydrophobic to hydrophilic surfaces in the monomer. As ΔG^t_{ppt} and ΔG^t_{sol} have large, comparable magnitudes with opposite signs, ΔG^t_{sys} is relatively small. As can be seen from the schematic, ΔG_{sys} is initially positive, due to placements in the dilute lattice. It becomes negative as an increasing number of favorable (p,p) contacts are formed. The generated data for ΔG^t_{sys} is plotted in Fig. 4 *a* on a reduced scale so that the nature of the curve and differences between curves as a function of monomer type for the energetics given in Table 1 can be seen. Fig. 4 *b* is a similar plot for type 1 monomers as $\Delta\Delta G(H\Phi, H\Phi)$ is varied from −1 to −5 kcal/mol. Values plotted are averaged over 500 runs. It should be noted that these curves describe the progress in ΔG^t_{sys} for the simulation algorithm and not necessarily for the actual aggregation *process*. The final value of ΔG^t_{sys} obtained for each molecule type governs the final aggregate properties.

All ΔG^t_{sys} trajectories show a small rise initially due to placements in isolated lattice regions where ΔG_{config} dominates and this is analogous to a nucleation process. Both crystallization and aggregation processes are thought to proceed through an initial nucleation step; the nature of the nuclei may determine whether a crystal or an aggregate is obtained (Georgalis et al., 1993). As the number of favorable (p,p) contacts increases, the curve for ΔG^t_{ppt} starts to flatten out, and finally, the contribution from these contacts offsets the configurational entropy term and ΔG^t_{sys} starts to decrease. As more (p,p) contacts are formed, ΔG^t_{sys} becomes negative and eventually reaches a minimum. Any aggregation beyond this point results in more unfavorable contacts and the ΔG^t_{sys} curve starts to increase monotonically. The aggregate structure is complete at this point. Simulations, however, proceed beyond this point until the lattice is statistically saturated as described before and are then returned to this point. The apparent increase in the noise level of these curves toward the jamming-limit occurs because the data becomes averaged by fewer data points as fewer runs attain the higher jamming limits.

Although the free energy trajectories we obtained resemble those for nucleation-controlled aggregation pathways, our simulation pathway involving nonlocal transfer is clearly nonphysical. However, as we are investigating the structures of the final thermodynamic states of the aggregates and not the aggregation process, we may choose any convenient pathway to arrive at these states. The main assumption associated with our deterministic path selection is that

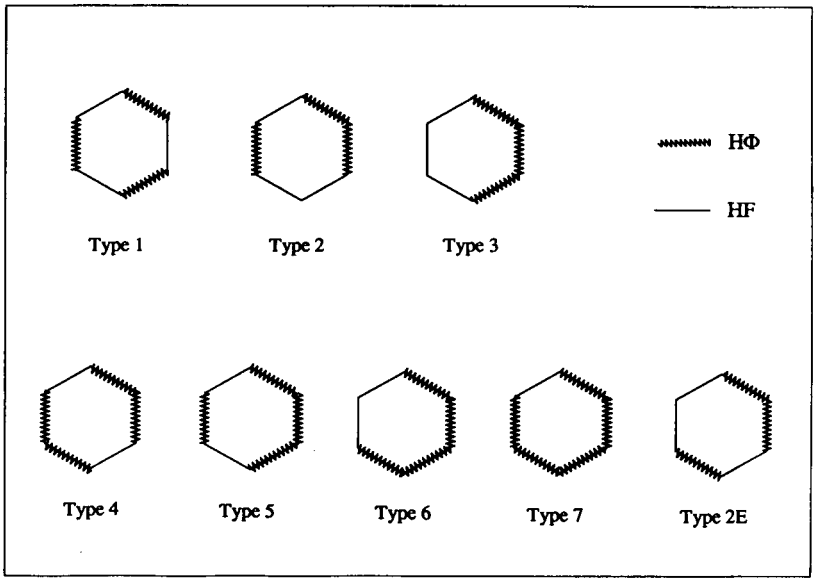

FIGURE 2 Monomer types considered. The enantiomer of the type 2 monomer is shown as type 2E.

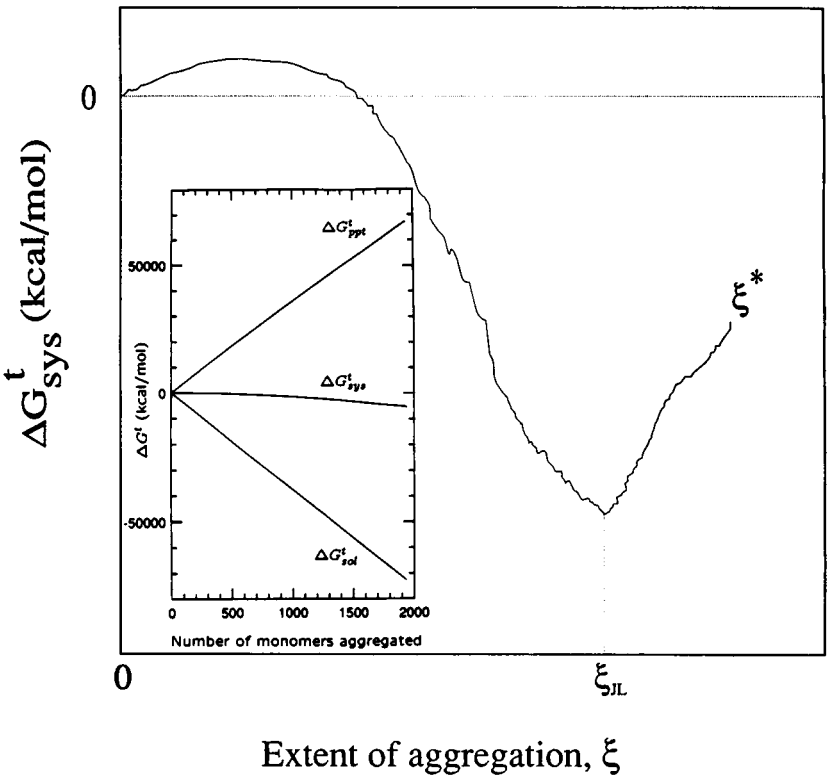

FIGURE 3 A schematic of ΔG^t_{sys} with respect to ξ. The curve is exaggerated for clarity. ξ_{JL} is the jamming limit and ξ^* is the saturation limit described in the text. The inset shows ΔG^t_{ppt}, ΔG^t_{sol}, and ΔG^t_{sys} trajectories for type 1 monomers averaged over 500 runs.

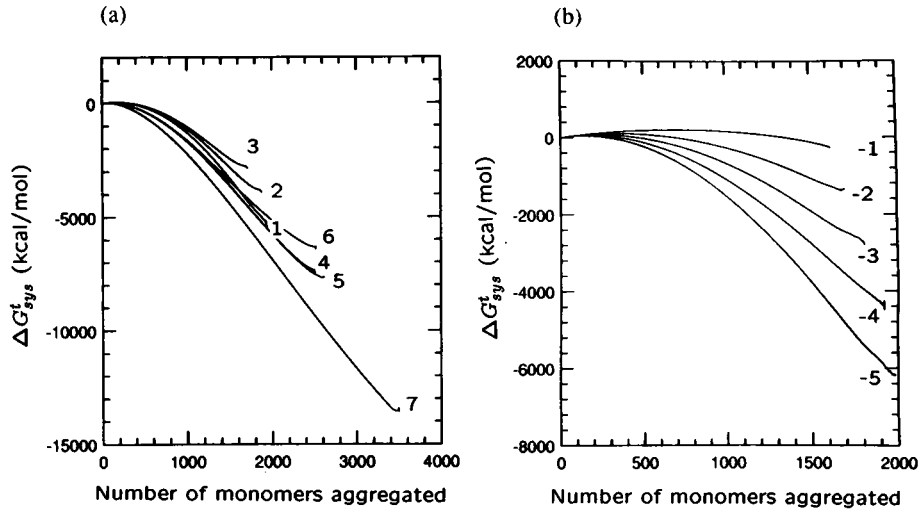

FIGURE 4 Variation of ΔG^t_{sys} with number of monomers transferred as a function of (*a*) monomer type and (*b*) $\Delta\Delta G(H\Phi, H\Phi)$. Labels correspond to monomer type in *a* and to values of $\Delta\Delta G(H\Phi, H\Phi)$ in kcal/mol in *b*. Note the initial rise in the curves for ΔG^t_{sys} due to placements in isolated regions of the lattice.

it terminates at a physically plausible final state that could have been reached by a local transfer pathway. By averaging over many such paths, we reduce the likelihood that any given path unduly biases the structure in the final state.

Simulation execution, lattice size, and number of runs

The computer code for the simulations was written in VAX FORTRAN and the simulations were carried out on a VAX cluster. The random number generator used plays a significant role in the simulation. To assure the generation of truely random numbers, and to eliminate any possible bias (Ferrenberg and Landau, 1992) the VAX random number generator was modified by a subroutine (Press et al., 1986).

Each run involves generating the aggregate structure from a lattice filled initially with solvent. To obtain statistically meaningful data, large numbers of runs must be averaged for each given configuration of the model protein molecule. The required number of runs depends on the size of the lattice due to the trade-off between spatial and run-to-run averaging. Also, the lattice size will restrict the maximum length scale of any structure formed. The lattice dimensions were chosen to minimize this effect and keep the computing-time tractable. A running average of the jamming limit in terms of the fractional lattice occupancy at $\xi = \xi_{JL}$, and denoted by $\langle JL \rangle$,

has been used as an index to examine the statistical fluctuations in the data for different lattice dimensions and for different numbers of runs. Lattice sizes of 16 × 16, 32 × 32, 64 × 64, and 128 × 128 were examined. As can be seen from Fig. 5, both 64 × 64 and 128 × 128 lattices show very little variation in ⟨JL⟩ at the end of 500 runs. A slight increase in the average jamming limit with increase in lattice size has been observed for smaller lattices, but the average jamming limit was about the same for both 64 × 64 and 128 × 128 lattices. The 64 × 64 lattice was chosen for all the simulations as a reasonable compromise between lattice size and computing time; the 128 × 128 lattice required more than six times the CPU time of the 64 × 64 lattice. Also, the standard deviation of ⟨JL⟩, when monitored as a function of the number of runs, reached a constant value well before 500 runs (data not shown), indicating that 500 runs were more than sufficient to define the true statistical distribution of the measured parameters. All the simulations were performed with 500 runs and the data averaged. The average CPU time for 500 runs with the 64 × 64 lattice is of the order of 24 h, with the actual time varying with the monomer type and energetics used.

Validation of configuration sensitivity of the model

The enantiomeric configuration for type 2, denoted as type 2E and shown in Fig. 2 was employed as a further check on the Monte-Carlo technique. We observed that all the data generated for type 2E aggregate simulations matched very closely with that for type 2 indicating that the observed surface distribution effects are due to the model protein type and are not compromised by numerical artifacts from the random number generator or the simulation technique used.

RESULTS AND DISCUSSION

Protein aggregate structure simulations were run for the seven monomer types shown in Fig. 2 using the interaction energies given in Table 1. Simulations were also run with $\Delta\Delta G(H\Phi, H\Phi) = -1$ to -5 kcal/mol using type 1 monomers. The simulated structures were analyzed with respect to jamming-limit, distribution of (p, p) contacts, solvent accessible perimeter, pore size distribution, and short- and long-range order. The monomer type and interaction energies used had a profound effect on these properties.

Fig. 6 shows a typical simulated aggregate structure for type 1 monomers with $\Delta\Delta G(H\Phi, H\Phi) = -5$ kcal/mol; only the first quadrant of the 64 × 64 lattice is plotted for clarity as no larger-scale aggregate structures were evident. Since order recognition from up to six unique monomer rotational orientations by visual inspection is difficult, we present only structures from type 1 monomer runs. The threefold symmetry of the type 1 model protein gives two unique orientations.

Jamming limit

The jamming limit provides the two-dimensional analog for aggregate density. Fig. 7 shows the dependence of $\Delta G'_{sys}$ on the jamming limit as a function of monomer type; here the jamming limit is expressed as the fraction of the total lattice area that is occupied by the protein. Each patch represents a collection of 500 data points for that particular monomer type. The presence of seven distinct patches indicates that the effects of both the extent and distribution of hydrophobic surfaces on the protein monomer are distinguishable. From Fig. 7 it is clear that the greater the monomer hydrophobic surface content, the greater the extent of aggregation and the more negative the resulting $\Delta G'_{sys}$. Increasing the monomer hydrophobic surface area increases the number of energetically favorable contacting surfaces.

A dispersed distribution of hydrophobic surfaces favors the formation of denser aggregates than does a single continuous hydrophobic patch. Monomer types 3 and 6 have contiguous hydrophobic patches and have lower jamming limits and less negative $\Delta G'_{sys}$ compared to the other types with equivalent hydrophobic contents. The presence of con-

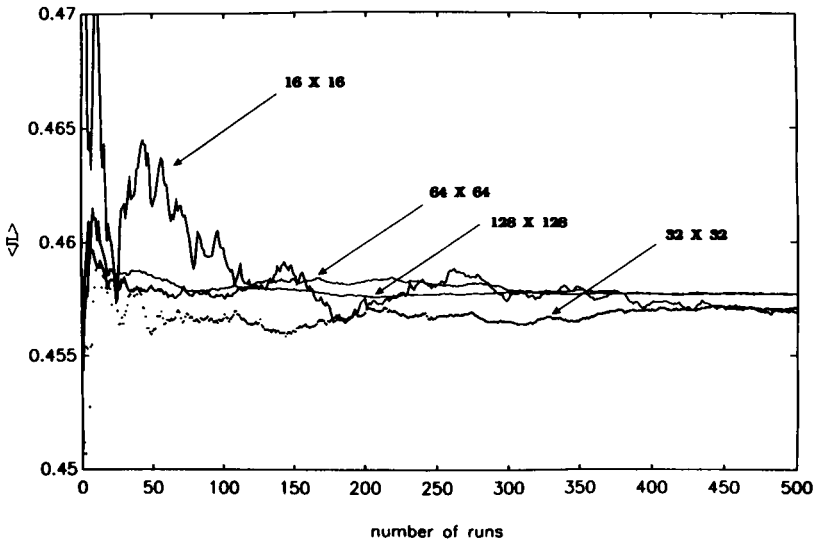

FIGURE 5 Effect of number of runs on the fractional lattice occupancy at the jamming limit as a function of lattice size for type 1 monomers with $\Delta\Delta G(H\Phi, H\Phi) = -4.5$ kcal/mol. Lattice sizes are marked on the figure. Notice that after 500 runs both 64 × 64 and 128 × 128 lattices level-off at about the same jamming limit. Note the dampening of the fluctuations as the lattice size increases.

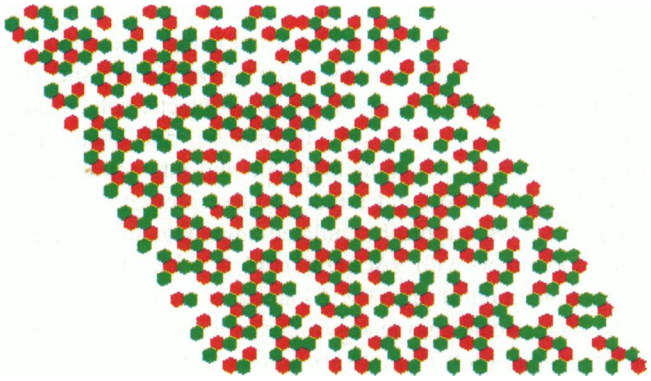

FIGURE 6 Typical simulated aggregate structure for type 1 monomers with $\Delta\Delta G(H\Phi, H\Phi) = -5$ kcal/mol. The red and green hexagons represent the two unique rotational orientations of type 1 monomers. Adjacent red and green hexagons represent ($H\Phi$, $H\Phi$) interactions exclusively.

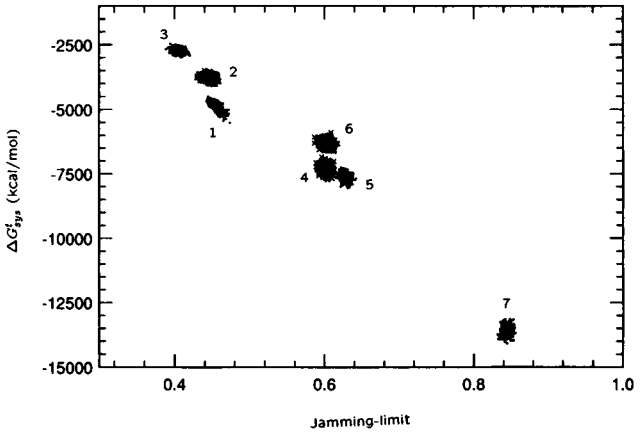

FIGURE 7 Variation of $\Delta G'_{sys}$ with jamming limit as a function of monomer type. Jamming limit is expressed as a fraction of the lattice occupied by protein. The labels correspond to the monomer types given in Fig. 2. Each patch represents a collection of 500 data points. Interaction energetics correspond to those given in Table 1.

tiguous hydrophobic patches decreases the system's degrees of freedom in the search for favorable contacting surfaces, yielding low jamming limits; one favorable contact between patches on adjacent monomers may preclude subsequent favorable contacts with additional monomers.

Fig. 8 is similar to Fig. 7, in this case showing the relationship between $\Delta G'_{sys}$ and the aggregate jamming limit for type 1 monomers as $\Delta\Delta G(H\Phi, H\Phi)$ is varied. Each patch again represents a collection of 500 data points obtained at a given value of $\Delta\Delta G(H\Phi, H\Phi)$. As in Fig. 7, a greater jamming limit generally results in a lower system free energy. Each patch describes an essentially linear relationship between $\Delta G'_{sys}$ and the jamming limit with a slope that is proportional to $\Delta\Delta G(H\Phi, H\Phi)$. The system attains ever lower ultimate free energies as $\Delta\Delta G(H\Phi, H\Phi)$ becomes more favorable because the majority of the (p, p) contacts are ($H\Phi$, $H\Phi$). The overall increase in jamming limit as $\Delta\Delta G(H\Phi, H\Phi)$ decreases is facilitated by the increased tolerance of unfavorable monomer placements; this effect is

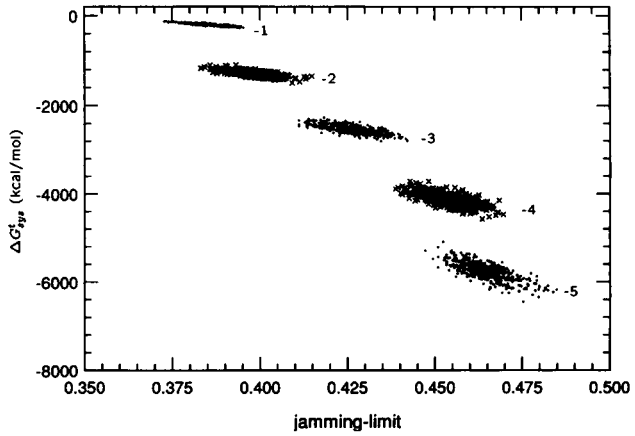

FIGURE 8 Variation of $\Delta G'_{sys}$ with jamming limit as a function of $\Delta\Delta G(H\Phi, H\Phi)$ for type 1 monomers. The labels give $\Delta\Delta G(H\Phi, H\Phi)$ in kcal/mol. All other interaction energetics as in Table 1.

also reflected in the increasing scatter in the individual patches as $\Delta\Delta G(H\Phi, H\Phi)$ decreases. However, the relationship between $\Delta G'_{sys}$, and the jamming limit is not a linear function of $\Delta\Delta G(H\Phi, H\Phi)$ in this case. The asymptotic jamming limit as $\Delta\Delta G(H\Phi, H\Phi)$ approaches an upper limit of $-\Delta G_{config}$ is finite and controlled by ΔG_{config} and its impact on the number of nuclei. The equilibrium protein aggregation model developed by Fields and co-workers (1992) also predicts a similar increase in the jamming limit when the hydrophobic fraction is increased and when the hydrophobic-hydrophobic interaction is made more favorable.

(p, p) contacts in aggregates

The nature and extent of (p, p) contacts within aggregates give insight into the energetic stabilization of the aggregate. Fig. 9 shows the variation of $\Delta G'_{sys}$ with the number of ($H\Phi$, $H\Phi$) contacts per monomer as a function of monomer type. As expected, the greater the number of ($H\Phi$, $H\Phi$) contacts per molecule in the aggregates, the lower the total free

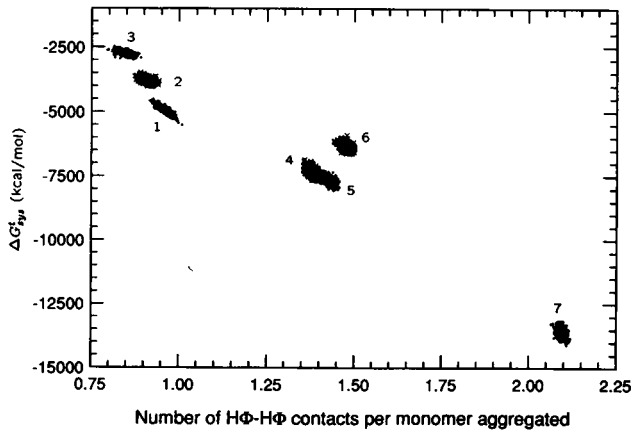

FIGURE 9 Relationship between $\Delta G'_{sys}$ and number of ($H\Phi$, $H\Phi$) contacts normalized by the total number of monomers aggregated as a function of monomer type. Labels refer to different monomer types.

energy attained. This strong correlation indicates that the primary mode of minimizing the system free energy is through burying the hydrophobic surfaces in the interior of the aggregate. This is consistent with the paradigm that the unfavorable solvation free energy for hydrophobic surfaces drives aggregation. Increasing the number of hydrophobic sides from three (monomer types 1, 2, and 3) to five (monomer type 7) results in an increased number of ($H\Phi$, $H\Phi$) contacts per protein molecule in the aggregates. It should be noted that the impact of monomer configuration is distinguishable and that the number of ($H\Phi$, $H\Phi$) contacts is roughly proportional to the square of the number of $H\Phi$ surfaces per monomer.

Fig. 10 is a plot of $\Delta G'_{sys}$ versus the number of ($H\Phi$, $H\Phi$) contacts per protein molecule aggregated as a function of $\Delta\Delta G(H\Phi, H\Phi)$ for type 1 monomers. Within every collection of data points, the higher the number of ($H\Phi$, $H\Phi$) contacts per molecule, the lower the resulting system free energy. As $\Delta\Delta G(H\Phi, H\Phi)$ becomes more negative, more unfavorable contacts can be accommodated, increasing the scatter in the data. The slopes of the patches of data points are again proportional to $\Delta\Delta G(H\Phi, H\Phi)$. For $\Delta\Delta G(H\Phi, H\Phi) = -1$ kcal/mol, most of the (p, p) contacts are ($H\Phi$, $H\Phi$) and very strong correlation is seen between the number of ($H\Phi$, $H\Phi$) contacts per molecule and the system free energy; the decrease in system free energy accompanying the increase in number of ($H\Phi$, $H\Phi$) contacts per molecule is small, but definite in this case. Fig. 10 also shows that the number of ($H\Phi$, $H\Phi$) contacts per protein molecule is sterically limited by the distribution of $H\Phi$ patches on the monomer. The average number of ($H\Phi$, $H\Phi$) contacts saturates in the neighborhood of one per monomer for $\Delta\Delta G(H\Phi, H\Phi) = -4$ kcal/mol; further decreases in $\Delta\Delta G(H\Phi, H\Phi)$ do not have an appreciable impact.

The distributions of (p, p) contacts in the aggregates were calculated and averaged over 500 runs for each monomer type and are shown in Table 2. For all types of monomers, ($H\Phi$, $H\Phi$) contacts are dominant. This may be expected from the relative numbers of $H\Phi$ and HF sides in the monomers considered from a mass action standpoint. The probability of forming each type of (p, p) contact in isolation has been computed and the values are shown in Table 2 for comparison. Multiple interaction effects and steric constraints in the growing aggregate limit the number of ($H\Phi$, $H\Phi$) contacts that actually occur as the pair-wise interaction frequencies always exceed the simulation results. Except for monomer type 1, about 81–86% of the contacts in the interior are found to be ($H\Phi$, $H\Phi$). This indicates that an increase in the number of hydrophobic surfaces on the monomer does not necessarily imply a corresponding increase in the fraction of ($H\Phi$, $H\Phi$) contacts in aggregates.

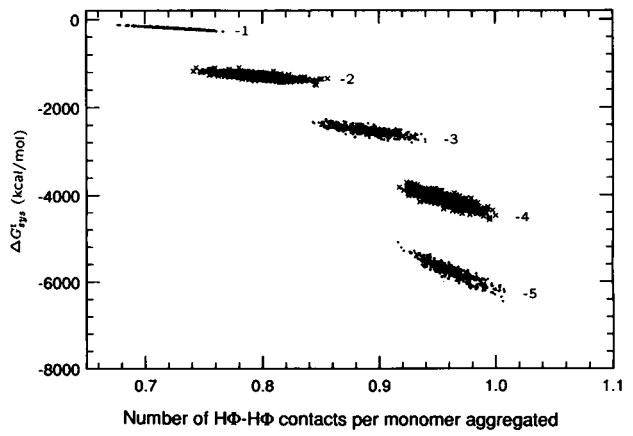

FIGURE 10 Relationship between $\Delta G'_{sys}$ and normalized number of ($H\Phi$, $H\Phi$) contacts as a function of $\Delta\Delta G(H\Phi, H\Phi)$ for type 1 monomers. Labels refer to $\Delta\Delta G(H\Phi, H\Phi)$ in kcal/mol.

With the exception of monomer types 2, 3, and 6, the number of ($H\Phi$, HF) contacts exceeds that of (HF, HF) contacts, although the latter is more favorable energetically in solution than the former. This conforms with the notion that hydrophilic groups prefer to remain solvated (Ben-Naim, 1991). But whenever there is a large contiguous patch of hydrophobic surfaces that forces a segregation of ($H\Phi$, $H\Phi$) and (HF, HF) contacts, as in monomer types 2, 3, and 6, (HF, HF) contacts occur more frequently than ($H\Phi$, HF) contacts. The more extensive the hydrophobic patch, the more segregation seems to occur for a given extent of hydrophobic surface area per monomer; the percentage of (HF, HF) contacts increases continuously from monomer type 1 through 3 and from monomer type 4 through 6. For type 7 monomers, however, we see that ($H\Phi$, HF) contacts outnumber the (HF, HF) contacts. The extent of the monomer hydrophobic surface area for type 7 monomer is so large that the jamming limit is reached well before many of the hydrophilic groups are forced out of solution. This mass action effect is qualitatively supported by the calculations shown in Table 2. This clearly confirms the importance of multiple interaction effects resulting from the combination of monomer configuration and interaction energetics. Thus we see that the orientation of hydrophobic patches along the monomer surface impacts the overall free energy change through the distribution of (p, p) contacts. This impact is greater than that of $\Delta\Delta G(H\Phi, H\Phi)$.

The effect of the magnitude of $\Delta\Delta G(H\Phi, H\Phi)$ on the (p, p) contact distribution is also shown in Table 2. As noted earlier, the vast majority of the contacts in the aggregates, more than 92%, are ($H\Phi$, $H\Phi$) in nature for all sets of energetics. As $\Delta\Delta G(H\Phi, H\Phi)$ becomes more favorable, the percentage of ($H\Phi$, $H\Phi$) contacts actually decreases as it becomes easier to offset unfavorable contacts. In the limit as $\Delta\Delta G(H\Phi, H\Phi) \rightarrow -\Delta G_{config}$, the jamming limit, and hence, the number of (p, p) interactions approaches a minimum. Thus, the distribution of (p, p) contacts approaches that calculated from mass action considerations.

Solvent accessible surface area

The aggregate solvent accessible surface area (SAS) provides insight into the driving force for aggregation and determines the nature of the surface that must be wetted in

TABLE 2 Distribution of (p, p) contacts as a function of monomer type and $\Delta\Delta G(H\Phi, H\Phi)$

	Monomer type	$\Delta\Delta G(H\Phi, H\Phi)$	(HF, HF)	$(HF, H\Phi)$	$(H\Phi, H\Phi)$
		kcal/mol	%	%	%
Simulated results	1	−4.5	0.49	6.15	93.37
	2	−4.5	9.45	4.97	85.58
	3	−4.5	11.83	4.95	83.22
	4	−4.5	4.96	9.85	85.20
	5	−4.5	7.01	7.82	85.18
	6	−4.5	9.87	8.37	81.76
	7	−4.5	5.01	11.23	83.76
	1	−1.0	0.0023	0.0046	99.99
	1	−2.0	0.008	0.069	99.92
	1	−3.0	0.037	1.522	98.44
	1	−4.0	0.22	5.7	94.08
	1	−5.0	1.0	6.94	92.06
Calculated results*	1, 2, 3		2.023×10^{-3}	7.344×10^{-4}	99.997
	4, 5, 6		5.056×10^{-4}	3.672×10^{-4}	99.9991
	7		8.091×10^{-5}	1.469×10^{-4}	99.9998

* Probability of pair-wise contacts made in isolation.

redissolution and suspension processes. Since solvent-exposed $H\Phi$ surface area is energetically costly, ΔG^{t}_{sys} increases with the extent of $H\Phi$ surface exposure. This may be inferred from Figs. 9 and 10 as the number of $(H\Phi, H\Phi)$ contacts per monomer and the aggregate solvent accessible $H\Phi$ perimeter vary inversely. This trend becomes more apparent at lower monomer hydrophobic surface contents. From 47 to 68% of the monomer hydrophobic surfaces were buried in the aggregate. Similar results were observed as a function of $\Delta\Delta G(H\Phi, H\Phi)$. As $\Delta\Delta G(H\Phi, H\Phi)$ decreases, the solvent exposed hydrophobic surface area decreases, indicating that small fractions of exposed hydrophobic surface area accompany high jamming-limits.

The computed loss in total SAS due to aggregation for monomer types 1 through 3, with three $H\Phi$ surfaces per monomer, were 34, 35, and 34%; the loss in SAS for type 1 monomers as a function of $\Delta\Delta G(H\Phi, H\Phi)$ varied between 24 and 35%. Crosio and co-workers (1990) have reported the loss of overall SAS during crystallization for RNase A and for RNase S to be 25 and 40%, respectively. If we attempt to match the monomer hydrophobicity of RNase with one of the model protein configurations, it very approximately corresponds to about three hydrophobic sides per hexagon (Miller et al., 1987). The loss in SAS with aggregation increases with increasing hydrophobic surface fractions; for monomer types 4, 5, and 6, with 66.67% hydrophobic SAS, the computed losses were 54, 56, and 60%, respectively. Similar data for a set of 39 monomeric proteins was reported to vary from 14.3 to 51.8% (Islam and Weaver, 1990). Aggregation, evidently leads to a greater loss in SAS than crystallization as many more unfavorable contacts occur.

Aggregate porosity

The overall porosity and distribution of pores within the aggregates relate to the ease with which aggregates may be redissolved. The aggregate porosity is inversely proportional to the jamming limit which was discussed earlier. The pore size distribution as a function of monomer type is shown in Fig. 11. As the monomer hydrophobic surface area increases from 50% for type 1 to 83.33% for type 7 monomers, the size of the largest pore decreases dramatically as shown in Fig.

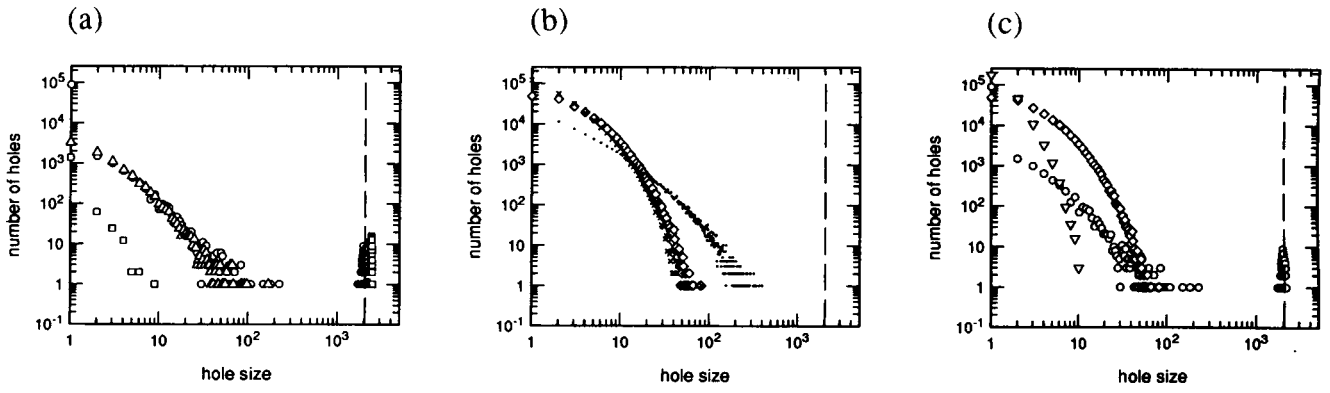

FIGURE 11 Frequency of pore size occurrence totalled over 500 runs versus pore size for different monomer types. Pore-size is measured by number of lattice points included in a single continuous pore. In *a* and *b*, pore-size distributions of monomers with equal $H\Phi$ content are compared to demonstrate the influence of surface distribution. In *c*, the influence of the extent of monomer hydrophobic surface area is presented. The symbols ○, △, □, ◇, ×, ●, and ▽ represent monomer types 1 through 7, respectively. The percolation limit, at a pore size of 2048, is shown as a dashed line.

11 *c*. Also, as the monomer hydrophobic fraction increases, the pore size distribution becomes almost continuous up to the largest observed pore size for that monomer type. For monomer types 4 through 7 the distribution was continuous while for monomer types 1 through 3 it was discontinuous.

For a given monomer hydrophobic content, the distribution of the hydrophobic groups also has an effect on the pore size distribution as shown in Fig. 11, *b* and *c*. The more contiguous the $H\Phi$ surfaces, the larger the size of the biggest pore. This effect was dominant at low $H\Phi$ contents, where the largest pore tends to percolate, or span the entire lattice, while the aggregates themselves typically do not percolate. From fractal geometric considerations, the minimum percolating pore size for a two-dimensional, 64 × 64 hexagonal lattice is 2048 (Feder, 1991) and is shown as a dashed line in Figs. 11 and 12. For type 3 monomers, where all the $H\Phi$ surfaces are contiguous, a large percolating pore, ranging in size from 2377 to 2500, was observed in nearly all simulations accompanied by a few very small pores of size less than 10. For monomer types 1 and 2, the pore size distribution was essentially continuous upto about 100 and exhibited a reduced frequency of percolating pores. As the hydrophobic content of the monomer increases, the effect of surface distribution seems to be mitigated; for monomer types 4, 5 and 6 the pore size distributions were continuous and similar except that the distribution for type 6 extended to a pore of size 399 while those for types 4 and 5 had no pores of size more than 100. The largest pore observed for type 7 monomer was of size 10 (see Fig. 11 *c*) and was observed only three times in 500 runs. This occurred because at increased monomer $H\Phi$ contents, high jamming-limits dominate the surface distribution effect.

Fig. 12 represents the pore-size distribution as a function of $\Delta\Delta G(H\Phi, H\Phi)$. As $\Delta\Delta G(H\Phi, H\Phi)$ decreases, the distribution becomes less discontinuous and the average size of the largest pore decreases. For all sets of energetics considered, the distribution has essentially the same characteristics; a few small pores followed by a large pore that usually percolates. The frequency of occurrence of the percolating pore, however, decreases as $\Delta\Delta G(H\Phi, H\Phi)$ becomes more favorable. For type 1 monomers with $\Delta\Delta G(H\Phi, H\Phi) = -5$ kcal/mol, the size of the largest pore rarely attained the percolation limit.

The nature of the pore-size distribution can also be altered by varying the configurational entropy contribution to ΔG_{sys}. A small decrease in the entropic penalty was found to make the distribution more continuous (results not shown). The effect of ΔG_{config} is more pronounced than the effect of $\Delta\Delta G(H\Phi, H\Phi)$ for type 1 monomers.

Configurational free energy

The configurational entropic penalty directly impacts the frequency of nucleation events. Although ΔS_{config} is properly a

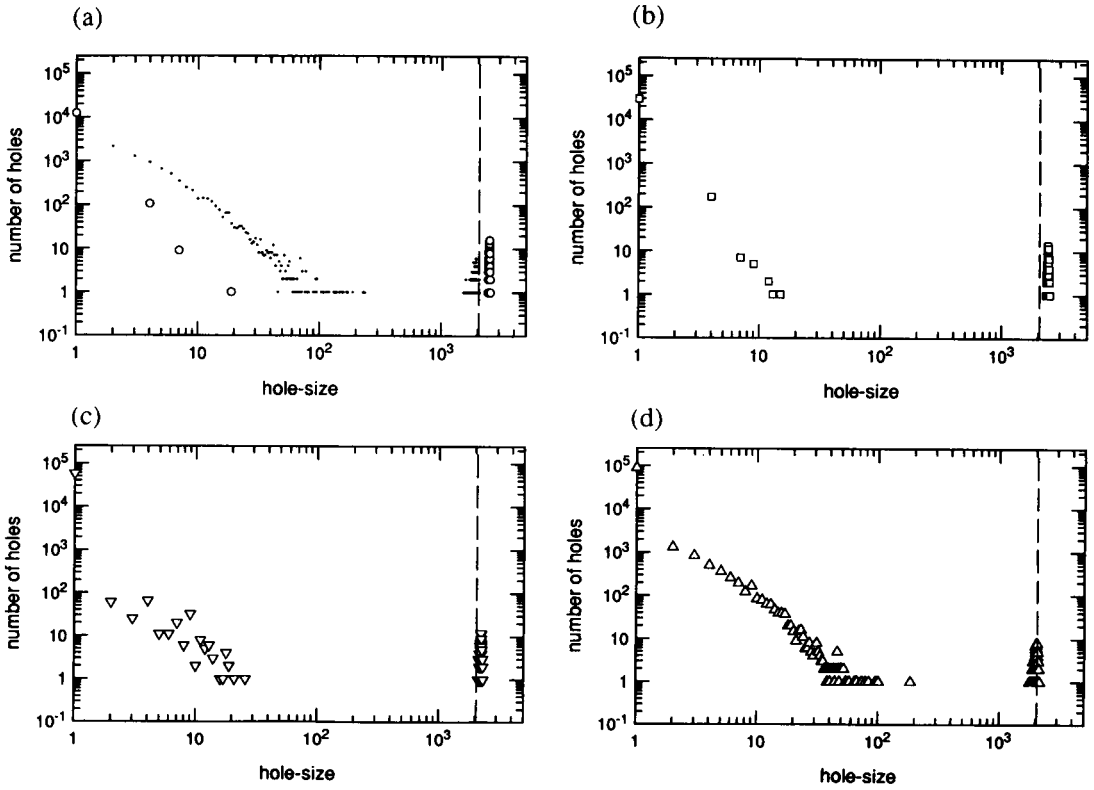

FIGURE 12 Frequency of pore size occurrence totalled over 500 runs versus pore size for type 1 monomers with $\Delta\Delta G(H\Phi, H\Phi)$ values of (*a*) −1 (○) and −5 (●) kcal/mol, and distributions for $\Delta\Delta G(H\Phi, H\Phi)$ = (*b*) −2, (*c*) −3, and (*d*) −4 kcal/mol. The percolation limit, at a pore size of 2048, is shown as a dashed line.

function of the protein concentration (Fields et al., 1992; Flory, 1953), we assumed a constant, order of magnitude value. However, in order to examine the impact of this term on aggregate structure, a range of values was also used. Increases in the entropic penalty resulted in a greater fraction of ($H\Phi$, $H\Phi$) contacts, lower jamming limits and less continuous hole-size distributions. Increasing the entropic penalty, therefore, is equivalent to making $\Delta\Delta G(H\Phi, H\Phi)$ less favorable in that it reduces the ability to compensate for unfavorable contacts. In addition, increasing the entropic penalty resulted in increased short-range order for all monomer types and for all sets of interaction energies. A practical upper bound to ΔG_{config} is given by $-\Delta\Delta G(H\Phi, H\Phi)$. For $\Delta G_{config} \geq -\Delta\Delta G(H\Phi, H\Phi)$, the energetic gain from ($H\Phi$, $H\Phi$) contact formation cannot offset the entropic penalty.

Short-range and long-range order

The involvement of specific interactions in protein aggregates gives rise to the possibility that aggregates may possess elements of crystallinity. Visual inspection of the simulated aggregate structures revealed traces of short- and long-range order. The jamming limit and (p, p) contact distribution data were used to achieve a qualitative comparison of order among structures belonging to different monomer types with equivalent $H\Phi$ contents. Higher jamming limits roughly correspond to structures with higher long-range order. A greater proportion of ($H\Phi$, $H\Phi$) contacts implies greater short-range order. Among monomer types 1, 2, and 3, structures arising from type 1 monomers showed the greatest long- and short-range order and those from type 3 exhibited the least. Although the monomer surface distribution and order at higher monomer $H\Phi$ contents were not strongly correlated, lumping all the hydrophobic surfaces in one stretch seems to decrease both long- and short-range order in aggregates. Type 6 monomers exhibited the lowest jamming limit and lowest proportion of ($H\Phi$, $H\Phi$) contacts compared to monomers of types 4 and 5.

The influence of $\Delta\Delta G(H\Phi, H\Phi)$ on aggregate order is more pronounced. Structures obtained for $\Delta\Delta G(H\Phi, H\Phi) = -1$ and -2 kcal/mol have almost uniformly favorable (p,p) contacts showing high short-range order, but are too dispersed to exhibit long-range order. They showed linear chains, rather than the rings that are characteristic of type 1 monomer structures with high long-range order. Structures for $\Delta\Delta G(H\Phi, H\Phi) = -4$ and -5 kcal/mol showed more long-range order, i.e., more rings, but some of the contacts are mis-matched resulting in lower short-range order than the structures obtained for less favorable values of $\Delta\Delta G(H\Phi, H\Phi)$. This is consistent with the belief that crystallization is favored by small binding energies (Durbin and Feher, 1991). Large interaction energies may lead to the irreversible binding of proteins in improper orientations, causing aggregates.

Aggregate structures were biased towards higher short-range order by increasing the entropic penalty term ΔG_{config}. This was confirmed by a corresponding increase in the proportion of ($H\Phi$, $H\Phi$) contacts found for all monomer types and $\Delta\Delta G(H\Phi, H\Phi)$ values examined. A high entropic penalty not only reduces the number of nucleation sites but also allows only highly favorable (p, p) contacts, thus decreasing the jamming limits. A notable exception to the apparent trade-off between long- and short-range order is exhibited by the type 1 monomers; in this case, the interaction between the symmetry of the monomer and that of the lattice enabled concomitant increases in long- and short-range order.

These observations lead us to an interesting hypothesis: Although small interaction energies enhance short-range order, they cannot promote long-range order. If more favorable binding energies are used to obtain higher long-range order, the resulting loss in short-range order can be overcome by increasing the entropic penalty. In other words, a high entropic penalty coupled with favorable energetics discourages the placement of molecules in improper orientations and should improve both short- and long-range order. We tested this hypothesis on type 1 monomers with $\Delta\Delta G(H\Phi, H\Phi) = -5$ kcal/mol. The highly symmetric nature of the molecule enables a regular formation of ($H\Phi$, $H\Phi$) contacts and consequently structures obtained by increasing the entropic penalty exhibited extensive long-range order. We have observed a significant increase in short-range order as well, as ΔG_{config} increases for this system. Fig. 13 shows representative aggregate structures obtained for a type 1 monomer with $\Delta\Delta G(H\Phi, H\Phi) = -5$ kcal/mol as ΔG_{config} is varied from $+1\ RT$ kcal/mol to $+\ 10\ RT$ kcal/mol in increments of $+\ 3\ RT$ kcal/mol. Only the first quadrant is shown in each case. For type 1 monomers, the number of pores of size 1 is indicative of the percent crystallinity or the extent of short-range order. By this measure, the short-range order of the aggregates shown in Fig. 13, *a–d*, was 19, 44, 71, and 85%, respectively. Although the quantification of percent crystallinity for other monomer types is more difficult, an improvement in short-range order was clearly observed.

CONCLUSIONS

The extent and orientation of monomer hydrophobic surface area, the magnitude of (p, p) interaction energies and the configurational entropy loss significantly impact the structure of kinetically irreversible protein aggregates. These impacts were manifested in terms of the free energy change of the system and the following aggregate structural characteristics: jamming limit, distribution of (p, p) contacts, solvent accessible hydrophobic surface area, porosity, and short- and long-range order.

An increase in the monomer $H\Phi$ content resulted in an increase in the aggregate density accompanied by an increase in the number of ($H\Phi$, $H\Phi$) contacts and a decrease in the $H\Phi$ surface that is exposed to the solvent. These findings reflect the central role that the ($H\Phi$, $H\Phi$) interaction plays in determining aggregate structure. This is in accord with expectations since $\Delta\Delta G(H\Phi, H\Phi)$ was the only favorable protein-protein interaction in solution, in our simulations.

It was also expected that the configuration of a monomer for a given $H\Phi$ content would impact aggregate structure.

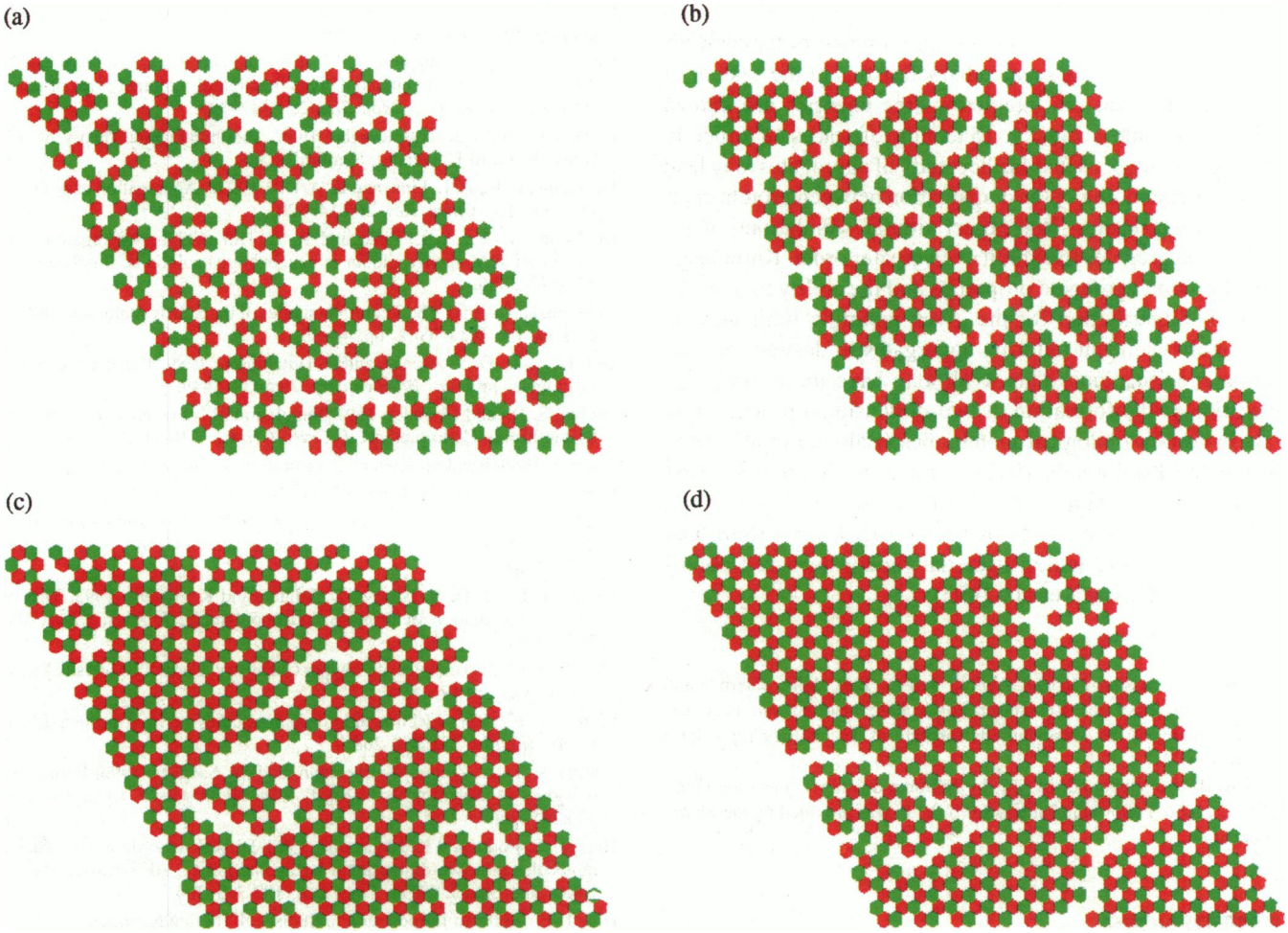

FIGURE 13 Representative aggregate structures for type 1 monomers with $\Delta\Delta G(H\Phi, H\Phi) = -5$ kcal/mol with ΔG_{config}, of (*a*) 1*RT*, (*b*) 4*RT*, (*c*) 7*RT*, and (*d*) 10*RT*. The approximate percent crystallinity (or short range order) for these structures is 19, 44, 71, and 85%, respectively. Note that all the (*p*, *p*) contacts in the highly ordered structures are ($H\Phi$, $H\Phi$).

However, this impact was more extensive than anticipated and could be resolved in terms of the adjacency of monomer $H\Phi$ patches. Pair-wise ($H\Phi$, $H\Phi$) interactions within aggregate structures for monomers with contiguous $H\Phi$ patches impose steric constraints on subsequent protein-protein interactions, leading to significant differences in aggregate density and order.

Several seemingly counter-intuitive phenomena were also noted. Increasing the monomer hydrophobic content did not necessarily increase the proportion of ($H\Phi$, $H\Phi$) contacts formed. Although $\Delta\Delta G(H\Phi, HF) > \Delta\Delta G(HF, HF)$, more ($H\Phi$, HF) contacts were found than (HF, HF) contacts for monomers with relatively dispersed $H\Phi$ patches. A more favorable $\Delta\Delta G(H\Phi, H\Phi)$ resulted in a lower proportion of ($H\Phi$, $H\Phi$) contacts.

The short- and long-range order of the aggregate structure could be controlled by manipulating $\Delta\Delta G(H\Phi, H\Phi)$ and ΔG_{config}. An increase in either of these values leads to a corresponding increase in short-range order at the expense of long-range order. Both short- and long-range order may be enhanced by increasing ΔG_{config} to maintain a high short-range order and decreasing $\Delta\Delta G(H\Phi, H\Phi)$ to ensure long-range order. This is a potential strategy to produce more crystalline structures.

Our model differs from earlier models mainly in its ability to account for the intermolecular interactions between monomers. Stenberg and Nygren (1991) have used a hexagonal lattice in their stochastic cellular automaton to study surface induced aggregation of ferritin. The hexagonal lattice was used only to decide the direction of the random time step and all protein-protein interactions were identical. Higo and coworkers (1992) have recently used a two-dimensional hexagonal lattice to model the effect of intermolecular interactions on the re-ordering of close-packed protein molecules in two-dimensional solid phases. They identified several solid phase transitions as a function of lattice temperature and monomer deviations from threefold symmetry. While their work addressed the structure of relaxed random aggregate states, we address the structure of nonequilibrium aggregate states.

Our two-dimensional model highlights many commonly observed properties of protein aggregates and several other interesting qualitative trends despite its simplicity. An extension of this model concept to computationally intensive

three-dimensional lattices is under development. The framework developed can be used to suggest a first order approximation of protein aggregate characteristics in the system of interest, provided the interaction free energies are known with reasonable accuracy. Detection of traces of order in aggregates suggests that further study of aggregates may lead to an understanding of the demarcation between protein crystals and aggregates. We expect similar simulations of reversible aggregates to exhibit even greater order. Knowledge about protein aggregate properties and the ability to alter the pore-size distribution for the given jamming limit may be extremely useful in tailoring aggregates to desired redissolution specifications. This work may have more direct application to the formation of two-dimensional protein crystals in certain biological membranes (Baldwin et al., 1988; Wang and Kuhlbrandt, 1991), at air-water interfaces (Darst et al., 1991; Fromherz, 1971; Sato et al., 1993; Ku et al., 1993), on the outer coats of bacteria (S layers) (Sara and Sleytr, 1987) and within bacterial rhodopsin based photoreceptors (Miyasaka et al., 1992).

We thank Professors Michael Abbott, Jeffrey Bell, John Salerno, and Georges Belfort of the Rensselaer Polytechnic Institute and Professor Jonathan King of the Massachusetts Institute of Technology for helpful discussions.

This work was supported, in part, by National Science Foundation Grant CTS-9211666. Fellowship support for S. Y. P. was provided by the Merck Foundation.

REFERENCES

Baldwin, J. M., R. Henderson, F. Beckman, and F. Zemlin. 1988. Images of purple membrane at 2.8 Å resolution obtained by cryo-electron microscopy. *J. Mol. Biol.* 202:585–591.

Becker, G. W., R. R. Bowsher, W. C. Mackellar, M. L. Poor, P. M. Tackitt, and R. M. Riggin. 1987. Chemical, physical and biological characterization of biosynthetic human growth hormone. *Biotech. Appl. Biochem.* 9:478–487.

Ben-Naim, A. 1990. Solvent effects on protein association and protein folding. *Biopolymers.* 29:567–596.

Ben-Naim, A. 1991. The role of hydrogen bonds in protein folding and protein association. *J. Phys. Chem.* 95:1437–1444.

Bowden, G. A., and G. Georgiou. 1990. Folding and aggregation of β-lactamase in the periplasmic space of *Escherichia coli*. *J. Biol. Chem.* 265:16760–16766.

Brems, D. N. 1988. Solubility of different folding conformers of bovine growth hormone. *Biochemistry.* 27:4541–4546.

Brems, D. N., L. A. Alter, M. J. Beckage, R. E. Chance, R. D. DiMarchi, L. K. Green, H. B. Long, A. H. Pekar, J. E. Shields, and B. H. Frank. 1992. Altering the association properties of insulin by amino acid replacement. *Protein Eng.* 5:527–533.

Cantor, C. R., and P. R. Schimmel. 1980. Biophysical Chemistry Part I: The Configuration of Biological Macromolecules. W. H. Freeman and Company, San Francisco, CA. pp. 285–288.

Casal, H. L., U. Kohler, and H. H. Mantsch. 1988. Structural and conformational changes of β-lactoglobulin B: an infrared spectroscopic study of the effect of pH and temperature. *Biochim. Biophys. Acta.* 957:11–20.

Chothia, C. 1974. Hydrophobic bonding and accessible surface area in proteins. *Nature (Lond.).* 248:338–339.

Cornette, J. L., K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky, and C. DeLisi. 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* 195:659–685.

Creighton, T. E. 1990. Protein folding. *Biochem. J.* 270:1–16.

Crosio, M., F. Rodier, and M. Jullien. 1990. Packing forces in Ribonuclease crystals. *FEBS Lett.* 271:152–156.

Darst, S. A., E. W. Kubalek, A. M. Edwards, and R. D. Kornberg. 1991. Two-dimensional and epitaxial crystallization of a mutant form of yeast RNA polymerase II. *J. Mol. Biol.* 221:347–357.

de Donder, and van Rysselberghe. 1936. Thermodynamic Theory of Affinity. Stanford University, Stanford, CA.

De Young, L. R., A. L. Fink, and K. A. Dill. 1993a. Aggregation of globular proteins. *Acc. Chem. Res.* 26:614–620.

De Young, L. R., A. L. Fink, and K. A. Dill. 1993b. Aggregation and denaturation of apomyoglobin in aqueous urea solutions. *Biochemistry.* 32:3877–3886.

Denbigh, K. G. 1950. The Thermodynamics of the Steady State. John Wiley & Sons, Inc. New York. pp. 44–46.

Dill, K. A., D. O. V. Alonso, and K. Hutchinson. 1989. Thermal stabilities of globular proteins. *Biochemistry.* 28:5439–5449.

Durbin, S. D., and G. Feher. 1991. Simulation of lysozyme crystal growth by the Monte-Carlo method. *J. Cryst. Growth.* 110:41–51.

Feder, J. 1980. Random sequential adsorption. *J. Theor. Biol.* 87:237–254.

Feder, J. 1991. Fractals. Plenum Press, New York. p. 112.

Ferrenberg, A. M., and D. P. Landau. 1992. Monte Carlo simulations: Hidden errors from "good" random number generators. *Phys. Rev. Lett.* 69: 3382–3384.

Fields, G. B., D. O. V. Alonso, D. Stigter, and K. A. Dill. 1992. Theory for the aggregation of proteins and copolymers. *J. Phys. Chem.* 96: 3974–3981.

Flory, P. J. 1953. Principles of Polymer Chemistry. Cornell University Press, Ithaca, NY. pp. 495–539.

Fromherz, P. 1971. Electron microscopic studies of lipid protein films. *Nature (Lond.).* 231:267–268.

Georgalis, Y., A. Zouni, W. Eberstein, and W. Saenger. 1993. Formation dynamics of protein precrystallization fractal clusters. *J. Cryst. Growth.* 126:245–260.

Higo, J., S. Endo, and K. Nagayama. 1992. Directional ordering of protein molecules in a two-dimensional hexagonal lattice. An approach with a poker chip model. *Chem. Phys. Lett.* 198:300–304.

Hill, T. L. 1960. An Introduction to Statistical Thermodynamics. Addison-Wesley Publishing Company, Menlo Park, CA. 1960. p. 44.

Islam, S. A., and D. L. Weaver. 1990. Molecular interactions in protein crystals: Solvent accessible surface and stability. *Protein Struct. Funct. Genet.* 8:1–5.

Jullien, R., and P. Meakin. 1989. Concentration effects in the off-lattice random ballistic deposition model. *J. Phys. A: Math. Gen.* 22: L1115–L1119.

Katchalsky, A., and P. F. Curran. 1965. Nonequilibrium Thermodynamics in Biophysics. Harvard University Press, Cambridge, Massachusetts.

Kauzmann, W. 1959. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* 14:1–63.

Kiefhaber, T., R. Rudolf, H. Kohler, and J. Buchner. 1991. Protein aggregation in vitro and in vivo: a quantitative model of the kinetic competition between folding and aggregation. *Bio/Technology.* 9:825–829.

Ku, A. C., S. A. Darst, C. R. Robertson, A. P. Gast, and R. D. Kornberg. 1993. Molecular analysis of two-dimensional protein crystallization. *J. Phys. Chem.* 97:3013–3016.

Lewis, U. J., E. V. Cheever, and B. K. Seavey. 1969a. Aggregate-free human growth hormone. I. Isolation by ultrafiltration. *Endocrinology.* 84: 325–331.

Lewis, U. J., D. C. Parker, M. D. Okerlund, R. M. Boyar, M. Litteria, and W. P. Vanderlaan. 1969b. Aggregate-free human growth hormone. II. Physico-chemical and biological properties. *Endocrinology.* 84:332–339.

Matthews, B. W. 1968. Solvent content of protein crystals. *J. Mol. Biol.* 33:491–497.

Miller, S., J. Janin, A. M. Lesk, and C. Chothia. 1987. Interior and surface of monomeric proteins. *J. Mol. Biol.* 196:641–656.

Mitraki, A., C. Haase-Pettingell, and J. King. 1991. Mechanism of Inclusion Body Formation *In* Protein Refolding. G. Georgiou, and E. De Bernardez-Clark, editors. American Chemical Society, Washington, DC. 35–49.

Miyasaka, T., K. Koyama, and I. Itoh. 1992. Quantum conversion and image detection by a bacteriorhodopsin-based artificial photoreceptor. *Science (Wash. DC).* 255:342–344.

Moore, W. V., and P. Leppert. 1980. Role of aggregated human growth hormone (hgh) in development of antibodies to hgh. *J. Clin. Endocrinol.* 51:691–697.

Mulkerrin, M., and R. Wetzel. 1989. pH dependence of the reversible and irreversible thermal denaturation of γ interferons. *Biochemistry.* 28:6556–6561.

Oosawa, F., M. Maruyama, and S. Fujime. 1972. Orientation distribution of globular protein molecules in a two-dimensional lattice: computer simulation. *J. Theor. Biol.* 36:203–221.

Orsini, G., and M. Goldberg. 1978. The renaturation of reduced chymotrypsinogen A in guanidine HCl. *J. Biol. Chem.* 253:3453–3458.

Pattou, D., B. Maigret, M. C. Fourniezaluski, and B. P. Roques. 1991. Computational analysis of conformational behavior of cholecystokinin fragments: I. Cck4, Cck5, Cck6 and Cck7 molecules. *Int. J. Pept. Protein Res.* 37:440–450.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1986. Numerical Recipes. Cambridge University Press. 195 pp.

Prevost, M., S. Wodak, B. Tidor, and M. Karplus. 1991. Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96 to Ala mutation in barnase. *Proc. Natl. Acad. Sci. USA.* 88: 10880–10884.

Prigogine, I. 1967. Introduction to Thermodynamics of Irreversible Processes. Interscience Publishers, Division of John Wiley & Sons, New York. pp. 14–17.

Przybycien, T. M., and J. E. Bailey. 1989. Aggregation kinetics in salt-induced protein precipitation. *AIChE J.* 35:1779–1790.

Sara, M., and U. B. Sleytr. 1987. Production and characteristics of ultrafiltration membranes with uniform pores from two-dimensional arrays of proteins. *J. Membr. Sci.* 33:27–49.

Sato, A., T. Furuno, C. Toyoshima, and H. Sasabe. 1993. Two-dimensional crystallization of catalase on a monolayer film of poly(1-benzyl-L-histidine) spread at the air/water interface. *Biochim. Biophys. Acta.* 1162:54–60.

Shih, Y., J. M. Prausnitz, and H. W. Blanch. 1992. Some characteristics of protein precipitation by salts. *Biotech. Bioeng.* 40:1155–1164.

Shortle, D., H. S. Chan, and K. A. Dill. 1992. Modeling the effects of mutations on the denatured states of proteins. *Protein Sci.* 1:201–215.

Stenberg, M., and H. Nygren. 1991. Computer simulation of surface induced aggregation of ferritin. *Biophys. Chem.* 41:131–141.

Voet, D., and J. G. Voet. 1990. Biochemistry. John Wiley & Sons, New York, NY. p. 30.

Wang, D. N., and W. Kuhlbrandt. 1991. High-resolution electron crystallography of light-harvesting chlorophyll *a/b*-protein complex in three different media. *J. Mol. Biol.* 217:691–699.

Weiner, S. J., P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. 1984. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* 106:765–784.

Wolfenden, R., L. Andersson, P. M. Cullis, and C. C. B. Southgate. 1981. Affinities of amino acid side chains for solvent water. *Biochemistry.* 20:849–855.

Yang, A., K. A. Sharp, and B. Honig. 1992. Analysis of the heat capacity dependence of protein folding. *J. Mol. Biol.* 227:889–900.

Yue, K., and K. A. Dill. 1992. Inverse protein folding problem: designing polymer sequences. *Proc. Natl. Acad. Sci. USA.* 89:4163–4167.

Zettlmeissl, G., R. Rudolph, and R. Jaenicke. 1979. Reconstitution of lactic dehydrogenase. Noncovalent aggregation vs. reactivation. 1. Physical properties and kinetics of aggregation. *Biochemistry.* 18:5567–5571.