Evolution of protein interfaces in multimers and fibrils

Citation for published version:
Zabel, WJ, Hagner, KP, Livesey, BJ, Marsh, JA, Setayeshgar, S, Lynch, M & Higgs, PG 2019, 'Evolution of protein interfaces in multimers and fibrils', The Journal of Chemical Physics, vol. 150, no. 22, pp. 225102. https://doi.org/10.1063/1.5086042

Digital Object Identifier (DOI):
10.1063/1.5086042

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
The Journal of Chemical Physics

Evolution of Protein Interfaces in Multimers and Fibrils

W. Jeffrey Zabel[1], Kyle P. Hagner[2], Benjamin J. Livesey[3], Joseph A. Marsh[3], Sima Setayeshgar[2], Michael Lynch[4], and Paul G. Higgs[1]*

1. Department of Physics and Astronomy, McMaster University, Hamilton, ON, L8S 4M1, Canada.
2. Department of Physics, Indiana University, Bloomington, Indiana 47405, USA,
3. MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine,
University of Edinburgh, Edinburgh EH4 2XU, UK and
4. Biodesign Center for Mechanisms of Evolution,
Arizona State University, Tempe, Arizona 85287, USA.

A majority of cellular proteins function as part of multimeric complexes of two or more subunits. Multimer formation requires interactions between protein surfaces that lead to closed structures, such as dimers and tetramers. If proteins interact in an open-ended way, uncontrolled growth of fibrils can occur, which is likely to be detrimental in most cases. We present a statistical physics model that allows aggregation of proteins as either closed dimers or open fibrils of all lengths. We use pairwise amino-acid contact energies to calculate the energies of interacting protein surfaces. The probabilities of all possible aggregate configurations can be calculated for any given sequence of surface amino acids. We link the statistical physics model to a population genetics model that describes the evolution of the surface residues. When proteins evolve neutrally, without selection for or against multimer formation, we find that a majority of proteins remain as monomers at moderate concentrations, but strong dimer-forming or fibril-forming sequences are also possible. If selection is applied in favour of dimers or in favour of fibrils, then it is easy to select either dimer-forming or fibril-forming sequences. It is also possible to select for oriented fibrils with protein subunits all aligned in the same direction. We measure the propensities of amino acids to occur at interfaces relative to non-interacting surfaces, and show that the propensities in our model are strongly correlated with those that have been measured in real protein structures. We also show that there are significant differences between amino acid frequencies at isologous and heterologous interfaces in our model, and we observe that similar effects occur in real protein structures.

* Corresponding author email: higgsp@mcmaster.ca
Submitted to Journal of Chemical Physics

PACS numbers:

I. INTRODUCTION

Many features of cellular biology are governed by the actions and interactions of proteins, and an understanding of their evolution is crucial to understanding the evolution of life itself. An important property of proteins is the formation of complexes consisting of two or more subunits, with 30-50% forming homo-oligomers composed of identical monomers [1]. Homo-dimers constitute the majority (41%) of oligomeric proteins of known structure [2]. Two identical proteins can aggregate in a closed way, with isologous (i.e. head-to-head) interfaces, or in an open way, with heterologous (i.e. head-to-tail) interfaces. If open, they have the possibility of forming infinite fibrils. Amyloid fibrils, formed by normally soluble proteins that assemble to form open insoluble fibers, are resistant to degradation and their formation can accompany a variety of human diseases, including Alzheimer's disease, type-2 diabetes and spongiform encephalopathies [3]. Given the importance of homo-oligomers in the cellular repertoire, from mediating gene expression, to functioning as enzymes, ion channels and receptors [4], it is important to understand the competition between these different ways of assembling. More generally, mutations of amino acids at protein-protein interfaces are known to have large effects on human health because they affect the formation of protein complexes [5].

Previous theoretical works have modeled protein fibrillogenesis based on mass action kinetics [6] and thermodynamics of peptide solutions including formation of protofilament intermediates [7, 8]. In this work, we present a simple model that allows both the physical and the evolutionary aspects of protein aggregation to be addressed. Our approach is similar to other previous works [9, 10] in adopting a transfer matrix approach to obtaining the equilibrium concentrations of oligomers of different lengths as a function of the free energies of interaction between proteins.

The novelty of our work is that it connects the statistical physics of protein aggregation to the evolution of higher-order protein structure by using population genetics theory to calculate the expected frequency of each protein in the ensemble of sequences generated by mutation and natural selection. We consider cases where the fitness is independent of whether the protein aggregates, and cases where fitness is a function of structure, including selection for the formation of dimers, and selection both for and against the formation of fibrils.

Our model considers a protein with two possible interacting surfaces, labelled A and B. There are two possible isologous interfaces (AA and BB), and one heterologous interface (AB). The energies of these interfaces depend on the amino acids on the two surfaces, as described in Section II. The model allows for the formation of closed dimers, which occur when one or the other of the isologous interfaces is strongly attractive and the other interfaces are weak. It also allows for the formation of fibrils with proteins oriented in the same direction in cases where the heterologous interface is strong, or fibrils with proteins aligned in alternating directions in cases where both isologous interfaces are strong. Using the transfer matrix method given in Section III, it is possible to calculate the probabilities P_n that a protein is found in an assembly of n units. These probabilities depend on the values of the three interface energies.

The multimeric states of proteins are sometimes observed to change rapidly on an evolutionary time scale [11, 12]. This may be an indication of selection for or against multimers, or may simply be a result of neutral evolution. Within our model, it is possible to ask how likely dimer and fibril structures are to form under neutral evolution. We include selection in the model using the strong-selection weak-mutation approximation [13], which allows the expected frequencies of sequences in the presence of selection to be calculated from their frequencies under neutral evolution. We use a Monte Carlo Markov Chain method (Section IV) to generate a set of representative protein sequences with frequencies given by evolutionary theory.

Thus, our model provides a simple way of linking evolutionary observations to the underlying statistical physics of protein aggregation. Within this framework, we consider probabilities of formation of dimers and fibrils, both under neutral evolution and under the action of several different kinds of selection. The model also predicts that the frequencies of amino acids at strongly-binding interfaces are significantly different from their frequencies under the mutation process alone, and from their frequencies at non-interacting, exposed surfaces. Furthermore, the frequencies of amino acids at isologous and heterologous interfaces are found to differ from one another. These predictions are compared with observations of amino acids frequencies at interfaces in databases of real proteins.

II. CALCULATION OF INTERFACE ENERGIES

We consider two opposing faces of the protein, denoted A and B, as potential binding surfaces (as shown in Fig. 1). There are two possible isologous interfaces (AA and BB), and one heterologous interface (AB). The energies of the three interfaces E_{AA}, E_{BB}, and E_{AB}, depend on the sequences of residues on the surfaces. Non-surface residues play no role in this model. A surface is modelled as a 4×4 array of amino acids. The energy of an interface is modelled as the sum of the 16 pairwise interactions between amino acids that are formed when two surfaces are brought together (see Fig. 1). We consider four possible 90^o rotations of two surfaces. The three energies E_{AA}, E_{BB}, and E_{AB} are defined to be the lowest of the four energies that arise from the four possible rotations.

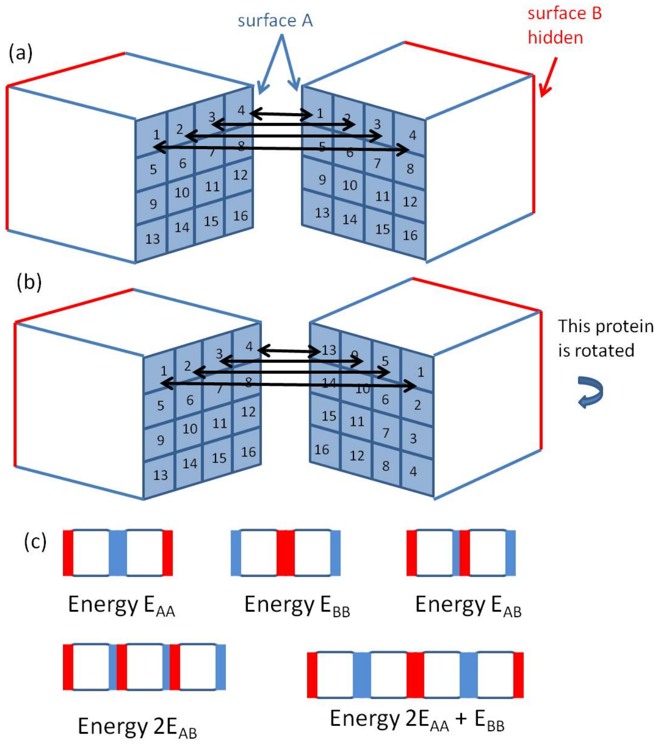


FIG. 1: Model of a protein with two opposing surfaces, A and B, that may interact, shown as blue and red, respectively. There are 16 amino acids on each surface. Interface energy is determined by the sum of the energies of the 16 pairwise contacts that are formed when the two surfaces are brought together, as indicated by arrows. (a) An AA interface is shown with the two proteins in the same rotational configuration. (b) An AA interface in which one protein has been rotated by 90^o. The energy E_{AA} of the AA interface is defined as the lowest energy of the four possible rotations. (c) When proteins aggregate in different configurations, the energy of the multimer is given by the sum of the energies of all the interfaces in the multimer structure.

The square array of 16 amino acids is used for convenience because we require a simple model for which energies can be calculated for hundreds of thousands of protein sequences during evolutionary simulations. However, the fact that we consider the lowest energy of the multiple rotations of the two surfaces is an important feature of the model. When two identical proteins form an isologous interface, each pairwise interaction between the two surfaces is present twice. This means that the variance of the energy of the interface is twice what it would be for an interface between two independent proteins with the same number of amino acid contacts. The relevant interaction energy controlling binding of two proteins is the lowest energy of the rotational configurations possible when they are brought into contact. As the distribution of energies is broader for homodimers than heterdimers, the lowest energy tends to be lower [14, 15]. This contributes to the excess of interactions between identical proteins and between closely related paralogs that is observed in the analysis of protein-protein interaction networks [16, 17]. This factor is relevant here, because we wish to consider relative probabilities of aggregation of multimer proteins in configurations that can involve either isologous or heterologous interfaces. If we simplified our model further by allowing only one rotational configuration, we would lose this effect.

As we wish to distinguish strongly and weakly interacting surfaces, it is useful to define the A surface such that the AA interface is stronger than the BB interface, i.e. $E_{AA} \leq E_{BB}$ (negative energies denote favourable interactions). For each amino acid sequence considered, we simply relabel the A and B surfaces if necessary, so that this condition is true.

We use a simple model of pairwise contact energies because we wish to study evolution of large numbers of protein sequences using a model where the fitness depends on the energies of the surface interactions (as described in Section III). Thus, it is necessary to be able to evaluate the surface energies of any given sequence very rapidly, which could not be done if a more realistic, three dimensional model of a protein surface was used (for example, as in [18–20]). Although pairwise amino acid potentials leave out many details (e.g. water and ion-mediated interactions, local flexibility of proteins, and the atomic structure of each residue), they have proved to be useful in many ways. The frequently-cited early work of Miyazawa and Jernigan [21] used the frequencies of contacts between amino acid pairs in globular protein structures to construct an effective pair potential matrix. This matrix has continued to be used in many applications such as coarse-grained simulations of protein complexes [22, 23], and is also used as a basis of a recent structural models of rates of amino acid substitutions [24–26].

Interactions between solvent and amino acids were not included originally [21], but Betancourt and Thirumalai [27] showed that this could be accounted for by shifting the elements relative to the amino acid threonine. The original matrix would not be suitable for our study here, because all the energies are negative, meaning all random surfaces would be attractive. This is not true in the transformed matrix, which has both positive and negative elements. The transformed matrix, B_{ij}, is shown in the Supplementary Table 1. It captures the fact that the interactions between pairs of hydrophobic amino acids are substantially negative, and those between hydrophobic and polar, or between two polar residues are, on-average, weaker, and can be either positive or negative. It also captures specific features such as attractions and repulsions between charged amino acids. As a concrete example, this matrix has been successfully used in a study of protein folding in the GroEL cavity [28].

It should also be noted that the B_{ij} matrix we use is derived from contact frequencies within globular protein structures, not from specific frequencies of amino acids at surfaces and interfaces. It is therefore essentially independent of data on interfaces propensities. We will show here that use of this energy matrix in our model leads to useful predictions on interface propensities that correlate with experimental observations. These predictions are non-circular, whereas they would be if we had used statistical potentials derived from surface data.

III. CALCULATION OF AGGREGATION PROBABILITIES

For any given sequence of surface residues, we calculate the interface energies as in Section II. We then use the interface energies to calculate the probabilities of protein-protein interactions. We consider a solution of a single kind of protein with total concentration ϕ moles per unit volume. We determine the equilibrium concentration of monomers c, and of aggregates of n subunits, C_n, in the following way.

For each of the three types of interface $ij \in \{AA, AB, BB\}$, we define

$$a_{ij} = \frac{1}{\omega} e^{-\beta E_{ij}}, \tag{1}$$

where ω is the number of possible orientational configurations of one protein relative to its neighbour. For the simple cubic lattice considered here, $\omega = 24$, which is the number of possible orientations of a cubic object on a cubic lattice. In the calculations below, the statistical weight of an interface of type ij is given by $a_{ij} c/c_0$, where $c_0 = 1\,\mathrm{M}$ is the reference concentration.

The concentrations of the different possible aggregates can be calculated by considering formation of chains that grow from one end only. If chains grow from both ends, or if chains can aggregate with other chains (rather than just chains with monomers), this does not alter the equilibrium frequencies of the different aggregates. Therefore we give the simplest case of the calculation here, which allows growth one monomer at a time from one end only.

Letting $C_n(A)$ and $C_n(B)$ denote the equilibrium concentrations of a chain of length n, with the A or the B face exposed at the growing end, the equilibrium concentrations can be calculated using a transfer matrix method:

$$\begin{pmatrix} C_n\,(A) \\ C_n\,(B) \end{pmatrix} = (c/c_0) \begin{pmatrix} a_{AB} & a_{BB} \\ a_{AA} & a_{AB} \end{pmatrix} \begin{pmatrix} C_{n-1}\,(A) \\ C_{n-1}\,(B) \end{pmatrix}. \tag{2}$$

We define $C_1(A) = C_1(B) = c/2$, so that the sum of the two orientations is equal to the total free monomer concentration, c. The eigenvalues of the transfer matrix, $\boldsymbol{a}$, are given by: $\lambda_\pm = a_{AB} \pm \sqrt{a_{AA} a_{BB}}$, and from these, the abundance of chains of length n, given by $C_n = C_n(A) + C_n(B)$, can be obtained as:

$$\frac{C_n}{c_0} = \left(\frac{c}{c_0}\right)^n \left[A_+ \lambda_+^{n-1} - A_- \lambda_-^{n-1}\right], \tag{3}$$

where

$$A_{\pm} = \frac{a_{AA} + a_{BB} \pm 2\sqrt{a_{AA}a_{BB}}}{4\sqrt{a_{AA}a_{BB}}}. \quad (4)$$

For example, it is easily verified from Eq. 3 that for $n = 2$, the dimer concentration is

$$\frac{C_2}{c_0} = \frac{1}{2}\left(\frac{c}{c_0}\right)^2 (a_{AA} + 2a_{AB} + a_{BB}) \,, \quad (5)$$

where the terms for the two dimer configurations with isologous interfaces and the two orientations of the dimer with the heterologous interface can be clearly seen.

The concentration of monomers, c, can now be determined. The concentration of proteins in clusters of size n is given by $\phi_n = nC_n$. The total protein subunit concentration is $\phi = \sum_n \phi_n$. This sum gives an equation from which the free monomer concentration, c, can be calculated:

$$\frac{\phi}{c_0} = \frac{A_+(c/c_0)}{(1 - \lambda_+ \, c/c_0)^2} - \frac{A_-(c/c_0)}{(1 - \lambda_- \, c/c_0)^2}. \quad (6)$$

There is always a single solution to Eq. 6 in the physical range where $0 < c < \phi$.

The probability of a subunit being present in an n-mer is $P_n = \phi_n/\phi$. The fractions of proteins present as monomers and dimers are P_1 and P_2. We refer to all aggregates of 3 or more units as fibrils, hence the fraction of proteins in fibrils is $P_{fib} = \sum_{n \geq 3} P_n$. In some cases we wish to distinguish closed dimers with the strong AA interface from other dimers. The fraction of proteins in closed dimers is

$$P_2^* = P_2 \frac{a_{AA}}{a_{AA} + a_{BB} + 2a_{AB}}. \quad (7)$$

Likewise, in other cases, we wish to distinguish oriented fibrils containing only AB interfaces from general fibrils containing mixtures of all three types of interface. The concentration of proteins in oriented fibrils of length n is

$$\phi_n^{ori} = \frac{nc^n a_{AB}^{(n-1)}}{c_0{}^{(n-1)}}, \quad (8)$$

and the fraction of proteins in oriented fibrils is

$$P_{ori} = \frac{1}{\phi} \sum_{n \geq 3} \phi_n^{ori}. \quad (9)$$

IV. EVOLUTIONARY COMPUTATIONS

We now consider the evolution of proteins whose interactions are described by the statistical physics model above. We consider a population of individuals, each with a gene for the protein in question. The fitness of an individual is a function of the protein sequence. If mutation is weak in comparison to selection, as we will assume below, there is a dominant variant of the protein in the population at any one time, and occasionally a new variant spreads through the population and replaces the old one. We would like to calculate the long-term steady state frequencies of sequences in the ensemble of sequences generated by this evolutionary process.

We consider protein sequences evolving under a mutational model in which the rate of mutation from amino acid i to j is $r_{ij} = u\pi_j$, where u is a rate constant and π_j is the steady state frequency of amino acid j under the mutational process. For simplicity, we deal with mutations at the level of the protein sequence, and do not consider the underlying DNA. In the neutral case, protein sequences evolve under the influence of mutation, and there is no selection. Let f_k^{mut} be the steady state frequency of sequence k under mutation. We consider the simplest case where all 20 amino acids have equal frequency ($\pi_j = 0.05$ for all j). Hence there are 20^{32} possible amino acid sequences, each with steady state frequency $f_k^{mut} = (0.05)^{32}$.

We define the fitness of a sequence as $w = 1 + s$, where positive and negative values of the selection coefficient, s, denote advantageous or deleterious sequences, and $s = 0$ for neutral variants. For any amino acid sequence, we assume that s is a function of the multimer configuration probabilities P_n for that sequence. We consider several choices of fitness functions: (i) a neutral case, where $s = 0$ for every sequence; (ii) positive selection in favour of dimer formation, where $s = \sigma P_2^*$; (iii) selection against fibril formation, where $s = -\sigma P_{fib}$; (iv) selection in favour of fibril formation, where $s = \sigma P_{fib}$; and selection in favour of oriented fibrils containing only AB interfaces, where $s = \sigma P_{ori}$. In all these cases, σ is a positive constant that determines the strength of selection.

In order to calculate the steady state frequencies of sequences in the presence of selection as well as mutation, we assume that mutations are rare enough so that only one mutation is segregating at a time in the same gene. This is a common approximation in population genetics that allows analytical progress in a simple way. In this approximation, the stationary frequency of a sequence k under the influence of selection is weighted by a factor $e^{2N_e s(k)}$ relative to the case with no selection [29, 30], where $s(k)$ is the selection coefficient for this sequence and N_e is the effective population size. The frequency of sequence k under selection and mutation is

$$f_k^{sel} = \frac{f_k^{mut} e^{2N_e s(k)}}{\sum_j f_j^{mut} e^{2N_e s(j)}}. \quad (10)$$

The practical issue with Eq. 10 is the exponential number of sequences in the sum. It is not possible to exhaustively consider all 20^{32} sequences. We therefore use a Markov Chain Monte Carlo (MCMC) sampling method that generates a large sample of representative protein sequences, such that the probability of any sequence arising in the sample is proportional to its steady state frequency.

The average properties of the full ensemble are closely approximated by the simple mean of the properties of the sequences in the sample.

The MCMC simulations begin with a random sequence of 32 amino acids. We then generate a descendant sequence via replication with mutation. The probability that an amino acid i in the parent is replaced by j in the descendant is r_{ij}. The probability that the amino acid remains unchanged is $r_{ii} = 1 - \sum_j r_{ij}$. The value of u is not critical, as it does not influence steady state frequencies. We found $u = 0.05$ to allow efficient exploration of the sequence space. If there is no selection, then every descendant sequence is accepted into the sample, and the method generates a sample with frequencies proportional to f_k^{mut}. If selection is acting, we accept or reject the descendant according to its fitness. Let the current sequence be k_1 and the descendant be k_2, and let the selection coefficients for these sequences be $s(k_1)$ and $s(k_2)$. The difference in fitness between the sequences is $\Delta s = s(k_2) - s(k_1)$. To insure that the frequency of any sequence k in the sample is proportional to $f_k^{mut} e^{2N_e s(k)}$, as is required, the ratio of acceptance of mutations that increase and decrease fitness must be $e^{2N_e \Delta s}$. Our MCMC algorithm does this in the simplest way: it accepts the new sequence with probability 1, if Δs is positive, and with probability $e^{2N_e \Delta s}$, if Δs is negative. If the new sequence is rejected, a second copy of the old sequence goes into the sample. This method is equivalent to the Metropolis algorithm used for Boltzmann-weighted sampling in physics. We also note that a similar method of evolutionary simulation was used in another model of protein evolution [31] in which the fitness of a sequence depends on its folding ability and its affinity to another target model.

V. PHENOTYPE DISTRIBUTIONS

The two most useful quantities to summarize the phenotype of a sequence are the frequency of AA dimers, P_2^*, and the frequency of fibrils, P_{fib}. Fig 2(a) shows the distribution of a sample of sequences generated by the MCMC evolutionary simulation in the neutral case with a total concentration $\phi = 0.01M$. The MCMC routine ran for 300000 generations, and the first 5000 were discarded to allow for equilibration. As all sequences have equal frequency under this mutational model when there is no selection, the sequences generated are simply random amino acid sequences. The figure shows that sequences are spread over a broad range of P_2^* and P_{fib}. Sequences close to the origin (where P_2^* and P_{fib} are close to zero) exist mostly as monomers (P_1 is close to 1). Sequences in the bottom right corner are mostly dimers. Sequences in the top corner are mostly fibrils. It can be seen, however, that strong fibril formers are rare under neutral evolution at this concentration. Thus, no points are found very close to the top corner in Fig 2a. The mean values of these probabilities for all sequences in the sample are $\langle P_2^* \rangle = 0.04$ and $\langle P_{fib} \rangle = 0.003$. Thus, typical sequences are usually monomers.

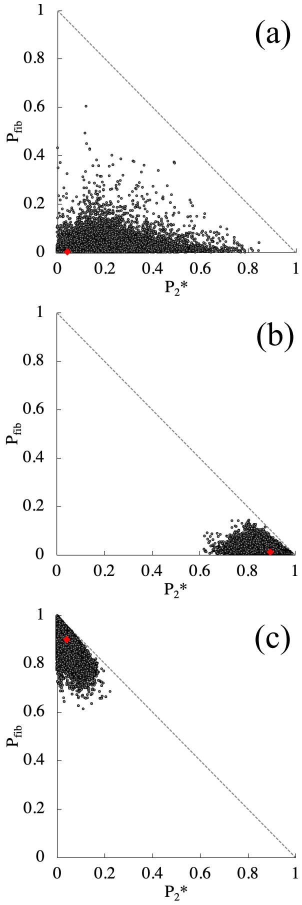


FIG. 2: Phenotype distribution in the space P_2^* versus P_{fib} for samples of sequences arising under evolution using the MCMC method. (a) neutral, (b) selection for dimers ($N_e\sigma = 25$), (c) selection for fibrils ($N_e\sigma = 25$). For each of these plots, the red symbol denotes the mean value of P_2^* and P_{fib} for all the sequences in the sample.

Figs 2(b) and 2(c) show the way the phenotype distribution shifts when selection is applied for dimers and for fibrils. When selection is applied for dimers, the distribution shifts close to the bottom right corner, with $\langle P_2^* \rangle = 0.89$ and $\langle P_{fib} \rangle = 0.01$. When selection is applied for fibrils, the distribution shifts close to the top corner, with $\langle P_2^* \rangle = 0.04$ and $\langle P_{fib} \rangle = 0.90$. This means that sequences that are either very strong fibril-formers or very strong dimer-formers are possible in this model, and that they arise easily when selection favours them. Nevertheless, they are relatively rare compared to the large number of random sequences with weaker interface interactions, so they do not arise frequently in the mixture of random sequences generated under neutral evolution.

Sequence	Description	E_{AA}/kT	E_{BB}/kT	E_{AB}/kT
A	Monomer	-2.68	-2.32	-2.80
B	Dimer-former	-12.06	-3.14	-5.25
C	Strong dimer-former	-15.82	-0.60	-2.80
D	Fibril former	-9.08	-8.67	-5.68
E	Strong fibril-former	-12.28	-12.06	-12.01
F	Oriented fibril-former	-4.53	-4.52	-8.18

TABLE I: Energies of the three interfaces for example sequences A-F discussed in Figures 3 and 4.

To illustrate the range of behaviors shown by individual sequences, we chose the six example sequences A-F described in Table I. For each of these sequences, the distribution of n-mer probabilities, P_n, is shown in Fig. 3 at concentration $\phi = 0.01M$. This value is consistent with cellular concentrations of the enzymes that are present at the highest quantities in cells, as these are the ones for which aggregation is most relevant. Various mechanisms of subcellular protein localization would additionally enhance their concentrations [32–34].

The probabilities P_n change significantly as the concentration is varied. The changes with concentration can be illustrated as trajectories in the P_2^* versus P_{fib} triangle. Fig 4 shows the trajectories for sequences A-F as the concentration is increased from 10^{-6} M to 1 M. All sequences begin at the origin (all monomers) for low concentration and eventually move towards the fibril corner for very high concentration. Dimer-forming sequences approach the dimer corner at intermediate concentrations. The concentration $\phi = 0.01$ M, which was used in Fig. 3 is shown as red diamonds in Fig 4. Extreme concentrations higher than this are included in order to illustrate the predictions of the model. The highest concentration point is 1 M, shown as a purple triangles.

Sequence A is a typical sequence chosen randomly from the sample generated by the neutral simulation (Fig. 2a). The energies of all three interfaces are weak; hence, this sequence is almost entirely monomers at $\phi = 0.01$ M (see Fig. 3A). The trajectory does not move close to the dimer corner at any concentration, and it is still not close to the fibril corner, even at $\phi = 1$ M.

Sequence B is a dimer-former found in the neutral sample. It is the sequence with the highest P_2^* in Fig. 2a. This sequence is mostly a dimer at $\phi = 0.01$ M (see Fig. 3B), and gradually becomes a fibril at concentrations higher than this. Sequence B forms dimers because the AA interface is strong. E_{AA} is much lower than the other two energies (see Table 1).

Sequence C is a strong dimer-former found in the sample generated under selection for dimer formation (Fig. 2b). It is almost entirely a dimer at the reference concentration, and remains very close to the dimer corner even at $\phi = 1$ M. The AA interface is even stronger than for Sequence B.

Sequence D is a fibril-former found in the neutral sample. All three interface energies are fairly strong. This sequence has $P_{fib} = 0.61$ at $\phi = 0.01$ M, which is the highest in Fig. 2a, and the distribution of P_n has significant weight at larger n. Sequence E is a strong fibril-former found in the sample of sequences selected for fibril formation (Fig. 2c). All three interface energies are very strong. This sequence has P_{fib} close to 1 already at $\phi = 0.01$ M.

Sequence F is a fairly strong fibril-former found in the neutral sample, which has $P_{fib} = 0.44$ at $\phi = 0.01$ M. It differs from the other fibril-formers (D and E) in that the heterologous interface energy E_{AB} is much lower than the others. This means that it forms mostly oriented fibrils. The frequency of closed AA dimers, P_2^* is very low at all concentrations, hence the trajectory in Fig 4, moves almost along the P_{fib} axis.

VI. PROPERTIES OF PROTEIN INTERFACES

Fig 5 shows the mean energies of the three possible interfaces for random sequences evolving neutrally (shown as horizontal lines) and compares these with the mean energies for sequences generated under four differend kinds of selection (shown as points). It can be seen that for neutral evolution, E_{AA} is significantly lower than E_{BB} even though the sequences are random. This occurs by definition, because we have labelled the surfaces A and B for each sequence such that A forms the stronger interface of the two. The heterologous interface energy E_{AB} is intermediate between the two isologous interface energies. All three energies are negative because the mean interaction energy of random amino acid pairs (from the B_{ij} matrix in Supplementary Table 1) is slightly negative: $\langle B_{ij} \rangle = -0.057$. The mean energy for an interface of 16 random pairs is therefore -0.912. The average energies of the three kinds of interface under neutral evolution are all lower than this because we consider four rotations of the two surfaces, as shown in Fig 1, and take the lowest of these to define the energy of the interface.

Fig 5 illustrates the way the energies of the interfaces change when selection is applied. In the case of selection for dimers, E_{AA} decreases substantially with respect to the neutral case, as we would expect, because we are selecting for sequences with high P_2^*. It can be seen that E_{BB} actually increases slightly with respect to the neutral case. It is important that the BB interface should remain weak, because if both AA and BB interfaces become strong, the sequence will form fibrils with proteins in alternating directions.

The second column in Fig 5 illustrates the case of selection *against* fibrils. We were interested in this case because we expect that uncontrolled fibril formation should be harmful to the cell. Selection against fibrils eliminates the rare sequences with high P_{fib} from the neutral phenotype distribution, but since these sequences are rare, and since the mean value of P_{fib} in the neutral case is already very low, selection against fibrils has only a small

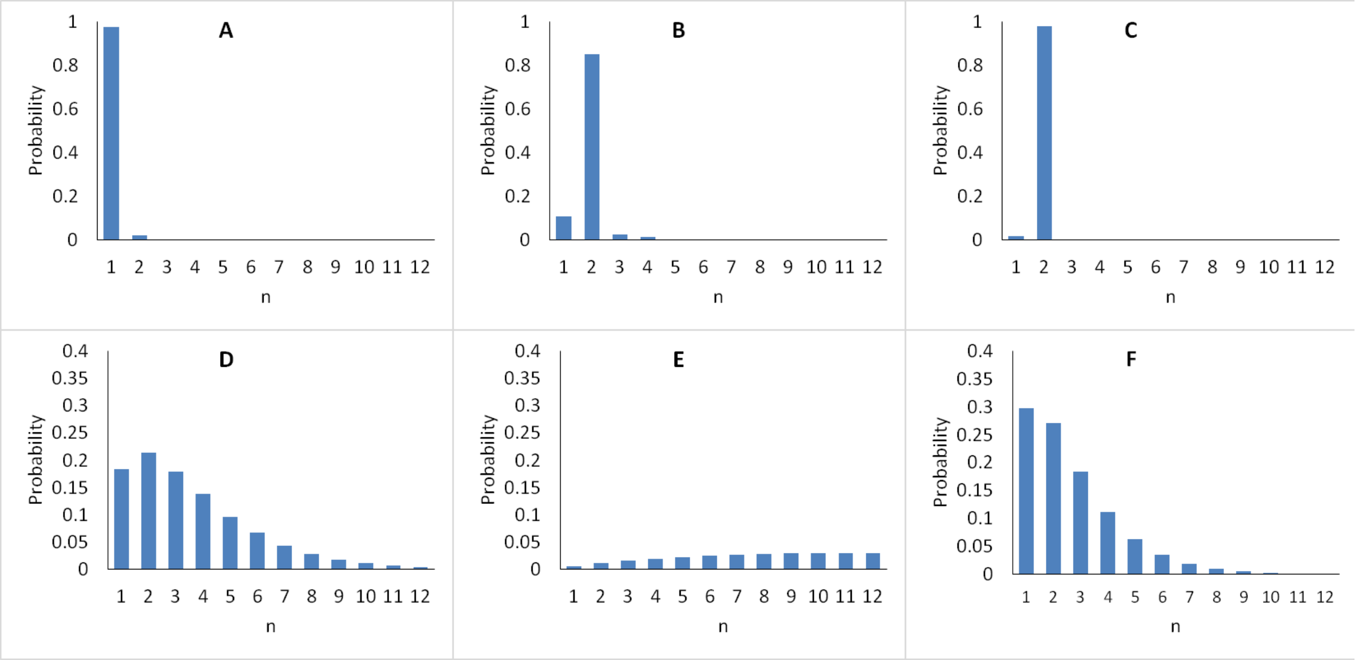


FIG. 3: Histograms of P_n for six example sequences illustrating different behaviors at $\phi = 0.01$ M. Panels A-F refer to the six sequences described in Table I.

effect on the mean energies of the interfaces. It can be seen that all three energies increase slightly with respect to their neutral values, making all kinds of multimers and fibrils less frequent.

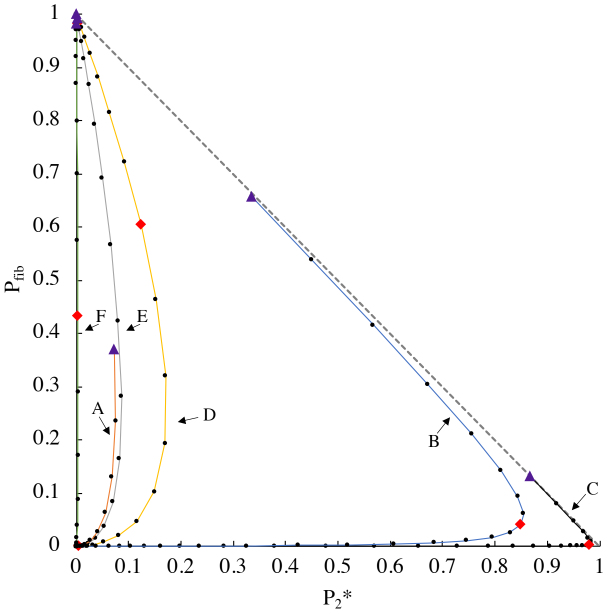


FIG. 4: Plot showing trajectories in the space P_2^* v. P_{fib} for the six sequences A-F in Table I. The concentration varies from $\phi = 10^{-6}$ M to 1 M, with the reference concentration $\phi = 0.01M$ labelled as a red diamond, and the highest concentration $\phi = 1M$ labelled as a purple triangle.

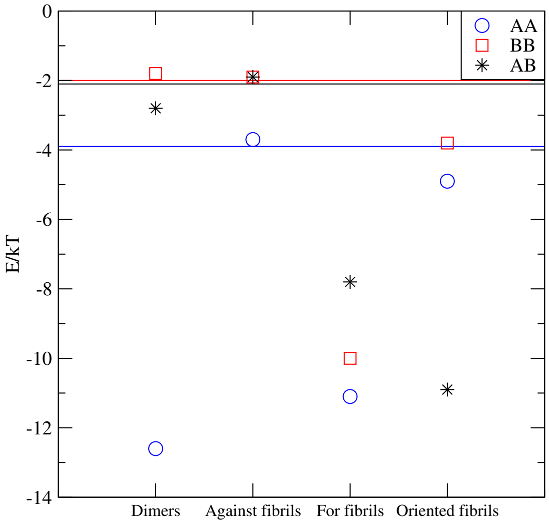


FIG. 5: Comparison of the mean energies of the three possible interfaces for random sequences evolving neutrally (shown as horizontal lines) with sequences generated under four differend kinds of selection (shown as points). Blue lines and circles E_{AA}; Red lines and squares E_{BB}; Black lines and stars E_{AB}.

There are some proteins whose function requires fibril formation, such as actin and tubulin. Therefore we also considered the case when selection acts *for* fibrils. In this case all three interface energies become much lower than the neutral values. It can be seen that E_{AA} and

E_{BB} decrease more than E_{AB}, meaning that the orientation of proteins in these fibrils will be rather random, but there will be relatively few heterologous interfaces. In section III, we defined P_{ori}, the fraction of proteins in oriented fibrils. For the case where selection is for fibrils of all kinds, we find $\langle P_{ori} \rangle$ is only 0.05, even though $\langle P_{fib} \rangle$, which includes fibrils with proteins in all possible arrangements, is 0.90. In contrast, the fourth case in Fig 5 shows selection for *oriented* fibrils only. In this case E_{AB} decreases much more than E_{AA} and E_{BB}. Hence, most interfaces will be heterologous. In this case we find $\langle P_{fib} \rangle$ is also 0.90, but $\langle P_{ori} \rangle$ is 0.87, meaning that almost all the fibrils are oriented.

Taken together, these results show that this model of protein interfaces is quite versatile. It allows selection for both increased or decreased strength of interfaces, and it allows separate selection for either heterologous interfaces (as in the case of dimers) or isologous interfaces (as in the case of oriented fibrils).

VII. INTERFACE PROPENSITIES OF AMINO ACIDS

Our model allows us to study the way that the frequencies of amino acids at interfaces vary with respect to the frequencies that would be expected under random mutation. The propensities of the amino acids to occur at interfaces have been measured in real proteins (Jones and Thornton [35]; Levy et al. [36]). In this section, we show that our 20-amino acid model generates interface propensities that are similar to these.

We generated 2×10^7 random sequences of amino acids on the A and B surfaces. We calculated E_{AA} and E_{BB}, if necessary relabelling A and B so that A is the stronger interface. We measured the frequencies p_A and p_B of amino acids on each surface, relative to the expected frequency under neutral mutations, $\pi_i = 0.05$. These are shown in Fig 6, and the data is given in Supplementary Table 2. When we compare pairs of interfaces in this way, and distinguish the stronger from the weaker, the frequencies of amino acids on the two surfaces are different, even though the average frequency on both surfaces has to be equal for all amino acids. Hence, in the figure, we see that p_A can be significantly higher or lower than p_B for many of the amino acids, even though the average of p_A and p_B has to be 1 for every amino acid.

From these probabilities, we define an interface propensity for each amino acid as

$$S_{int} = \ln(p_A/p_B). \tag{11}$$

This score is positive for amino acids that have increased frequency at strong interfaces, and negative for those that have decreased frequency.

The S_{int} scores are related to the energies in the B_{ij} matrix, as shown in Fig. 7(a). We define B_{self} as the self interaction energy of the amino acid (the diagonal element B_{ii} of the matrix), and B_{ave} as the mean of the interaction of one amino acid with the 20 possible partners. The more negative B_{self} and B_{ave}, the higher the interface propensity. The data are given in Supplementary Table 2. The correlation coefficients r and the p values for the t-test of correlation are given in the caption and in Supplementary Table 3. These correlations are highly significant, and the correlation with B_{ave} is stronger than with B_{self}, as can be seen graphically in in Fig. 7(a).

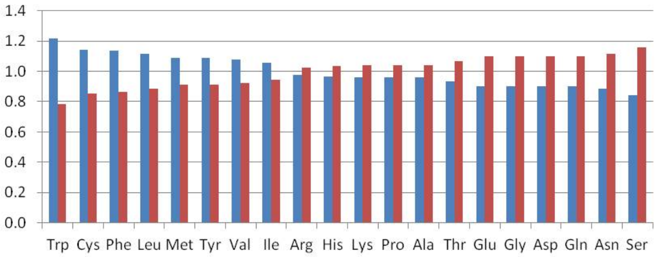


FIG. 6: Relative amino acid frequencies at the A surface (p_A, blue) and the B surface (p_B, red) for neutrally evolving proteins. The frequencies are not equal on the two surfaces because the A and B surfaces have been defined as the stronger and weaker of the two, respectively.

The S_{int} scores are also related to two previous scales of interface propensities measured from protein structure data. In Supplementary Table 2, ln(RIP) is the “relative interface propensity” from [35], and “stickiness” is the interface propensity scale from [36]. Both of these scales are derived from the observed frequencies of amino acids at protein-protein interfaces relative to their frequencies at non-interacting surfaces. The score from our model, obtained from the relative frequencies at the A and B surfaces, is directly comparable to these. Fig. 7(b) shows that there is a strong positive correlation between S_{int} and the two other interface propensities. The correlation coefficients and t-test parameters are given in Supplementary Table 3, showing that there is highly significant correlation between all three scores.

This observation tells us that the B_{ij} matrix contains detailed information about the strengths of interactions between the different amino acids that is sufficient to quantitatively predict which amino acids increase or decrease in frequency at interfaces. It also tells us something about why this occurs. Since surface amino acids interact with other copies of themselves and with all possible other amino acids at the interface, those which have the most negative B_{self} and B_{ave} increase the most in frequency at the strongly-binding interface.

Fig. 8 and Supplementary Table 4 show the frequencies of amino acids at surfaces A and B in the sets of sequences generated by the MCMC sampling method when selection is present. Selection for dimers (as in Fig. 8a) accentuates the difference between the A and B surfaces that is already seen in the neutral case (Fig. 6). In the dimer case, the hydrophobic amino acids on the left are

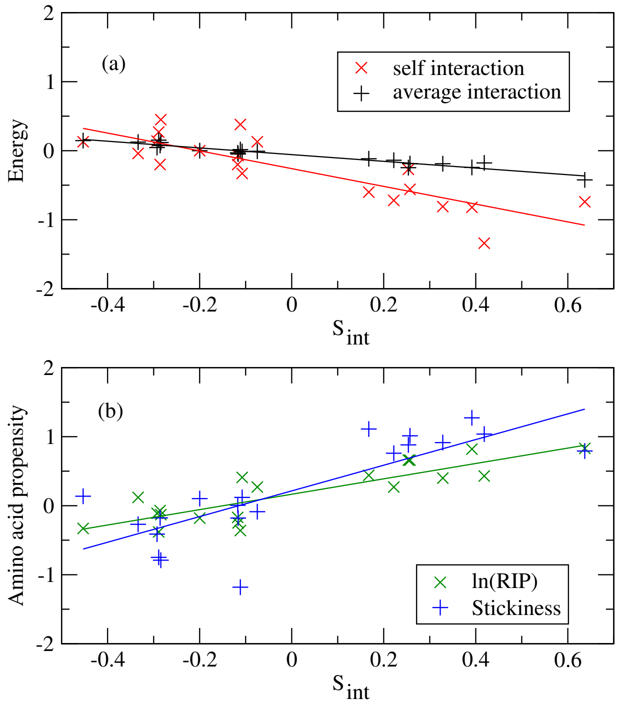


FIG. 7: (a) Correlation of S_{int} with the self energy B_{self} and the average energy B_{ave} of the B_{ij} matrix. (b) Correlation of S_{int} with propensities measured in protein data sets. The correlation coefficients and significance values for these plots are: $B_{self}, r = -0.836, p = 4.4 \times 10^{-6}$; $B_{ave}, r = -0.966, p = 5.2 \times 10^{-12}$; ln(RIP), $r = 0.850, p = 2.1 \times 10^{-6}$; stickiness, $r = 0.797, p = 2.6 \times 10^{-5}$

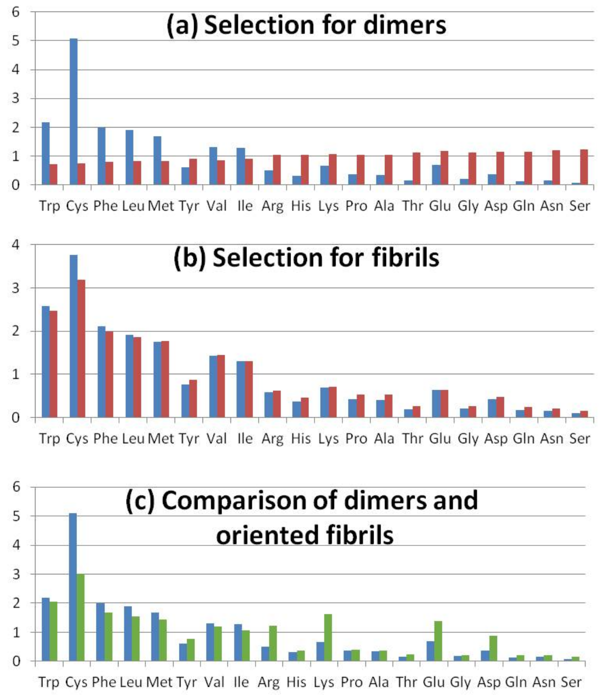


FIG. 8: (a) Relative amino acid frequencies at the A surface (p_A, blue) and the B surface (p_B, red) in the case of selection for dimers. (b) Same in the case of selection for fibrils. (c) Relative amino acid frequencies at the A surface of isologous interfaces in dimers(blue) and the A surface of heterologous interfaces in oriented fibrils (green).

very much more frequent on the dimer-forming A interface than on the non-interacting B interface, whereas the hydrophilic amino acids on the right are much more frequent on the non-interacting B interface.

In the case of selection for fibrils, both AA and BB interfaces become strong. Thus p_A and p_B both show a decreasing trend from left to right in Fig. 8b, and there is not much difference between p_A and p_B. The case of selection *against* fibrils is shown in Supplementary Table 4. The frequencies do not change very much from the neutral case, because most sequences have a low fibril forming probability, as we saw previously.

Fig. 8c compares the p_A frequencies in the case of selection for dimers, with the p_A frequencies in the case of selection for *oriented* fibrils. In the dimers case, we select for isologous AA interfaces, whereas in the case of oriented fibrils, we select for heterologous AB interfaces. Comparison of these two shows that amino acids differ significantly in frequency between isologous and heterologous interfaces. In particular, it can be seen that Cys is more frequent at isologous interfaces and the charged amino acids (Arg, Lys, Glu and Asp) are more frequent at heterologous interfaces. From this, we define a propensity for amino acids at isologous versus heterologous interfaces as

$$S_{iso} = \ln\left(p_A(\text{dimers})/p_A(\text{oriented fibrils})\right). \qquad (12)$$

This propensity is highest for Cys, and most negative for the charged amino acids (see Supplementary Table 4).

This effect occurs because amino acids in isologous interfaces have a significant probability of interacting with the copy of themselves on the other side of the interface, whereas amino acids in heterologous interfaces only interact with themselves if there is an independently-evolved amino acid of the same kind on the other surface. We define the difference between average and self energies as $\Delta B = B_{ave} - B_{self}$. We expect amino acids with positive ΔB to be favoured at isologous interfaces, and vice versa. Fig. 9(a) shows that there is a highly-significant correlation of S_{iso} and ΔB. The significance values are given in the caption and in Supplementary Table 3. From the B_{ij} matrix in Supplementary Table 1, it can be seen that Cys has a particularly low value of B_{self}, presumably reflecting the presence of disulphide bridges in the data from which the B_{ij} matrix was derived. This results in a large positve ΔB for Cys. The charged amino acids have positive values of B_{self}, presumably due to repulsions between like charges. This results in negative ΔB for the charged amino acids.

It can be seen in Fig. 1b that when there is a 90^o rotation of one protein with respect to the other, four of the

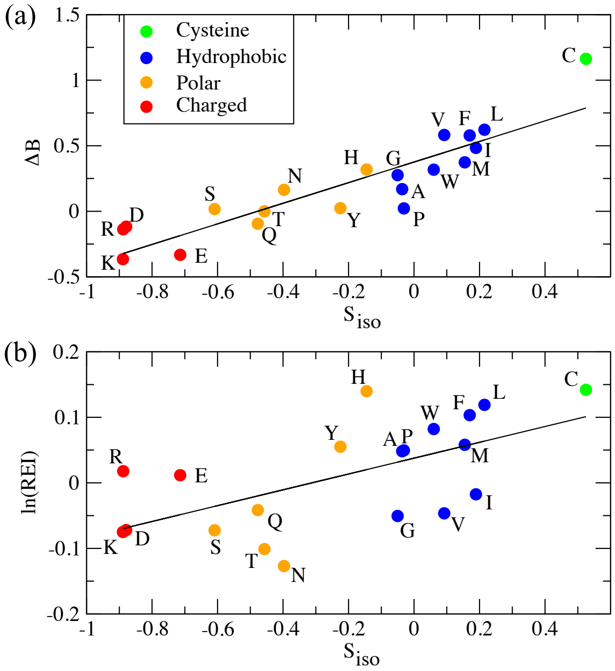


FIG. 9: (a) Correlation of the difference between average and self energies $\Delta B = B_{ave} - B_{self}$ with the isologous versus heterologous interface propensity, S_{iso}; (b) Correlation of the relative enrichment of amino acids in isologous versus heterologous interfaces, ln(REI), found in analysis of real protein complex structures with S_{iso} obtained from our model. The correlation coefficients and significance values for these plots are: $\Delta B, r = 0.897, p = 8.7 \times 10^{-8}$; ln(REI), $r = 0.615, p = 3.9 \times 10^{-3}$;

sixteen amino acids (numbered 1, 6, 11, 16) form pairwise contacts with themselves. The same happens in the 270^o rotation, but does not happen in the 0^o and 180^o rotations. Although some of these details are particular to the square lattice we are using, the point that amino acids in isologous interfaces can interact with copies of themselves is still true in real proteins where there is no square lattice. For example, the same effect occurs in the circular patch model used in [14, 15]. Therefore, it is reasonable to ask whether the systematic difference in amino acid frequencies between isologous and heterologous interfaces that we observe in our model also arises in real proteins. To address this, we performed a systematic analysis of the amino acid residues present in the homomeric interfaces of real protein complex structures present in the Protein Data Bank.

Starting from a snapshot of all of the structures in the Protein Data Bank (9-26-2018), all protein residues present in homomeric interfaces were identified as those burying any solvent-accessible surface area with an identical polypeptide chain. Incomplete residues missing any non-hydrogen atoms from the side chain were excluded. Isologous interfaces were classified as those where the correlation between the residue-specific buried surface area for each subunit in an interacting pair was > 0.7, as defined previously [42]. Since the dataset of interface residues initially contained data from many proteins with closely related or identical sequences, we used the PISCES protein sequence culling server [43] to remove chains with 90% or greater similarity. This resulted in a total number of 932,536 interface residues from 18,777 non-redundant chains. The total number of occurrences of each amino acid was then counted for isologous and heterologous interfaces, and the proportion was calculated by dividing by the total number of residues at each type of interface. Relative enrichment at isologous interfaces, REI, was calculated by dividing the proportion of each amino acid at isologous interfaces by the proportion at heterologous interfaces (see Supplementary Table 4).

Interestingly, we observe a signficant correlation between S_{iso} and ln(REI) (see Supplementary Table 3 and Fig. 9(b)), thus validitating the utility of our simplified model and demonstrating its power in capturing genuine sequence differences between the different types of interfaces. The deviations between our model and the pattern observed in real structures could be due to a number of factors. In particular, there are likely to be systematic differences in the functions of homo-oligomers with isologous versus heterologous interfaces, as there is a strong association between symmetry and function [4]. For example, transmembrane channels will be enriched in heterologous interfaces due to their strong association with higher-order cyclic symmetry. Thus if the interfaces of transmembrane proteins tend to differ in amino acid composition compared to other proteins, this could add a degree of bias.

VIII. DISCUSSION AND CONCLUSIONS

This work presents a first attempt at a theoretical decription of the evolution of multimers and fibrils. The pairwise contact-energy matrix that we used is a simple way of defining interface energies that does not account for three dimensional structure of surfaces. It was not optimized in any way for the present model. We are therefore very satisfied that several features of the interface propensities and the isologous/heterologous propensities are quite close to those seen in real proteins.

The inevitable presence of hydrophobic residues means that all proteins will be aggregation prone to some extent. Hydrophobic residues in the interior are necessary for proper folding of proteins, and hydrophobic residues on the surface can lead to formation of functional multimeric states. While uncontrolled protein aggregation has been shown to be associated with an increasing number of pathological conditions, including human diseases, due to loss of normal function or gain in toxic activity, fibril formation can also serve functional roles in cases such as adhesion and biofilm formation in bacteria [37] and defense against micro-organisms [38]. Cells employ a range of strategies to control aggregation, at both the sequence level (for example, through modulation of aggregation-prone regions or protein stability), and at the cellular level (for example, through compartmentalization and modulation of protein abundance) [39].

With the interaction energies used here, we find that strongly aggregating proteins will be rare under neutral evolution at concentrations that are likely to arise in the cell. This conclusion needs to be treated with caution because of the simplicity of the interaction energy rules. We have only considered solutions of a single kind of protein. It may be possible to extend the model to consider mixtures of many kinds of proteins in the future. It should also be noted that there is a parameter ω for rotational entropy in equation 1 that is not known with certainty. Lower values of ω would lead to higher probabilities of aggregation at any given concentration. Also, we have arbitrarily chosen surfaces with 16 amino acids. Increasing or decreasing the number of interacting amino acids in a patch would increase or decrease the strength of interactions, which would also affect the frequency of strongly-aggregating proteins expected under neutral evolution.

A further caveat is that the evolutionary calculations were done under the approximation that a single mutation is segregating in the sequence at once, which is not always true. This could be improved using full-scale population genetics simulations in the future. It would also be possible to consider evolution at the DNA level and determine the portein sequence by translation of the gene. This would allow us to consider cases where the steady state frequencies of the amino acids in the proteins and the four nucleotides in the genes are biased by mutation.

Future extensions of this work include developing this model to consider other multimer structures, such as cyclic and dihedral tetramers, by allowing more than two sticky faces on each protein, or by considering proteins with two sticky faces at an angle of 90^o to one another. An important aim will be to predict the relative frequencies of multimers of different symmetries and different numbers of subunits, as is tabulated in the "periodic table" classification of the protein structure database [40]. Furthermore, as our model is able to predict the way the multimer structures will change when mutations are made to the surface residues, we will be able to study evolution of multimer structures over time in a family of related species, and compare this with studies of structural evolution [41]. The present approach therefore introduces a method from which a wide range of new developments will be possible for the study of the evolution of higher order protein structure.

IX. SUPPLEMENTARY MATERIAL

The file Supplementary Tables.xlsx contains the data in the four supplementary tables. The file Supplementary Table Descriptions.docx contains descriptions of these tables.

X. ACKNOWLEDGMENTS

Author contributions: PGH and ML developed the theory linking the statistical physics of protein aggregation to the evolutionary theory. KPH and SS carried out simulations of this model using sequences of HP monomers that were included in an earlier version of this manuscript. JZ and PGH carried out simulations of this model with the 20 amino acid B_{ij} matrix, leading to the results in this version of the manuscript. SS and PGH wrote the majority of the manuscript. BJL and JAM calculated the REI score from real protein data in order to compare with the predictions of the model.

This work was supported by Discovery Grant 2017-05911 to Paul Higgs from the Natural Sciences and Engineering Research Council of Canada. This collaboration was initiated at KITP and supported in part by the National Science Foundation Grant NSF PHY-1125915, Multidisciplinary University Research Initiative Awards W911NF-09-1-0444 and W911NF-14-1-0411, and National Institutes of Health Award R35-GM122566-01. JAM is supported by a Medical Research Council Career Development Award (MR/M02122X/1) and a Lister Institute Research Prize Fellowship.

[1] E. D. Levy and S. Teichmann, Structural, evolutionary, and assembly principles of protein oligomerization, in Progress in Molecular Biology and Translational Science, pp. 25–51, Elsevier, (2013).

[2] J.-P. Changeux, Allostery and the Monod-Wyman-Changeux Model After 50 Years, *Annu. Rev. Biophys.*, **41**, 103–133, (2012).

[3] D. Eisenberg and M. Jucker, The Amyloid State of Proteins in Human Diseases, *Cell*, **148**, 1188–1203, (2012).

[4] L. T. Bergendahl and J. A. Marsh, Functional determinants of protein assembly into homomeric complexes, *Scientific Reports*, **7**, 4932, (2017).

[5] H. C. Jubb, A. P. Pandurangan, M. A. Turner, B. ochoa-Montano, T. L. Blundell and S. B. Asher, Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health, *Progress Biophys Mol Biol*, **128**, 3-13, (2017).

[6] F. Oosawa and M. Kasai, A theory of linear and helical aggregations of macromolecules, *Journal of Molecular Biology*, **4**, 10–21, (1962).

[7] I. A. Nyrkova, A. N. Semenov, A. Aggeli, M. Bell, N. Boden, and T. C. B. McLeish, Self-assembly and structure transformations in living polymers forming fibrils, *The European Physical Journal B*, **17**, 499–513, (2000).

[8] J. D. Schmit, K. Ghosh, and K. Dill, What Drives Amyloid Molecules To Assemble into Oligomers and Fibrils?, *Biophysical Journal*, **100**, 450–458, (2011).

[9] J. van Gestel, P. van der Schoot, and M. Michels, He-

lical transition of polymer-like assemblies in solution, *The Journal of Physical Chemistry B*, **105**, 10691–10699, (2001).

[10] J. van Gestel and S. W. de Leeuw, A Statistical-Mechanical Theory of Fibril Formation in Dilute Protein Solutions, *Biophysical Journal*, **90**, 3134–3145, (2006).

[11] M. Lynch, Evolutionary diversification of the multimeric states of proteins, *Proceedings of the National Academy of Sciences*, **110**, E2821–E2828, (2013).

[12] M. Lynch, The evolution of multimeric protein assemblages, *Molecular biology and evolution*, **29**, 1353–1366, (2011).

[13] J. H. Gillespie, *The causes of molecular evolution*, Oxford University Press, 1991.

[14] D. B. Lukatsky, K. B. Zeldovich and E. I. Skakhnovich, Statistically enhanced self-attraction of random patterns, *Phys Rev Lett*, **97**, 178101, (2006).

[15] D. B. Lukatsky, B. E. Skakhnovich, J. Mintseris and E. I. Skakhnovich, Structural similarity enhances interaction propensity of proteins, *J. Mol. Biol.*, **365**, 1596-1606, (2007).

[16] I. Ispolatov, A. Yuryev, I. Mazo and S. Maslov, Binding properties and evoution of homodimers in protein-protein interaction networks, *Nucl Acids Res*, **33**, 3629-3635, (2005).

[17] K. B. Zeldovich, P. Chen, B.E. Shakhnovich and E.I. Shakhnovich, A first-priciples model of early evolution: Emergence of gene families, species and preferred protein folds, *PLoS Comp Biol*, **3(7)**: e139, (2007).

[18] T. Kortemmea, A. V. Morozov, and D. Baker, An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes, *Journal of Molecular Biology*, **326**, 1239-1259, (2003).

[19] W. Zheng, N. P. Schafer, A. Davtyan, G. A. Papoian and P. G. Wolynes, Predictive energy landscapes for protein-protein association, *Proceedings of the National Academy of Sciences*, **109**, 19244-19249, (2012).

[20] M.M. Gromiha and K. Yugandhar, Integrating computational methods and experimental data for understanding the recognition mechanism and binding affinity of protein-protein complexes, *Progress Biophy Mol Biol*, **128**, 33-38, (2017).

[21] S. Miyazawa and R. J. Jernigan, Residue-residue potentials with a favourable contact pair term and an unfavourable high packing density term, for simulation and threading. *Journal of Molecular Biology*, **256**, 623-644, (1996).

[22] Y.C. Kim and G. Hummer, Coarse-grained models for simulations of multiprotein complexes: Application to Ubiquitin binding, *J Mol Biol*, **375**, 1416-1433, (2008).

[23] G.L. Dignon, W. Zheng, Y.C. Kim, R.B. Best, and J. Mittal, Sequence determinants of protein phase behavior from a coarse-grained model, *PLoS Comp Biol*, **14(1)**: e1005941, (2018).

[24] D. M. Taverna and R. A. Goldstein, Why Are Proteins So Robust To Site Mutations? *Journal of Molecular Biology*, **315**, 479-484, (2002).

[25] P. D. Williams, D. D. Pollock and R. A. Goldstein, Functionality and the evolution of marginal stability in proteins: Inferences from lattice simulations, *Evolutionary Bioinformatics*, **2**, 91-101, (2006).

[26] R.A. Goldstein and D.D. Pollock, Sequence entropy of folding and the absolute rate of amino acid substitutions, *Nature Evol Evol*, **1**, 1923-30, (2017).

[27] M. R. Betancourt and D. Thirumalai, Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes, *Protein Science*, **8**, 361–369, (1999).

[28] M. R. Betancourt and D. Thirumalai, Exploring the kinetic requirements for enhancement of protein folding rates in the groel cavity1, *Journal of molecular biology*, **287**, 627–644, (1999).

[29] M. Bulmer, The selection-mutation-drift theory of synonymous codon usage., *Genetics*, **129**, 897–907, (1991).

[30] P. G. Higgs and W. Ran, Coevolution of Codon Usage and tRNA Genes Leads to Alternative Stable States of Biased Codon Usage, *Molecular Biology and Evolution*, **25**, 2279–2291, (2008).

[31] M. Manhart and A. V. Morozov, Protein folding and binding can emerge as evolutionary spandrels through structural coupling, *Proceedings of the National Academy of Sciences*, **112**, 1797–1802, (2015).

[32] K. R. Albe, M. H. Butler, and B. E. Wright, Cellular concentrations of enzymes and their substrates, *Journal of theoretical biology*, **143**, 163–195, (1990).

[33] R. Milo, What is the total number of protein molecules per cell volume? a call to rethink some published values, *BioEssays*, **35**, 1050–1055, (2013).

[34] E. D. Levy, J. Kowarzyk, and S. W. Michnick, High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation, *Cell reports*, **7**, 1333–1340, (2014).

[35] S. Jones and J. M. Thornton, Analysis of protein-protein interaction sites using surface patches, Journal of Molecular Biology, **272**, 121-132 (1997).

[36] E. D. Levy, S. De and S. A. Teichmann, Cellular crowding imposes global constraints on the chemistry and evolution of proteomes, *Proceedings of the National Academy of Sciences*, **109**, 20461-20466, (2012).

[37] O. Vidal, R. Longin, C. Prigent-Combaret, C. Dorel, M. Hooreman, and P. Lejeune, Isolation of an Escherichia coli K-12 Mutant Strain Able To Form Biofilms on Inert Surfaces: Involvement of a New ompR Allele That Increases Curli Expression, *Journal of Bacteriology*, **180**, 2442–2449, (1998).

[38] B. L. Kagan, H. Jang, R. Capone, F. T. Arce, S. Ramachandran, R. Lal, and R. Nussinov, Antimicrobial Properties of Amyloid Peptides, *Molecular Pharmaceutics*, **9**, 708–717, (2011).

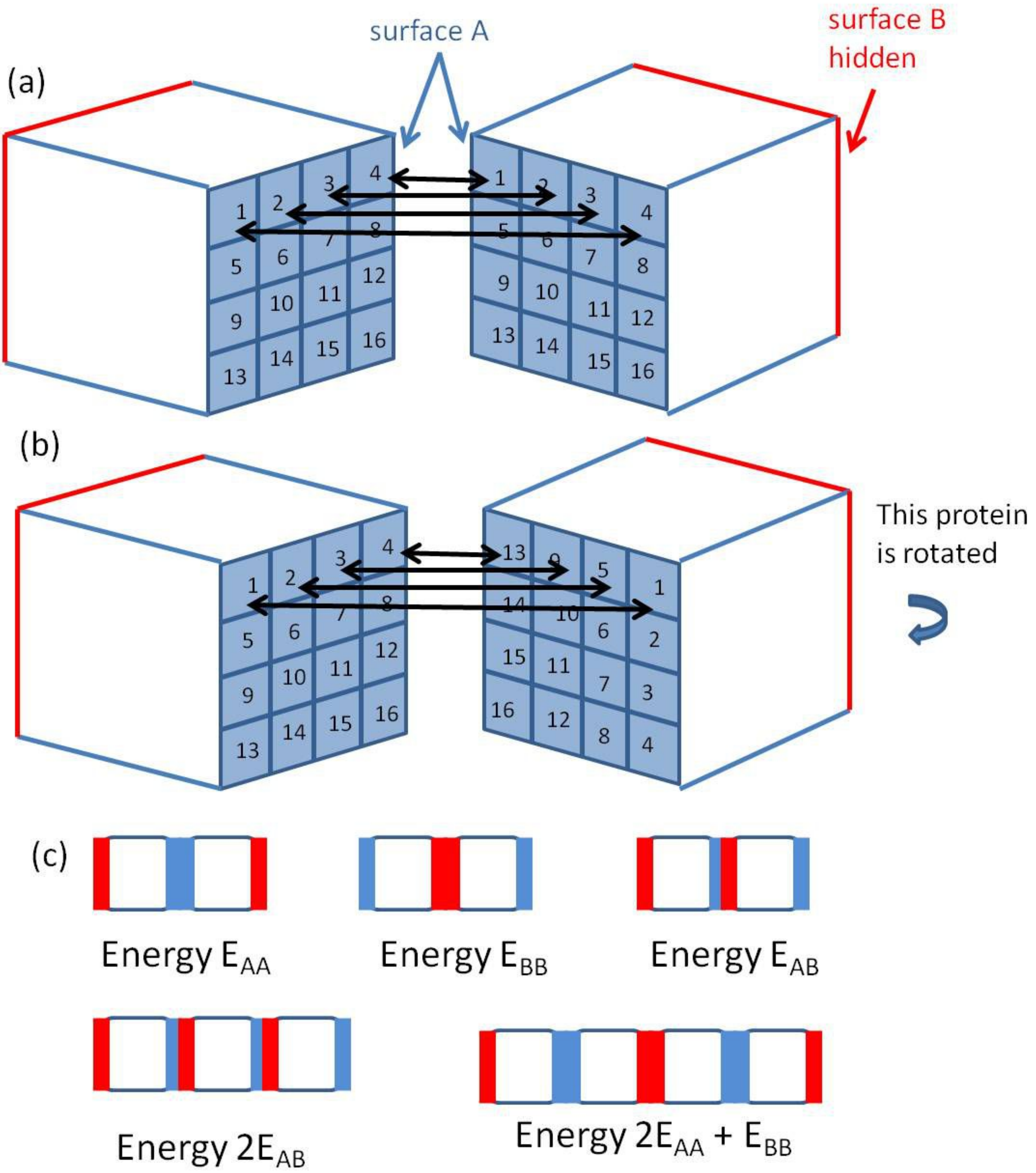
[39] N. S. de Groot, M. Torrent, A. Villar-Piqué, B. Lang, S. Ventura, J. Gsponer, and M. M. Babu, Evolutionary selection for protein aggregation, *Biochemical Society Transactions*, **40**, 1032–1037, (2012).

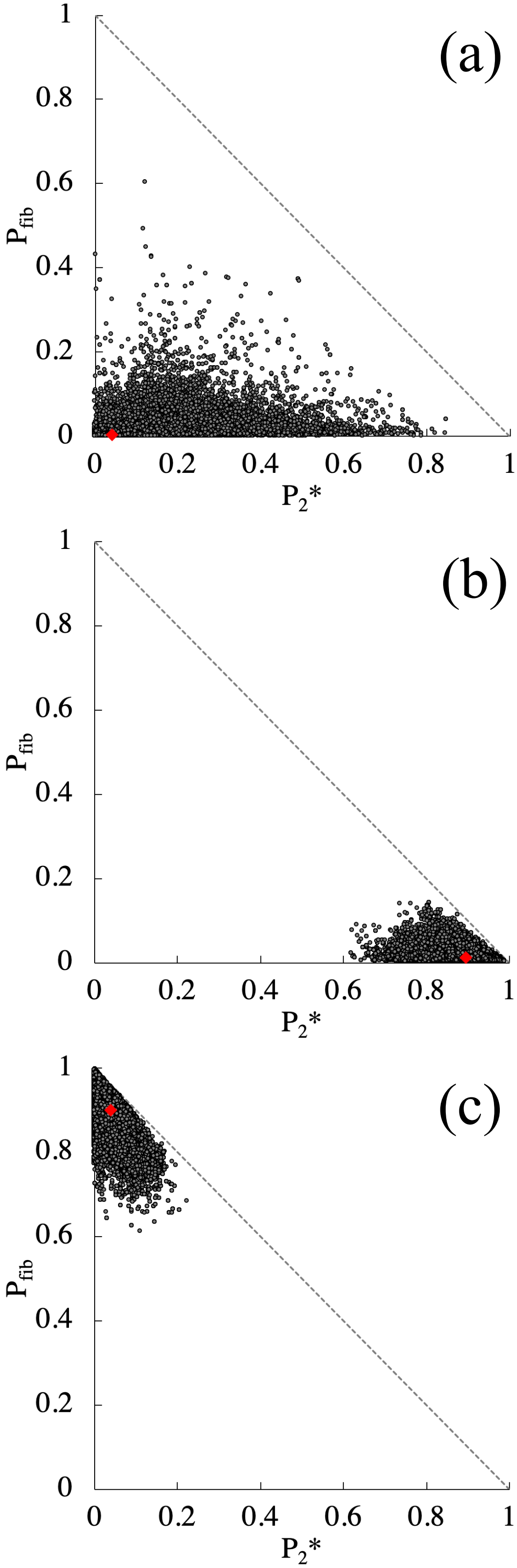
[40] S. E. Ahnert, J. A. Marsh, H. Hernández, C. V. Robinson, and S. A. Teichmann, Principles of assembly reveal a periodic table of protein complexes, *Science*, **350**, aaa2245, (2015).

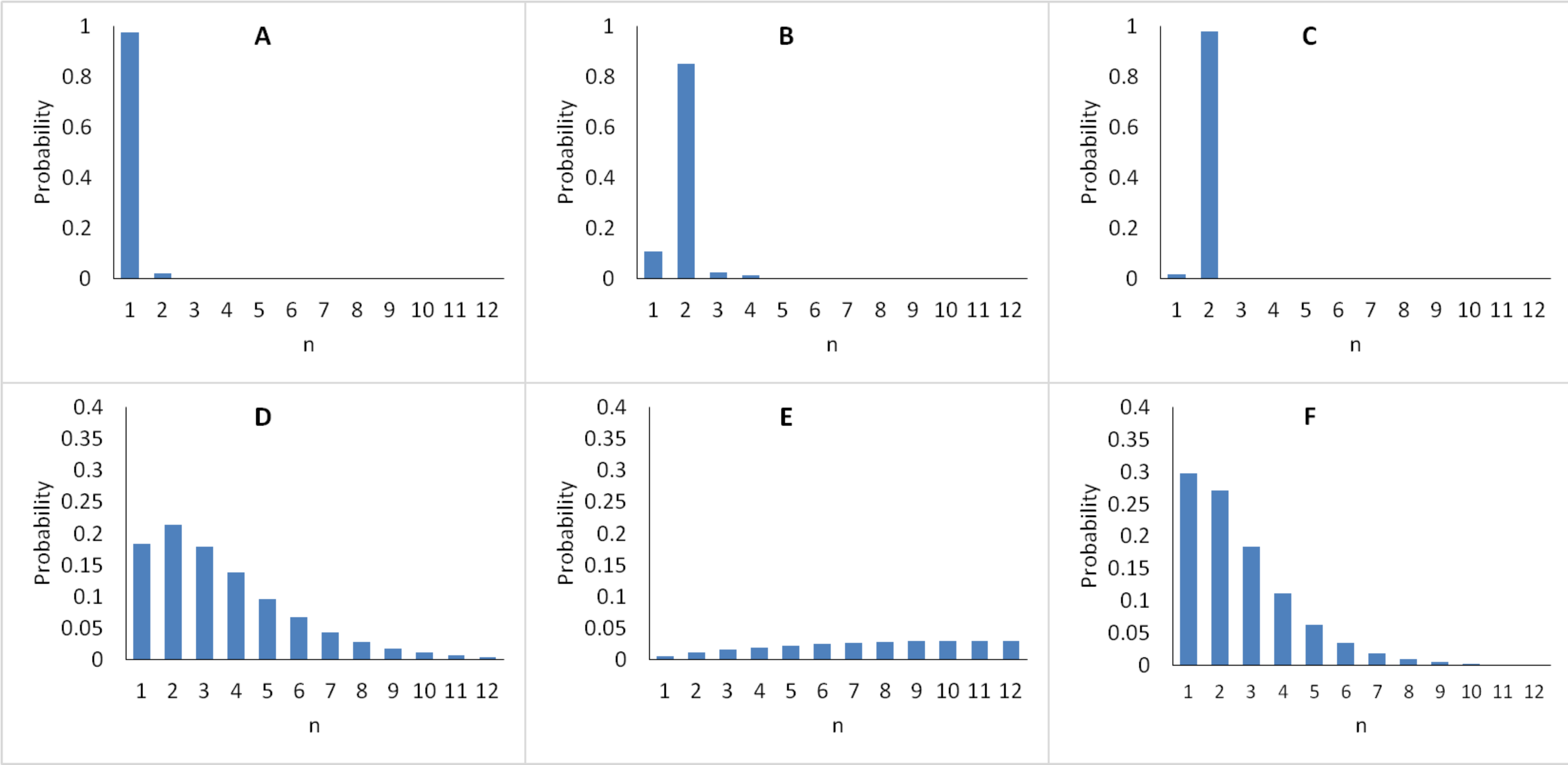
[41] J. E. Dayhoff, B. A. Shoemaker, S. H. Bryant, and A. R. Panchenko, Evolution of Protein Binding Modes in Homooligomers,'*Journal of Molecular Biology*, **95**, 860–870, (2010).

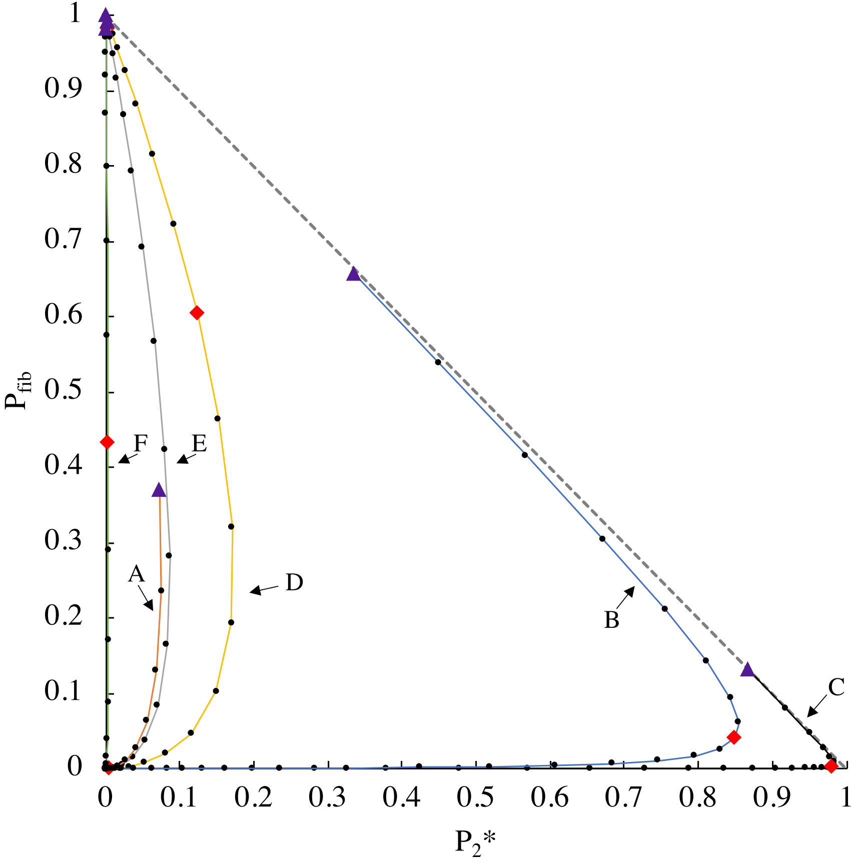
[42] J. A. Marsh and S. A. Teichmann, Protein flexibility facilitates quaternary structure assembly and evolution, *PLoS biology*, **12**, e1001870, (2014).

[43] G. Wang and R. L. Dunbrack Jr, Pisces: a protein sequence culling server, *Bioinformatics*, **19**, 1589–1591, (2003).

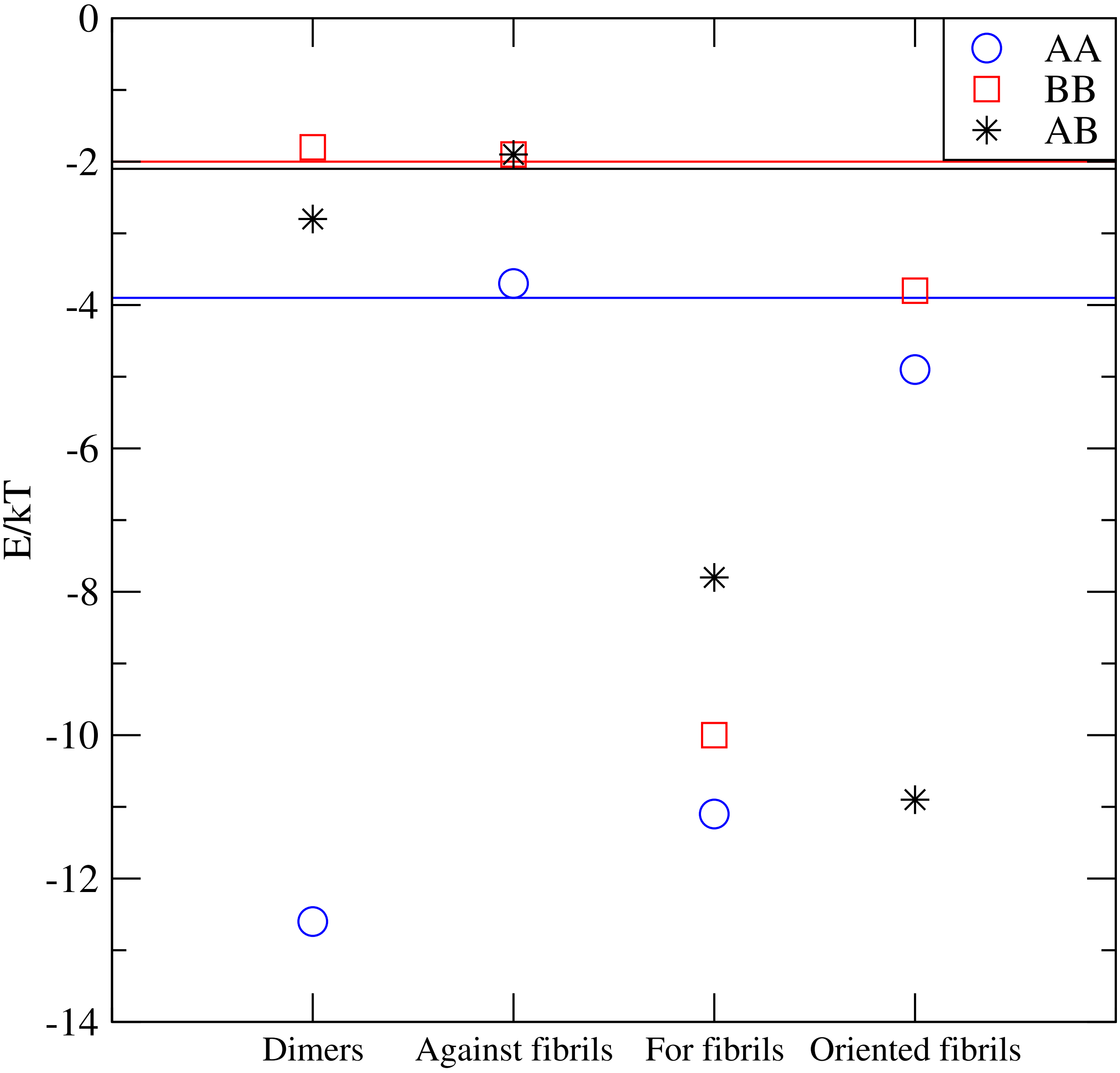

surface A
surface B hidden
(a)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
(b)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
13 9 5 1 14 10 6 2 15 11 7 3 16 12 8 4
This protein is rotated
(c)
Energy E_{AA}
Energy E_{BB}
Energy E_{AB}
Energy $2E_{AB}$
Energy $2E_{AA} + E_{BB}$

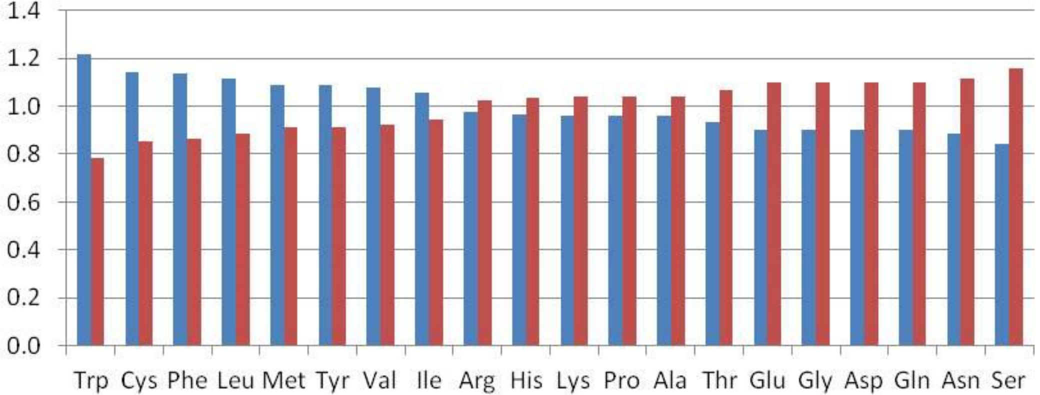

(a)
1
0.8
0.6
0.4
0.2
0
P_{fib}
0
0.2
0.4
0.6
0.8
1
P_2*
(b)
1
0.8
0.6
0.4
0.2
0
P_{fib}
0
0.2
0.4
0.6
0.8
1
P_2*
(c)
1
0.8
0.6
0.4
0.2
0
P_{fib}
0
0.2
0.4
0.6
0.8
1
P_2*

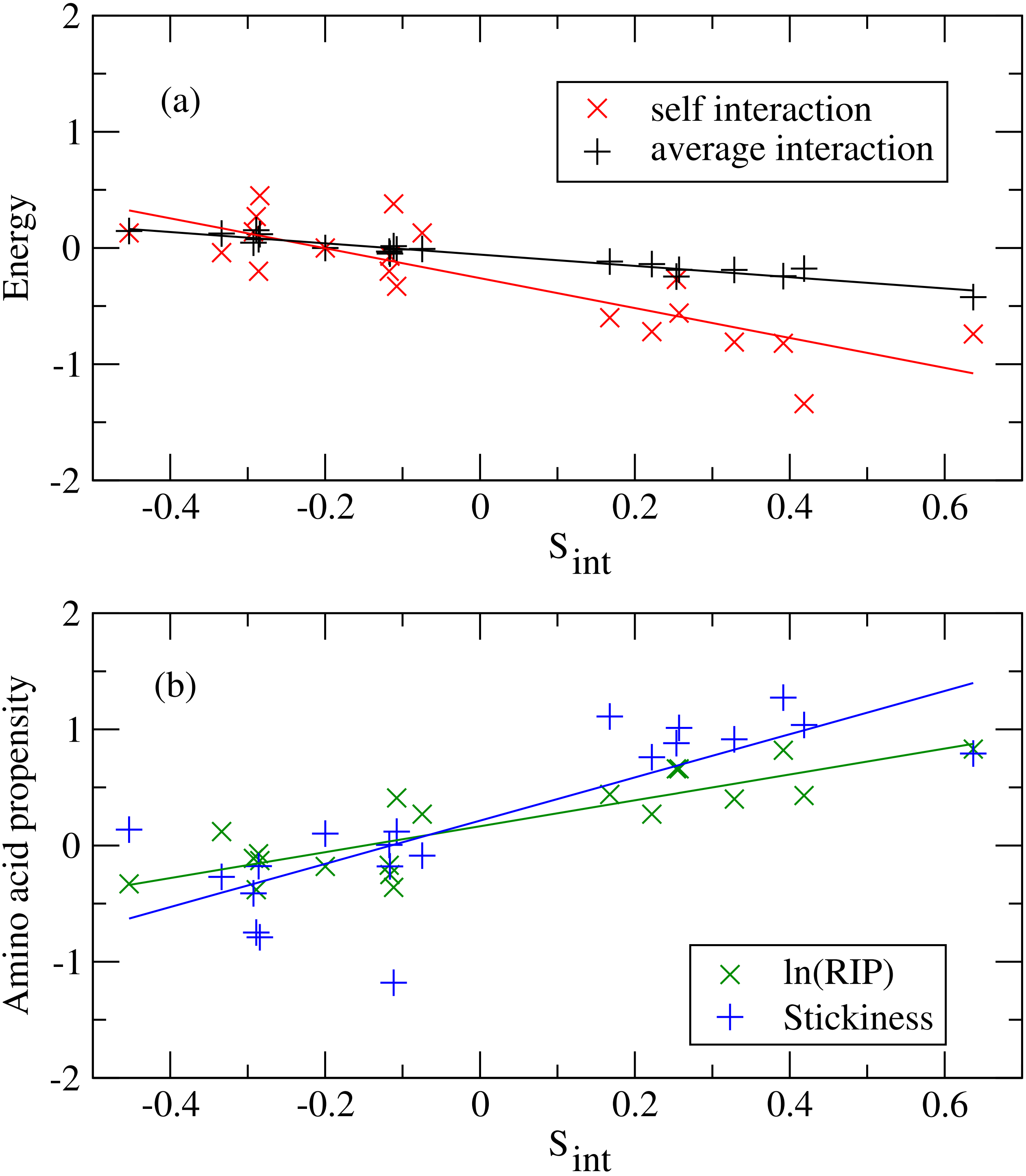

A
B
C
D
E
F
Probability
n
1
0.8
0.6
0.4
0.2
0
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
1 2 3 4 5 6 7 8 9 10 11 12

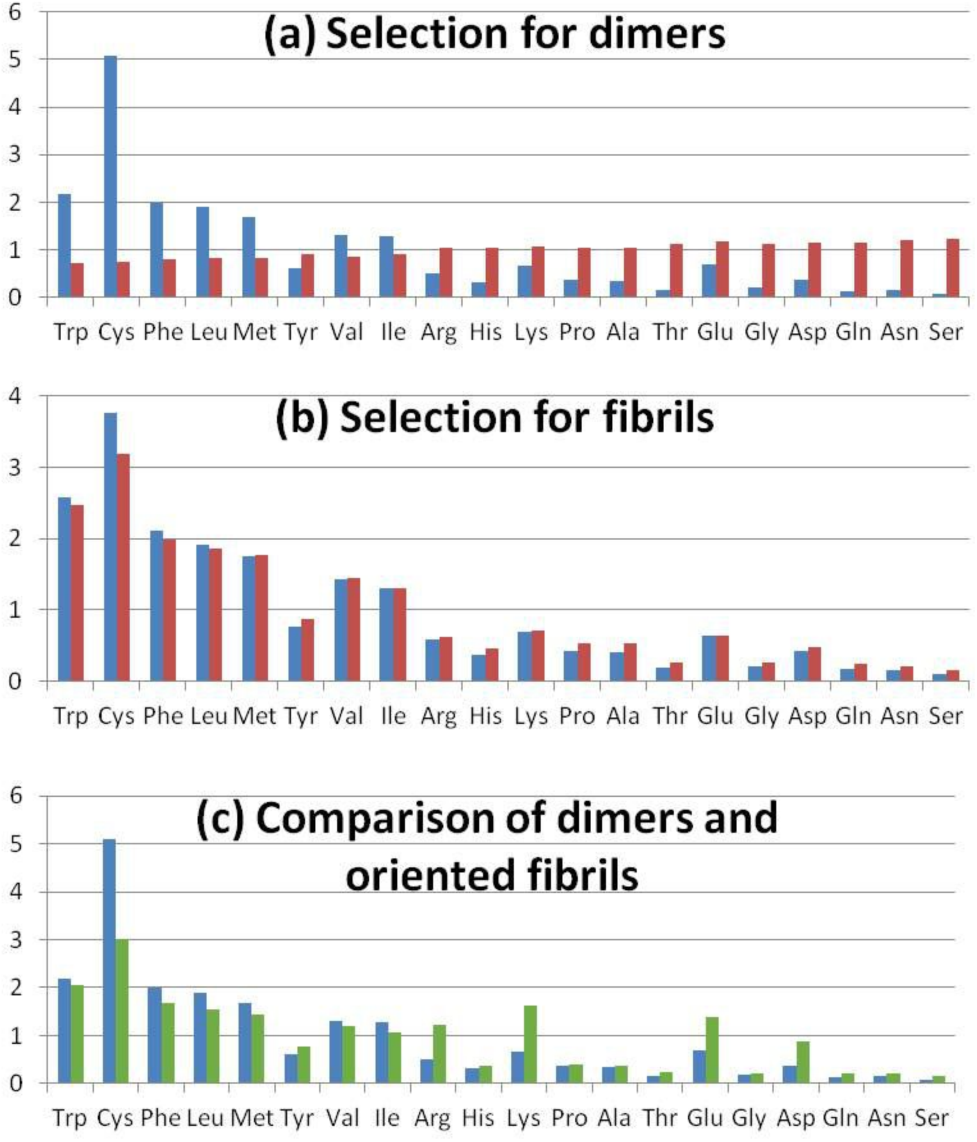


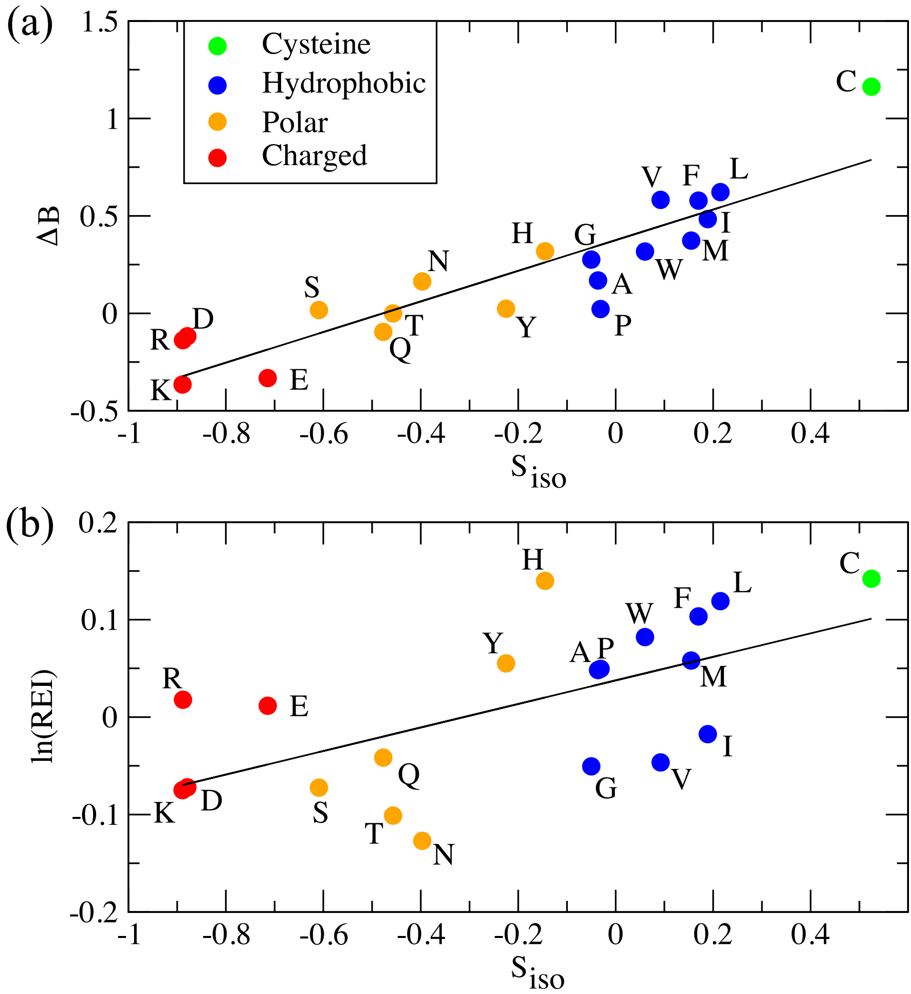
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P_{fib}
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
P_2*
A
B
C
D
E
F


0
-2
-4
-6
-8
-10
-12
-14
E/kT
Dimers
Against fibrils
For fibrils
Oriented fibrils
AA
BB
AB


1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Trp
Cys
Phe
Leu
Met
Tyr
Val
Ile
Arg
His
Lys
Pro
Ala
Thr
Glu
Gly
Asp
Gln
Asn
Ser


(a)
self interaction
average interaction
Energy
2
1
0
-1
-2
-0.4
-0.2
0
0.2
0.4
0.6
S_{int}
(b)
ln(RIP)
Stickiness
Amino acid propensity
2
1
0
-1
-2
-0.4
-0.2
0
0.2
0.4
0.6
S_{int}


(a) Selection for dimers
0
1
2
3
4
5
6
Trp Cys Phe Leu Met Tyr Val Ile Arg His Lys Pro Ala Thr Glu Gly Asp Gln Asn Ser
(b) Selection for fibrils
0
1
2
3
4
Trp Cys Phe Leu Met Tyr Val Ile Arg His Lys Pro Ala Thr Glu Gly Asp Gln Asn Ser
(c) Comparison of dimers and oriented fibrils
0
1
2
3
4
5
6
Trp Cys Phe Leu Met Tyr Val Ile Arg His Lys Pro Ala Thr Glu Gly Asp Gln Asn Ser



(a)
● Cysteine
● Hydrophobic
● Polar
● Charged
ΔB
1.5
1
0.5
0
-0.5
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
S_{iso}
C
V
F
L
I
M
W
H
G
A
P
N
S
Y
T
Q
D
R
K
E
(b)
ln(REI)
0.2
0.1
0
-0.1
-0.2
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
S_{iso}
H
C
L
F
W
Y
A
P
M
R
E
I
Q
V
G
K
D
S
T
N