Classification: Biological Sciences, Biophysics

The network of sequence flow between protein structures

Leonid Meyerguz, Jon Kleinberg, and Ron Elber[1]
Department of Computer Science, Cornell University, Ithaca NY 14853

[1]corresponding author: Ron Elber, Department of Computer Science, Cornell University, Ithaca, NY 14853. e-mail: ron@cs.cornell.edu; Phone: 607-255-7416; FAX: 607-255-4428

Manuscript information: The number of text pages 11, number of figures 5, number of tables 0.

Abstract

Sequence-structure relationships in proteins are highly asymmetric since many sequences fold into relatively few structures. What is the number of sequences that fold into a particular protein structure? Is it possible to switch between stable protein folds by point mutations? To address these questions we compute a directed graph of sequences and structures of proteins, which is based on 2060 experimentally determined protein shapes from the Protein Data Bank. The directed graph is highly connected at native energies with “sinks” that attract many sequences from other folds. The sinks are rich in beta sheets. The number of sequences that transition between folds is significantly smaller than the number of sequences retained by their fold. The sequence flow into a particular protein shape from other proteins correlates with the number of sequences that matches this shape in empirically determined genomes. Properties of strongly connected components of the graph are correlated with protein length and secondary structure.

Introduction

As data on protein sequences and their variations become more accessible (following the abundance of large scale sequencing and gene expression projects) it is clear that protein structures serve as evolutionary templates. Similar protein backbones are used again and again to create proteins with adjusted functions in response to environmental variations, or at random. This asymmetric relationship is of considerable interest in the study of protein evolution and design, and has received considerable attention. How many sequences fold to a common structure, or equivalently what is the sequence capacity (or designability) of a known fold? Past theoretical and computational studies are primarily focused on the thermal stability of the proteins. The stability is estimated by an energy calculation of threaded sequences in a known structure. The theory and calculations can be divided (roughly) into two categories: (i) General theories (1-6) and exhaustive simulations of simple model systems (7-11), and (ii) Accurate and detailed modeling of a few proteins (12-16). The studies of class (i) provide a universal view of sequence-structure matches and their variations. Investigations of class (ii) made specific predictions on protein folds that are straightforward to test experimentally. The function of interest, protein designability or sequence capacity, was estimated theoretically and by computations. However, neither of these calculations consider explicitly *all* structures of the Protein Data Bank (PDB) (17). Quantitative extrapolations from approximate theories, lattice models, or detailed simulations of a few proteins to other folds may not be obvious. Furthermore, collective behavior of the evolutionary process, not restricted to a single or a few proteins, may go unnoticed.

Explicit calculation of sequence capacity of all protein folds is of particular interest as genomic-scale experiments are emerging, making it possible to determine sequence selection mechanisms (18-20). The experiments assess the contribution of sequence capacity, estimated from theory or simulation, and compare it to natural mutation rates. We have developed a computational model in which the sequence capacity was directly computed for a representative set of structures from the PDB (3660 folds) (21). In the calculation only the energy function is approximate while the PDB structures and their corresponding sequences are sampled significantly. The sampling allowed for statistical convergence of the capacity. In addition to sequence capacities of all folds we computed an intriguing temperature relationship between the folds.

We sampled only experimentally determined structures from the PDB, so an obvious question is the completion of the set. Arguments were made that the PDB is indeed complete (22, 23) with the current thousands of distinct folds. This argument further supports the creation of a comprehensive model of protein structure space and their sequence capacities, and the progressive refinement of this model. In (21) we did not consider the possibility that mutated sequences of a particular structure will fold to different shapes (sequence migration - we use “migration” to denote sequences that evolve in one fold and end up in another structure). This analysis of sequence migration is particularly timely with the growing experimental evidence for pairs of proteins with high percentage of sequence identity and *alternate structures*. These “interface” sequences were illustrated experimentally on model systems (24-26) and on proteins (26-30). What is the impact of the interfaces on protein evolution and design? Interesting

analyses of existing structures and identification of continuous evolutionary changes are in (31, 32) suggesting the mixing of folds during the evolutionary processes. The major goal of the present manuscript is the development of a complete computational model for protein space as a network, with the nodes of the graph representing the protein folds and directed edges accounting for the flow of sequences in and out of the folds. The in-degree of a fold is the number of edges that point to it, or the number of other folds that lose sequences to that structure as a result of point mutations. Similarly the out-degree of a fold is the number of edges that carry sequences from that fold to other proteins shapes. An edge indicates loss of sequences that are energetically compatible with one structure to another fold. While other interesting network models for protein space were proposed in the past (33-36), they were not based on explicit modeling of the kinetics of evolution (i.e. sequence mutations and migration between structures) which is done here.

Computational model

We first summarize the basic components of the computational model. We directly compute the absolute number of sequences that fold to each member of a comprehensive sample of protein structures. We also calculate the number of *transitional* sequences between folds. A transitional sequence allows, with a single point mutation, to flip between alternative stable structures. The computation is based on a stochastic sampling of sequences with provable polynomial convergence in the sequence length. The sampling is done one structure at a time. A sequence evolves in one particular structure and the number of sequences (sequence capacity) below a particular energy is estimated as well as the number of sequences that were lost to other folds. Our selection criterion is based on a threading energy function that makes it possible to estimate microcanonical partition functions of sequence space in the neighborhood of each stable basin (a protein structure) and the entropy of the transition state between pairs of folds. The numbers of sequences of a fold and of the transitional sequences between pairs of folds form the network of sequence flow which is the prime result of the present manuscript.

In order to estimate the sequence capacity of a representative set of structures we selected chains from the Protein Data Bank (PDB) covering over 90% of the protein families in the SCOP database classes a-e (37) (representing single and multi-domain alpha and beta proteins). We started with a large subset of 14,000 chains from the PDB, chosen so that no two chains have more than 70% sequence identity. We compared this subset against the families in SCOP and eliminated chains that yielded redundant representations while making sure that our coverage of SCOP families remained as high as possible. Afterward, we compared the remaining chains using the TM-Align algorithm (38). The range of the TM-score is between 0 and 1, where 1 is identity. Out of every pair of proteins with TM-Score > 0.8, we removed one, thereby eliminating structural redundancies. The resulting data set used in this work contains 2060 protein chains.

In our earlier study (21) we considered the sequence capacity $N_i(E)$, which is the number of sequences with energy lower than E of fold *i*. To characterize the properties of the new set we define the sequence capacity with competition, $C_i(E)$ as follows: it is

the number of sequences for which the energy E_i in fold i is lower than E and also lower than the energy E_j in any of the competing folds j. In the present study we are using a model close to gapless threading (39) to check for competing folds. We do not allow general alignments with deletions and insertions when we fit sequences into structures. The shorter sequence of two matched proteins is always continuous and considered in full. Deletions and insertions at the beginning and the end of sequences score zero, which is a penalty since the THOM2 energy function (see below) is negative on the average. Hence, gaps do not make energetic contributions in our model.

The energy used in the present study is THOM2 (40). It was used in an earlier study of $N_i(E)$ which is extended here to the study of $C_i(E)$ and the stability network. It is also an integral part of our structure prediction program LOOPP (http://cbsuapps.tc.cornell.edu/loopp.aspx) and provides a useful signal to detect similarities between folds of proteins. THOM2 captures the environment of each structural site by assigning a score, $u(\alpha, m)$, for each *contact* to a structural site. A *contact* is assumed if the distance between the geometric centers of two amino acid side chains is less than 6.4 angstrom. The score is determined from a lookup table using the type of amino acid, α, at the site of interest and the number of neighbors m to the contact site. The total energy of a protein is a sum of the site contributions: $E = \sum_{l=1,...L} \sum_k u_{lk}(\alpha_l, m_{kl})$ where the index l is running over the structural sites and the index k over the contacts of the site l. THOM2 performs quite well on the set of folds we considered. It recognizes in 1885 of the 2060 proteins the native structure as the best fold of the native sequence. The remaining 175 structures are not competitive for sequences within the network and therefore do not influence its behavior significantly. The 175 folds have fewer than ten sequences with better energy than the native sequence. From a bioinformatic perspective the THOM2 energy is particularly useful since an efficient alignment algorithm (dynamic programming (41)) is known. From the perspective of estimating $N(E)$ and $C(E)$ this energy function is also of significant value. It was shown (42) that the Markov chain of the algorithm described below that relies on THOM2 is well mixed (approaching the desired distribution) after a polynomial number of steps in the sequence length. It is therefore expected that an efficient calculation of the sequence capacity can be made with THOM2. In contrast we cannot demonstrate that the Markov chain for pairwise potentials is ergodic.

Consider a protein sequence A_t of length L, a fold X_i, and energy function $E_t \equiv E(A_t, X_i)$. We set two upper energy boundaries for an intermediate estimate of the number of sequences, E_s and E_{s+1}, ($E_s < E_{s+1}$). Both boundaries are chosen empirically, such that $E_t < E_{s+1}$ and the ratio $N(E_{s+1})/N(E_s)$ is of order one (between one and twenty). Note that typically $N(E)$ grows exponentially with the protein length L and its maximum is 20^L which can make the choices of the energies a little tricky. The determination of the ratios of $N(E_{s+1})/N(E_s)$ for different energies E_s and E_{s+1} is the

target of the current calculation. Each individual ratio is estimated with a randomized algorithm (21, 42) as described below: Starting from A_t we modify at random one of its amino acids. If the energy of the new sequence is larger than E_{s+1}, the new sequence is rejected and a new trial is made based on A_t. If the new energy is lower or equal E_{s+1} the step is accepted and we add one to the counter l_{s+1}. In this way, we performed a random sampling on the space of sequences with energy below E_{s+1}. We keep track of the number of steps that this walk spends below energy E_s in a second counter l_s; the ratio l_{s+1}/l_s can then be shown to give a polynomial-time approximation to $N(E_{s+1})/N(E_s)$.

In a previous study we approximate $N(E_n)$, the number of sequences with energy lower than the energy of the native sequence, A_n, by a successive ratio $N(E_n) = N(E_{ref})\frac{N(E_1)}{N(E_{ref})}\frac{N(E_2)}{N(E_1)}\ldots\frac{N(E_S)}{N(E_{S-1})}\frac{N(E_n)}{N(E_S)}$, where $N(E_{ref})$ is the number of sequences at a reference energy which we can estimate directly. For example, the average energy, E_{mean}, and $N(E_{mean})$ can be determined by direct random sampling of sequences. The capacity $N(E_{mean})$ is quite close to $1/2\cdot 20^L$ (though we compute it for every fold in the set). Since the difference between $N(E_{ref})$ and $N(E_n)$ can be exponential in the protein length, we establish S intermediate ratios to satisfy the above requirement that each individual ratio is $O(1)$. For each intermediate ratio we typically generate a sample of a few million sequences. We repeat this calculation for every fold in the set.

In the calculations of $C_i(E)$ the number of sequences that fit a fold X_i with competition, we adjust the counting as follows. As before, we generate a Markov chain in sequence space such that $E_t < E_{s+1}$ and we compute the ratio l_{s+1}/l_s. In addition to previous calculation (21, 42) we check for each sequence whether one of the alternative folds X_j has energy $E_j < E_t$. We define c_{s+1} as the number of sequences sampled below E_{s+1} which do not have better energies in other folds, and r_{s+1} as the total number of sequences sampled below the energy E_{s+1}. We compute the estimate $C(E_{s+1}) \approx N(E_{s+1})c_{s+1}/r_{s+1}$, and estimate $C(E_s)$ in the same manner. We also compute $m_s(i \to j)$, the number of sequences that migrate from structure i to structure j. This information is sufficient to describe the directed graph we are after.

The above procedure can increase the computational cost by several orders of magnitude compared to counting without competition (instead of a single energy evaluation per sequence, we may need thousands of evaluations). However, we can employ additional heuristics to significantly reduce the running time of our algorithm. For instance, we observe that a structure j that cannot compete with structure i in the energy interval E_s

is highly unlikely to be competitive at any interval $E_{s'}$ such that $E_{s'} < E_s$. Therefore, whenever a structure j has been non-competitive for three successive energy intervals, we will remove it from the list of competitors for the next two intervals (rounds). After "sitting out" for two rounds, the structure j will re-enter the competition; however, if it remains non-competitive, it will be eliminated again for four rounds, then for eight rounds, etc. Empirically, we observe that re-entering structures are almost always eliminated outright, and never have any significant effect on the competition (e.g. $m_s(i \to j)$ is close to zero for all $s' < s$). The heuristic allows us to significantly cut down on the number of structures we need to consider at any given time, and generally increases the efficiency of our algorithm.

Results

In figure 1 we show a schematic view of the two largest strongly connected components of the graph (a strongly connected component of a directed graph is a maximal set inside which every node has a path to every node). A directed edge is drawn from protein *i* to *j* if at least a fraction c of the sequences with energy below the native energy of *i* migrate to *j*. The minimal fraction c to establish an edge in figure 1 (0.00375) was chosen for clearer visualization. The size of the largest component is 320 structures, and the second largest 90, (the third is 39). For clarity in this image, at most five out-going edges with the largest weights were kept for each node.

Examining the properties of the components we realize that the largest of the two includes proteins that are unusually long. The average length of proteins of the largest component is 516 amino acids, while the average length of the whole set is 260. However, the average length of proteins in the second component is only 200 amino acids. This length is shorter than the average length of the set, which suggests that the length is not the only factor leading to the strong connectivity. We define the secondary structure content as the fractions of residues in an alpha helix or a beta sheet configuration according to the DSSP program (43). The largest component is slightly richer with helical content compared to the second largest component (0.276 versus 0.252) and slightly poorer with sheets (0.232 versus 0.276). The contribution of the secondary structure to in-degrees is clearer when correlations are examined. We emphasize that no correlation is observed between secondary structure and protein length ($\rho = -0.002$ and p-value of 0.1). The Spearman correlation coefficient of beta sheet content and in-degree is $\rho = 0.215$ with p-value less than 10^{-12}. Another piece of evidence for the importance of protein length and beta sheet content for in-degree are the "sinks". The graph clearly shows the existence of structures that attract sequences from many other folds. The potential existence of sinks was noted in the past based on 3x3x3 lattice simulations (6, 33). All the top attractors are in the largest component. The PDB identifiers of the proteins with at least 100 in-degrees are: 1TYV (272 in-degrees, length of 542 amino acids, and high beta sheet content), 1IDK is similar in its characteristics (152 in-degrees, length of 359 amino acids, and high beta sheet content), and so is the next in line -- 1OFL:A (100 in-degrees, 481 amino acids, and a beta protein). Yet another example is 1RWR which is the fifth strongest attractor with 64 in-degrees and moderate

length (301 amino acid and a beta protein). A useful property that strongly correlates with the graph in-degree is the contact density (the total number of contacts of a protein divided by the sequence length). It is higher for the largest component (1.626 versus 1.531) which is not surprising since higher contact density is expected for longer proteins with smaller surface-volume ratio. Indeed the contact density is highly correlated with the protein length. One may expect that the in-degree of a structure will strongly correlate with the sequence capacity. The correlation however is not so strong once the length effect is factored out. The correlation coefficient of $\log\left[C(E)/20^{L}\right]$ with the in-degree is 0.468 and of $\log\left[N(E)/20^{L}\right]$ is -0.169.

The above analysis focused on a particular definition of an edge which was useful for graphical purposes. Another definition that we examined in detail is based on the size of the network. Two nodes i and j are connected by a directed edge if the fraction of sequences that migrate from i to j is larger or equal to $1/K$ ($K = 2060$ is the total number of folds in our set). In figure 2 we show the distribution of the number of "in" edges.

The total number of in-edges is 785182 suggesting that the connectivity of the graph is dense. Besides the dominant feature at zero, the distribution also shows a long tail to much higher values (up to 2002 in edges!). The proteins that feature in this high number of in-degree class (e.g. 1K32:A) are (again) enriched with beta sheet structures compared to the rest of the proteins. This observation suggests that the properties of the sinks are not sensitive to edge cutoff value. The Spearman correlation coefficient of length and in-degrees is highly significant $\rho = 0.623$ with p-value significantly less than 10^{-12}. It is interesting to note that the distribution of the number of out-edges is considerably more focused and no long tails are observed. It is peaked at a value of about 470 for the degree and is not correlated with the number of in edges of a particular fold.

Since the in-degrees show such a striking behavior we examined if there are correlations between the distribution of in-degrees and the number of sequences for each fold family that are observed experimentally. For every native sequence in our database we identify all related sequences in the NR database (44). The matching was done with BLAST (45) with an E-value of 0.001 and the BLOSUM 60 substitution matrix (46) (no significant changes in the results reported below were found for E value of 0.01). Since longer proteins may have more than one domain (and therefore may have independent blast hit to different domains) we divided the number of sequence hits by the number of domains. The sampling of sequences is significant and on the average we assigned about 340 sequences to one fold. We have found that the number of sequences that match a particular fold correlates with the number of in-edges with Spearman's correlation coefficient of 0.223. While this correlation suggests that many other factors are involved in evolutionary processes (besides stability) in accord with observations of others (19), it is nevertheless highly significant (p-value of less than 10^{-12}).

For every fold we can also determine an ideal energy E^* where the fraction of retained sequences – i.e. the quantity $C(E^*)/N(E^*)$ – is maximized. This energy is always lower than the energy of the native sequence and for a large number of proteins $C(E^*)/N(E^*)=1$. Hence, some proteins are able to retain all their sequences at the ideal energy. In a sharp contrast the other folds retain only a small fraction of their sequences as demonstrated in figure 3, dividing the fold family into two broad classes. The proteins with high retention factors are also with high contact density.

From the discussion above it is clear that the edges of the graph are a function of an ad-hoc cutoff value of the transmission probability between nodes and the energy of the calculation (at E^* the number of edges is likely to be minimal). It is therefore useful to explore different values of cutoff and of sequence-counting energy (between E^* and E_{nat}). In Figure 4 we plot the number of components of the graph computed as we varied these parameters

In figure 5 we show the functions $\log\left[N(E_n)/20^L\right]$ and $\log\left[C(E_n)/20^L\right]$ for all proteins in the set plotted as a function of the contact density (the total number of contacts of a protein molecule divided by the protein length L). We call these functions *the density of capacity* (with or without competition).

We observe that $\log\left[N(E_n)/20^L\right]$ is a non-increasing function (on the average) of the contact density. This is easy to explain since structural sites of amino acids with higher contact density are more selective and a smaller fraction of sequences is found below the native energy. The function $C(E_n)/20^L$ behaves differently and shows a maximum as a function of the contact density. The deviation of $C(E_n)/20^L$ from $N(E_n)/20^L$ is the clearest for low contact density while at high values both functions are more similar. At low contact density the native structure is only marginally stable making $N(E_n)/20^L$ large (it is easy to find sequences with better or comparable energy to the native energy for this particular fold). However, the marginal stability of structures with low contact density suggests that it is easy to find alternative folds with lower energies for the probe sequence. The availability of alternative folds in the calculation of $C(E_n)$ significantly reduces the number of sequences for marginally stable proteins compared to the results of $N(E_n)$. On the other hand when the contact density is large, the fraction of sequences acceptable to that fold is smaller, and their energies are lower making it more difficult to find alternative folds for a particular sequence. Hence for large contact density the two densities of capacity are more similar. The more accurate function for estimating sequence capacity, $C(E)$, has an intriguing maximum at about 1.5 for the contact density, which is the largest value observed for the density of sequence capacity or protein designability.

We finally discuss potential sources of errors in our calculations. While the convergence of our sampling procedure is mathematically sound, two other components of the model may have significant errors. First, our set of alternative structures is incomplete even if the PDB is, since in our studies we use only gapless alignments. The presence of gaps will significantly increase the number of alternate structures (47). Second, our energy function, which is a one-body potential is less accurate than more sophisticated energy models that are available. The first point will tend to make the network denser while the second point probably more diluted. Both of these choices were made to facilitate the construction of the network. We examine tens to hundreds of millions of sequences for each particular fold. It would not be practical for us to create a network at the same level of sampling accuracy and a comprehensive view of the PDB with significantly more complex models.

Discussion

Perhaps the most striking observation of the present manuscript is the high connectivity between protein folds induced by sequence migration. Both the usual notion of a unique and stable protein fold and the success of homology modeling (many sequences fold into one particular shape) are in conflict with the picture of a densely connected space of protein structures by sequence evolution. This conflict is easily resolved. The number of sequences that migrates between folds is significantly smaller than the total number of sequences available to a particular fold. For longer proteins the probability of structural flip is particularly small. Given that the number of homologous proteins known today to a particular fold is in the hundreds to thousands, experimental detection of a transition would be hard to come by. Nevertheless, a number of intriguing sequence migration and structural shifts were already observed (26-30). Further searches for such transitions can benefit from interactions between experiment and simulations; the simulations might be able to guide the search for these rare events.

The high connectivity that we observed is for energies that are below the native energies of these proteins. These transitions are therefore direct with a single point mutation. They are possible, within our energy model, without causing unfolding. Obviously additional possibilities for these exchanges will be open once more complex moves are considered (such as domain swap). However, even with the highly restricted move set the connectivity is quite significant and may serve a purpose. It is well known that proteins have native sequences that are far from optimal in their correct folds. A number of speculations were suggested to explain this observation, like making sure the protein folds even after potentially damaging point mutation (large sequence capacity), retaining flexibility necessary for function, etc. Here we are adding one more speculation. The native sequences of experimental folds are far from optimal to retain the possibility of *structural* flexibility, adjusting protein shapes by local mutations in response to environmental pressure. These transitions will be obviously rare but still possible according to our calculations and to a few experimental examples listed above. Hence, the present paper opens the way for speculation on structural evolution that results from point mutations. The observed structural flips are not necessarily restricted to proteins and it is possible that molecules like RNA will show similar behavior.

Acknowledgements
This research was supported by NIH grant GM067823 to RE. The calculations were performed on a computer cluster purchased with NIH grant RR020889.

Figure legends
Figure 1: The two largest strongly connected components of the network of sequence flow between protein folds. Protein space is presented as a directed graph in which a node is a protein shape and the directed edge denotes a flow of sequences from one fold to another. Sequence flow is created when a sequence that is energetically compatible with one structure becomes more compatible with another structure as the result of a single point mutation.
Figure 2: The log of the number of proteins (the number of nodes in the directed graph) as a function of in-degree (the total number of edges directed into a fold). The in-degree is an indicator of the stability of a particular shape and its ability to "steal" sequences from other structures.
Figure 3: Sequence retention at the energy E^* as a function of the contact density. For every fold, E^* is the energy at which the fraction of sequences retained by that fold is maximal. In our model, some proteins retain all sequences at E^* and all energy levels below. For other proteins, the fraction of retained sequences reaches a maximum at their E^*, and then falls again as energy is lowered. Some protein folds even have zero sequence retention rate throughout the energy landscape, meaning that they are almost entirely energetically dominated by other folds.
Figure 4: A contour plot of the number of strongly connected components in the graph as a function of the log of the cutoff value for establishing an edge (y axis) and as a function of the energy in the range E^* and E_{nat} (x axis). An edge is established when the fraction of sequences that flow between one protein to the next exceed a cutoff value. (A strongly connected component of a directed graph is a maximal set inside which every node has a path to every node).
Figure 5: The density of sequence capacity (without and with competition $\log\left[N(E_n)/20^L\right]$ and $\log\left[C(E_n)/20^L\right]$) as a function of the contact density (the total number of contacts divided by the number of amino acids).

Figure 1

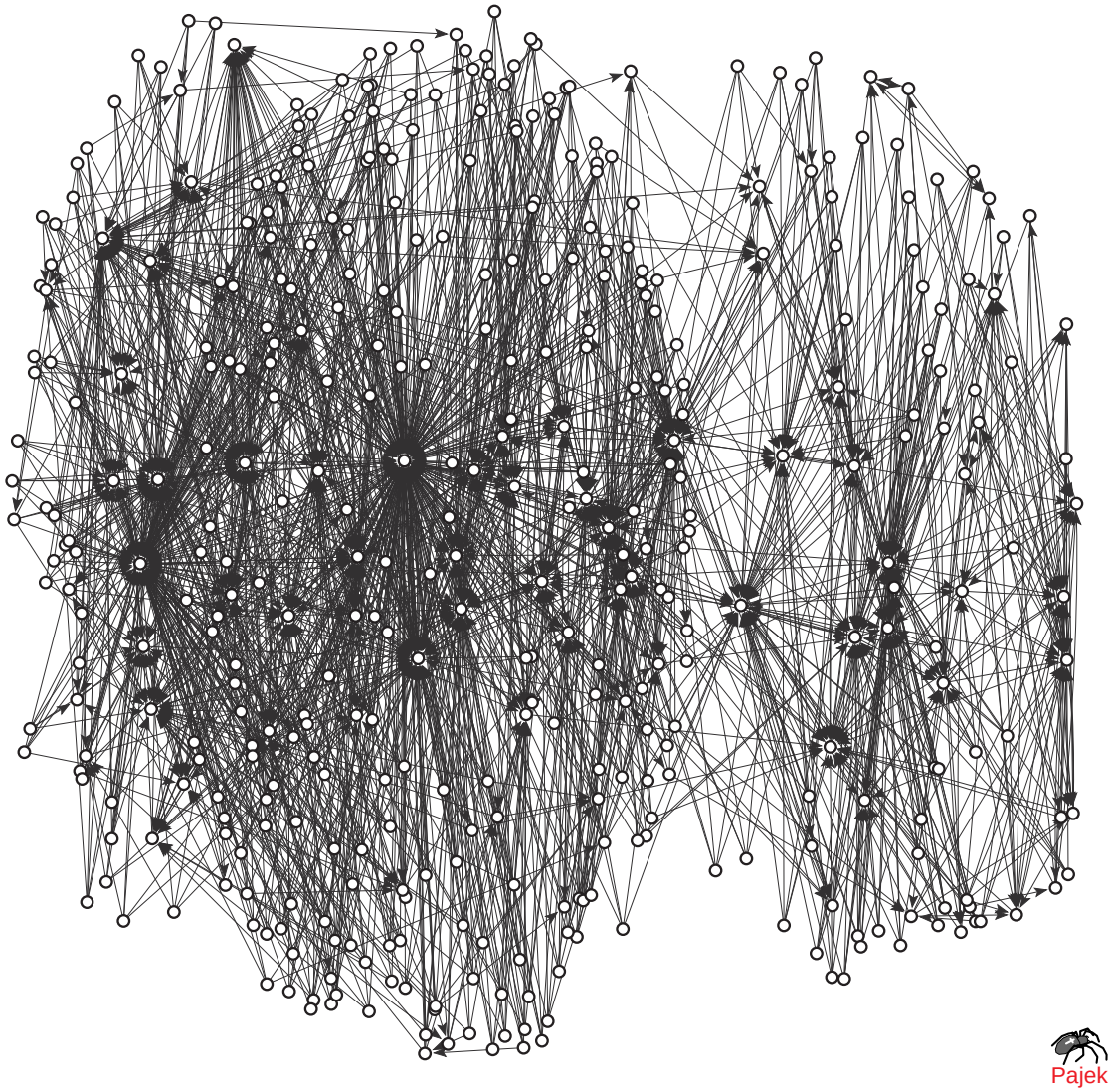


Figure 2

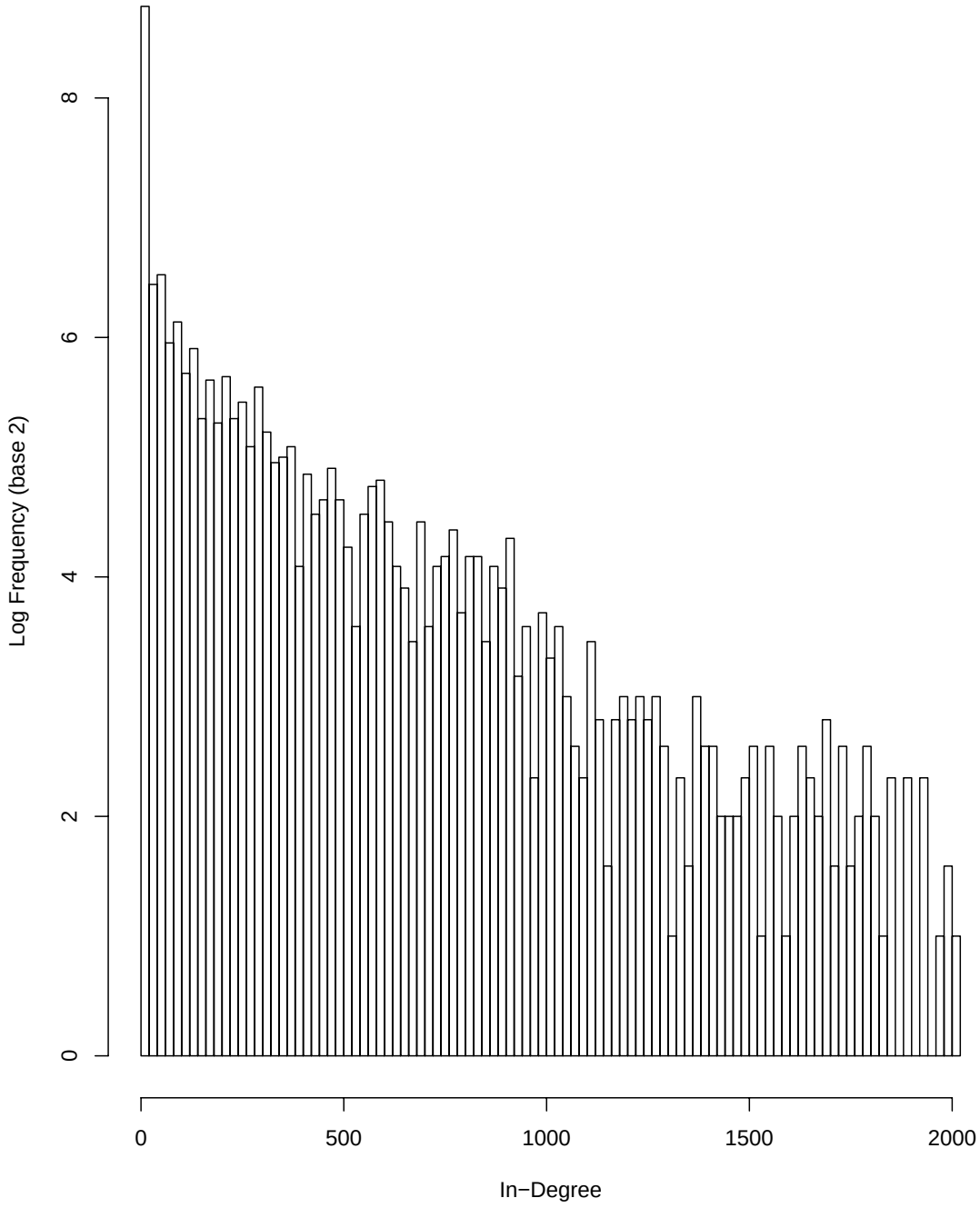


Figure 3

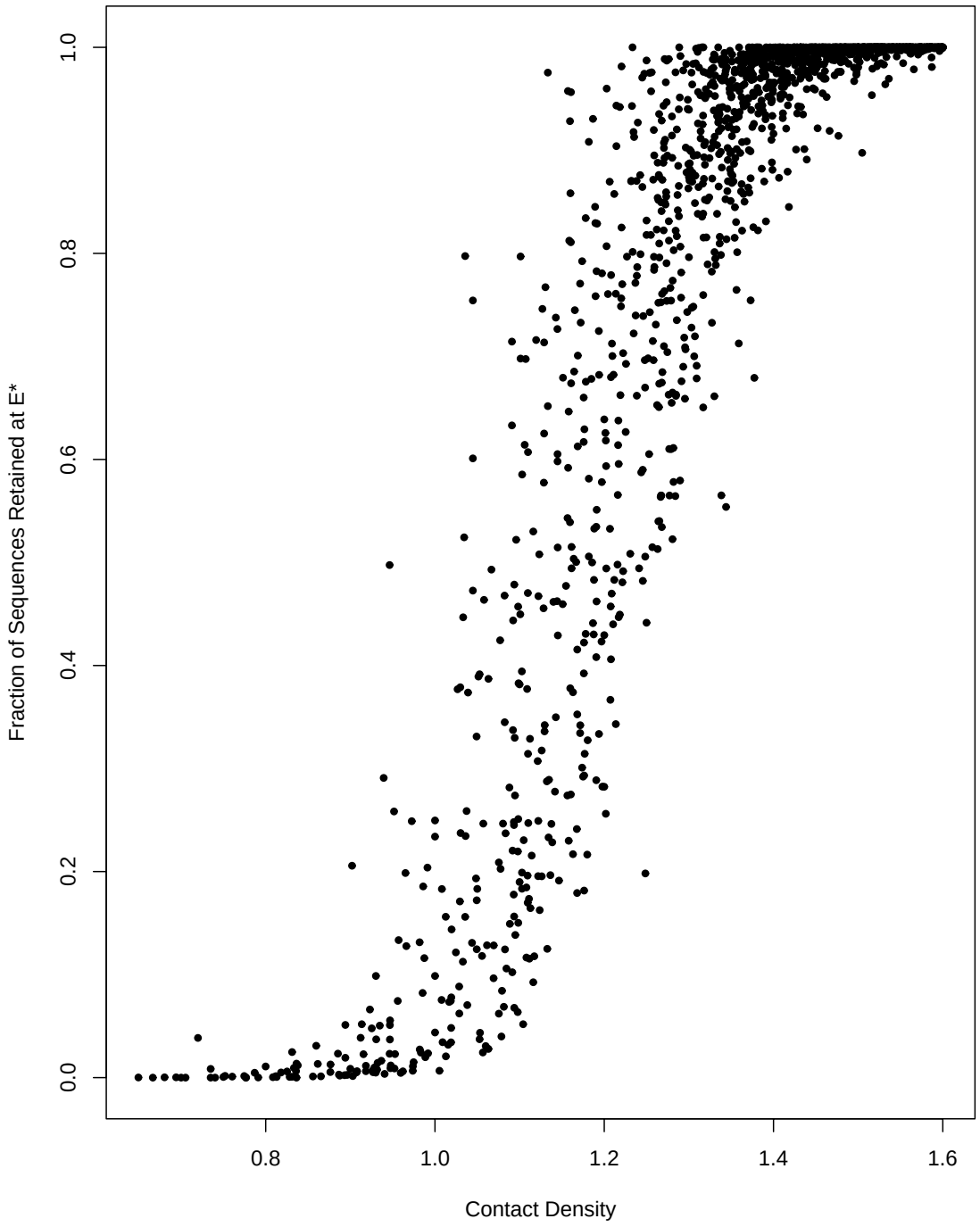

Fraction of Sequences Retained at E*
0.0
0.2
0.4
0.6
0.8
1.0
0.8
1.0
1.2
1.4
1.6
Contact Density

Figure 4

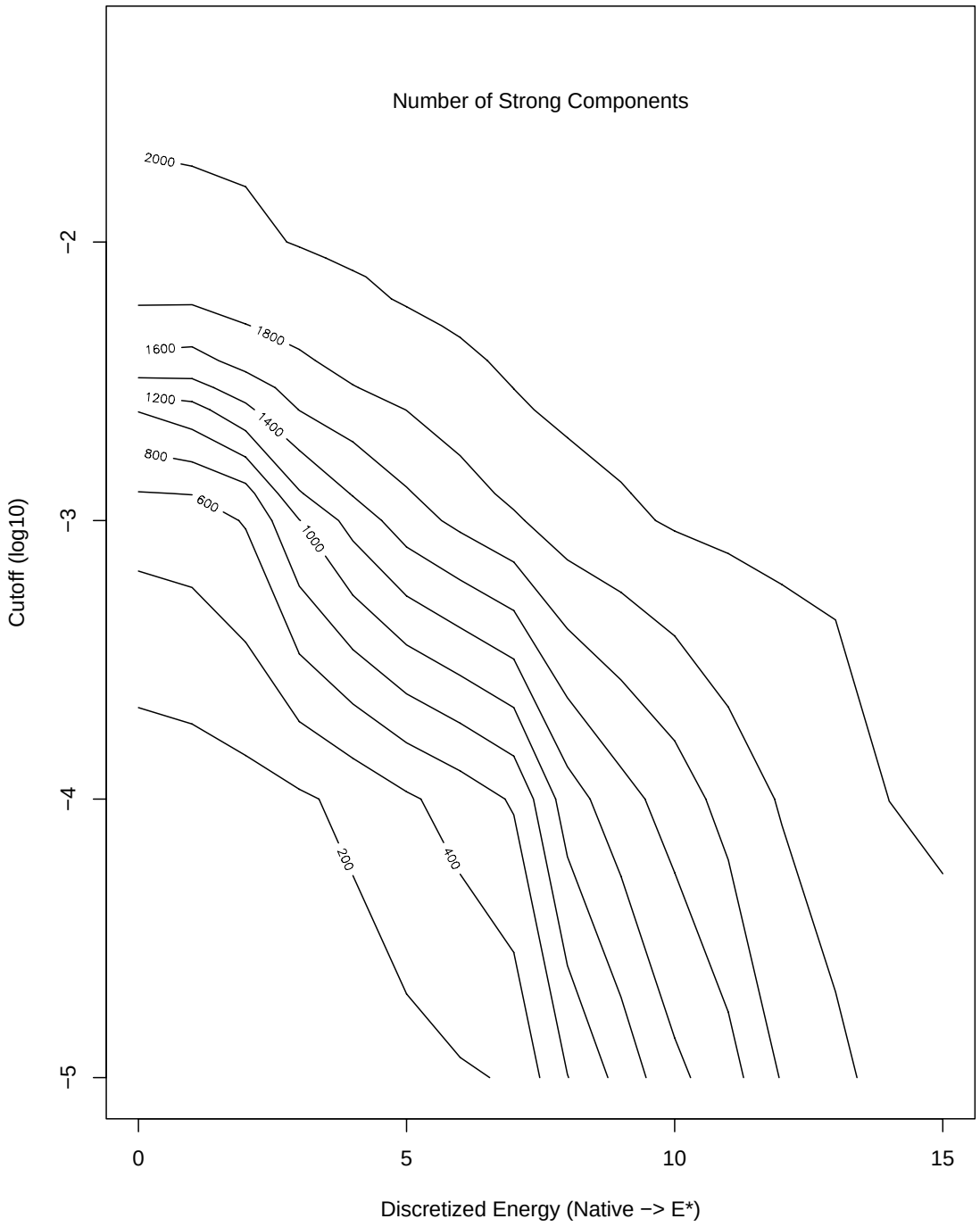
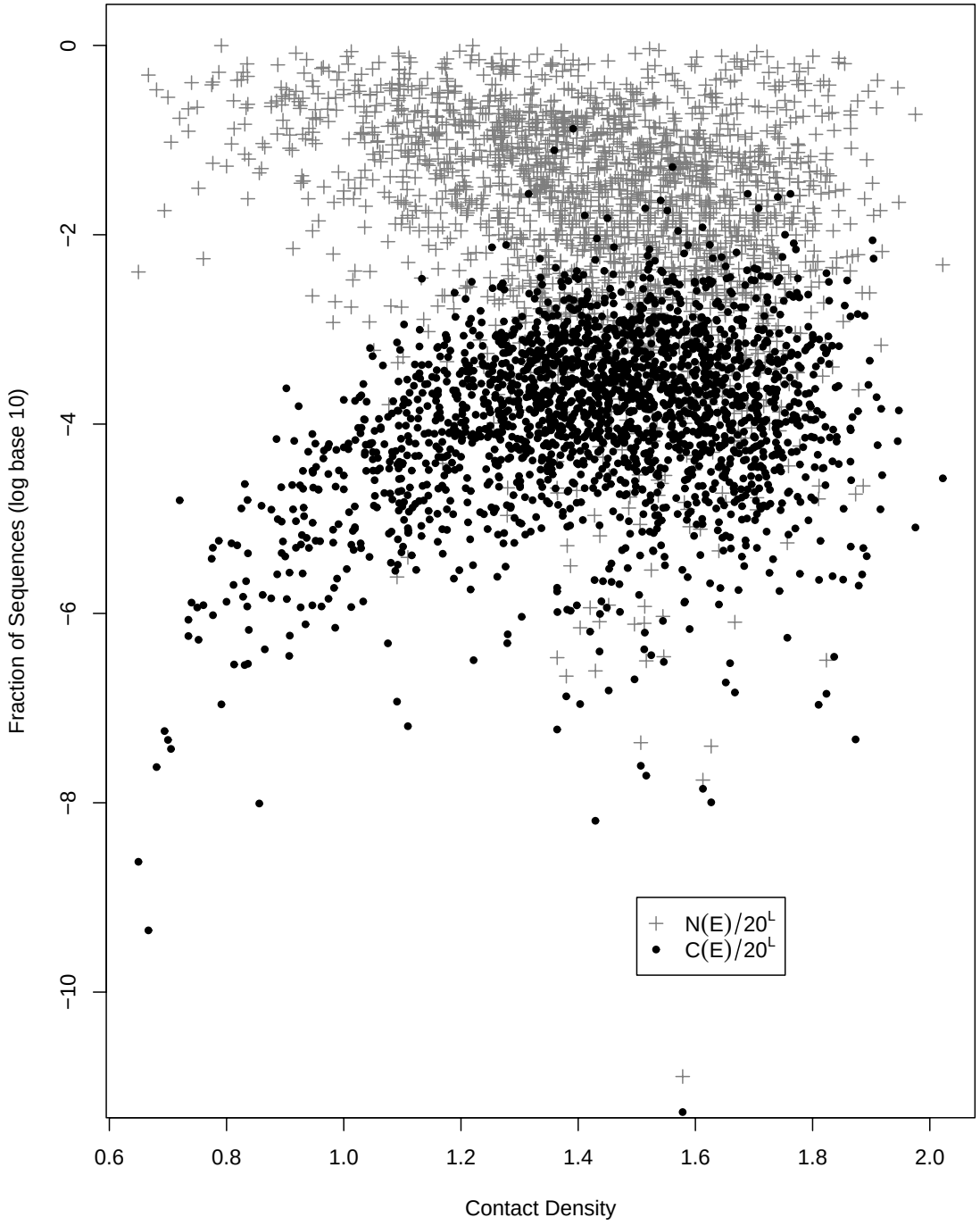

Number of Strong Components
2000
1800
1600
1400
1200
1000
800
600
400
200
−2
−3
−4
−5
Cutoff (log10)
0
5
10
15
Discretized Energy (Native −> E*)

Figure 5


Fraction of Sequences (log base 10)
0
−2
−4
−6
−8
−10
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
Contact Density
N(E)/20^L
C(E)/20^L

References

1. Saven, J. G. & Wolynes, P. G. (1997) *Journal of Physical Chemistry B* **101,** 8375-8389.
2. Shakhnovich, E. (1998) *Folding and Design* **3,** 45-58.
3. Betancourt, M. R. & Thirumalai, D. (2002) *Journal of Physical Chemistry* **106,** 599-609.
4. Lau, K. F. & Dill, K. (1990) *Proceeding of the National Academy of Science USA.* **87,** 638.
5. Shakhnovich, E. I. (1994) *Physical Review Letters* **72,** 3907-3911.
6. Govindarajan, S. & Goldstein, R. A. (1996) *Proceedings of the National Academy of Sciences of the United States of America* **93,** 3341-3345.
7. Li, H., Helling, R., Tang, C. & Wingreen, N. (1996) *Science* **273,** 666-669.
8. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. (2006) *Proceedings of the National Academy of Sciences of the United States of America* **103,** 5869-5874.
9. Xia, Y. & Levitt, M. (2004) *Proteins-Structure Function and Bioinformatics* **55,** 107-114.
10. Sun, S. J., Brem, R., Chan, H. S. & Dill, K. A. (1995) *Protein Engineering* **8,** 1205-1213.
11. Kleinberg, J. (1999) in *ACM RECOMB*, Vol. 3.
12. Park, S., Xi, Y. & Saven, J. G. (2004) *Current Opinion in Structural Biology* **14,** 487-494.
13. Saven, J. G. (2002) *Current Opinion in Structural Biology* **12,** 453.
14. Koehl, P. & Levitt, M. (2002) *Proceedings of the National Academy of Sciences of the United States of America* **99,** 1280-1285.
15. Larson, S. M., England, J. L., Desjarlais, J. R. & Pande, V. S. (2002) *Protein science* **11,** 2804-2813.
16. Bradley, P., Misura, K. M. S. & Baker, D. (2005) *Science* **309,** 1868-1871.
17. Berman, H. M., J., W., Z., F., G., G., T.N., B., H., W., I.N., S. & P.E., B. (2000) *Nucleic acids research* **28,** 235.
18. Pal, C., Papp, B. & Lercher, M. J. (2006) *Nature Reviews Genetics* **7,** 337-348.
19. Bloom, J. D., Drummond, D. A., Arnold, F. H. & Wilke, C. O. (2006) *Molecular Biology and Evolution* **23,** 1751-1761.
20. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. (2005) *Proceedings of the National Academy of Sciences of the United States of America* **102,** 14338-14343.
21. Meyerguz, L., Grasso, C., Kleinberg, J. & Elber, R. (2004) *Structure* **12,** 547-557.
22. Zhang, Y., Hubner, I. A., Arakaki, A. K., Shakhnovich, E. & Skolnick, J. (2006) *Proceedings of the National Academy of Sciences of the United States of America* **103,** 2605-2610.
23. Kihara, D. & Skolnick, J. (2003) *Journal of Molecular Biology* **334,** 793-802.
24. Regan, L. & Jackson, S. (2003) *Current Opinion in Structural Biology* **13,** 479-481.
25. Dalal, S., Balasubramanian, S. & Regan, L. (1997) *Nature Structural Biology* **4,** 548-552.

26. Ambroggio, X. I. & Kuhlman, B. (2006) *Current Opinion in Structural Biology* **16,** 525-530.
27. Cordes, M. H. J., Walsh, N. P., McKnight, C. J. & Sauer, R. T. (1999) *Science* **284,** 325-327.
28. Van Dorn, L. O., Newlove, T., Chang, S. M., Ingram, W. M. & Cordes, M. H. J. (2006) *Biochemistry* **45,** 10542-10553.
29. Alexander, P. A., Rozak, D. A., Orban, J. & Bryan, P. N. (2005) *Biochemistry* **44,** 14045-14054.
30. Anderson, T. A., Cordes, M. H. J. & Sauer, R. T. (2005) *Proceedings of the National Academy of Sciences of the United States of America* **102,** 18344-18349.
31. Grishin, N. V. (2001) *Journal of Structural Biology* **134,** 167-185.
32. Kinch, L. N. & Grishin, N. V. (2002) *Current Opinion in Structural Biology* **12,** 400-408.
33. Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. (2006) *Journal of Molecular Biology* **357,** 1335-1343.
34. Shakhnovich, B. E., Deeds, E., Delisi, C. & Shakhnovich, E. (2005) *Genome Research* **15,** 385-392.
35. Dokholyan, N. V., Shakhnovich, B. & Shakhnovich, E. I. (2002) *Proceedings of the National Academy of Sciences of the United States of America* **99,** 14132-14136.
36. Kim, P. M., Lu, L. J., Xia, Y. & Gerstein, M. (2006) *Science* **314,** 1938-1941.
37. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *Journal of Molecular Biology* **247,** 536-540.
38. Zhang, Y. & Skolnick, J. (2005) *Nucleic Acids Research* **33,** 2302-2309.
39. Meller, J. & Elber, R. (2002) in *Advances in chemical physics*, ed. Friesner, R. (John Wiley & Sons, Vol. 120, pp. 77-130.
40. Meller, J. & Elber, R. (2001) *Proteins, Structure, Function and Genetics* **45,** 241.
41. Durbin, R., Eddy, S., R.,, Krogh, A. & Mitchison, G. (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids* (Cambridge University Press, Cambridge UK).
42. Meyerguz, L., Kempe, D., Kleinberg, J. & Elber, R. (2004) in *RECOMB*.
43. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577-2637.
44. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. (2006) *Nucleic Acids Research* **34,** D16-D20.
45. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & al., e. (1990) *Journal of Molecular Biology* **215,** 403-410.
46. Henikoff, S. & Henikoff, J. G. (1992) *Proceeding of the National Academy of Science USA.* **89,** 10915-10919.
47. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1996) in *Recent developments in the theoretical studies of proteins*, ed. Elber, R. (World Scientific, Singapore), pp. 359-388.