Mutation-induced fold switching among lattice proteins

Christian Holzgräfe, Anders Irbäck, and Carl Troein

Citation: J. Chem. Phys. **135**, 195101 (2011); doi: 10.1063/1.3660691
View online: http://dx.doi.org/10.1063/1.3660691
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v135/i19
Published by the American Institute of Physics.

Mutation-induced fold switching among lattice proteins

Christian Holzgräfe,[a] Anders Irbäck,[b] and Carl Troein[c]
Computational Biology and Biological Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden

(Received 16 September 2011; accepted 26 October 2011; published online 18 November 2011)

Recent experiments uncovered a mutational pathway between two proteins, along which a single mutation causes a switch in fold. Searching for such paths between real proteins remains, despite this achievement, a true challenge. Here, we analyze fold switching in the minimalistic hydrophobic/polar model on a square lattice. For this analysis, we generate a comprehensive sequence-structure database for chains of length ≤ 30, which exceeds previous work by five units. Single-mutation-induced fold switching turns out to be quite common in the model. The switches define a fold network, whose topology is roughly similar to what one would expect for a set of randomly connected nodes. In the combinatorially challenging search for fold switches between two proteins, a tempting strategy is to only consider paths containing the minimum number of mutations. Such a restricted search fails to correctly identify 40% of the single-mutation-linked fold pairs that we observe. The thermodynamic stability is correlated with mutational stability and is, on average, markedly reduced at the observed fold switches. © *2011 American Institute of Physics*. [doi:10.1063/1.3660691]

I. INTRODUCTION

The traditional view that functional amino acid sequences adopt specific and mutationally robust three-dimensional structures provides a remarkably good starting point in rationalizing the complex sequence-structure relationship of proteins.[1] This picture is, nevertheless, known to be incomplete, as shown by the existence of intrinsically disordered proteins[2,3] and of proteins that can switch between alternative ordered states.[4,5] An example of a protein populating two distinct folds under physiological conditions is the chemokine lymphotactin.[6] A slightly different kind of fold switching is exemplified by two proteins from the Cro family of bacteriophage transcription factors, which have a high sequence identity (40%) but very different folds.[7]

A great deal has been learned about fold switching from stepwise mutagenesis experiments.[8–14] Of particular interest is the mutational pathway recently found between two binding domains of protein G, G_A and G_B, with 3α and $\alpha + \beta$ folds, respectively.[12–14] Along this pathway, an abrupt change in fold and binding properties takes place, caused by a single mutation. The stability is reduced near the switch point, but complete unfolding never occurs. The sequence determinants of this fold switch were analyzed by spectral methods.[15] Clearly, this system poses a challenge to structure prediction methods. Interestingly, the ROSETTA program was shown to capture the switch in fold when including chemical shifts as input (CS-ROSETTA).[16]

In this article, we study fold switching in the two-dimensional hydrophobic/polar (HP) lattice model.[17] This model is too coarse-grained to permit studies of specific proteins, but has provided insights into various generic aspects of protein folding.[18] This and similar models have, for instance, been used to study protein evolution[19–36] and the statistical properties of protein sequences.[37–39] One noteworthy finding from these studies is that the number of sequences that fold to a given structure correlates with their average thermodynamic stability.[22] A related but different result is that among sequences folding to the same structure, there is a correlation between thermodynamic and mutational stability.[26]

The HP model has the useful property that the sequence-to-structure mapping can be explored in an exact and complete manner for short chains, by means of exhaustive enumerations. Sequences that have a unique ground-state structure are called designing. In previous work, we determined all designing HP sequences of length $N \leq 25$.[40] Here, we extend these calculations to $N \leq 30$.

In addition to ground-state properties, we also determine the full density of states, $g(E)$, for all designing $N = 27$ sequences and for selected designing $N = 30$ sequences. Using these data, we assess the correlation between thermodynamic and mutational stability, and the loss of thermodynamic stability upon fold switching.

II. MODEL AND METHODS

A. Model and definitions

In the HP model,[17] a protein is represented by a self-avoiding chain of hydrophobic (H) or polar (P) beads on a lattice. The energy function is a contact potential. It can be expressed as

$$E = \sum_{i<j} U(\sigma_i, \sigma_j) C_{ij}, \tag{1}$$

[a]Electronic mail: christian.holzgraefe@thep.lu.se.
[b]Electronic mail: anders@thep.lu.se.
[c]Electronic mail: carl@thep.lu.se.

where σ_i denotes the type of bead i (H or P), $U(\sigma_i, \sigma_j)$ sets the interaction strengths, and C_{ij} is defined by

$$C_{ij} = \begin{cases} 1, & \text{if beads } i, j \text{ are neighbors on the lattice but } |i-j| \neq 1, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Two beads i and j with $C_{ij} = 1$ are said to form a contact. Following Lau and Dill,[17] we put $U(\sigma_i, \sigma_j) = -1$ if both σ_i and σ_j are H, and $U(\sigma_i, \sigma_j) = 0$ otherwise. With this choice, E is simply minus the number of HH contacts.

A sequence is *designing* if it possesses a unique minimum energy structure; it then designs that structure. The *designability* of a structure is the number of sequences designing it. A structure is *designable* if it has a non-zero designability.

In simplified protein models, it is customary to consider a mutation *neutral* if it preserves structure. The set of all sequences designing a given structure is referred to as the *neutral set*, which may be connected or fragmented with respect to single-point mutations. The largest connected component is called the *neutral net*. The *prototype sequence* of a neutral net is the sequence that can accommodate the largest number of neutral single-point mutations. This definition need not be unique. If not, we select the sequence having the lowest average Hamming distance to other sequences in the neutral net; the Hamming distance between two sequences of equal length is the number of positions where they differ.

Below, we examine the neutral nets of highly designable $N = 30$ structures. The above definitions of neutral net and prototype sequence are unambiguous for these structures.

Our analysis is performed on a square lattice, where the fraction of designing sequences is a few percent (see below). This fraction is significantly lower on a triangular lattice, where the chains are more flexible.[41]

B. Computational approach

We study this model by enumeration methods that rely on several optimizations, but are approximation-free. To simply go through all possible structures for each of the 2^N possible sequences is unfeasible for $N = 30$; the number of structures is $\sim 8 \times 10^{11}$ for this N (see below). The optimizations we use to overcome this problem are largely as described in our previous $N \leq 25$ study.[40] Somewhat related techniques were used in an $N \leq 20$ study.[42] In short, the structures are reduced to contact sets, which are further reduced through a breadth-first exploration of sequence space. In each step, a sequence position is set to H or P, the contact sets are reduced accordingly, and any redundant contact sets are eliminated essentially as described before.[40] To reduce memory requirements, we had to split the 2^{30} sequences into subsets and deal with each subset separately. With this approach, it was possible for us to determine all designing sequences and the corresponding structures for $N \leq 30$.

For all designing $N = 27$ sequences and for selected designing $N = 30$ sequences, we also determine the density of states, $g(E)$, using similar methods. The reduction of structures to contact sets is useful in this case as well, as the energy of a structure is determined by the contact set. To be able to compute $g(E)$ for all designing $N = 30$ sequences, further optimization would have been needed.

III. RESULTS

A. Sequence and structure statistics

We begin by discussing how the number of designing HP sequences, S_N, and the number of designable structures, D_N, depend on chain length, N. As far as we know, these quantities have not been determined before for $N > 25$. A complete listing of S_N, D_N, the total number of structures, and the number of contacts sets for $N \leq 30$ can be found in the supplementary material.[43]

Both D_N and S_N grow exponentially with N. Fig. 1 shows D_N along with the total number of structures and the number of contact sets, plotted against N. A fit shows that D_N scales with N as μ^N with $\mu \approx 1.86$. The total number of structures, that is the number of self-avoiding walks on a square lattice, is known to behave as $\sim N^{\gamma-1}\mu^N$ for large N, with $\gamma = 43/32$ and an effective coordination number of $\mu \approx 2.63$.[44] It follows that the fraction of all structures that are designable decreases rapidly with N. For $N = 30$, this fraction is $\sim 3 \times 10^{-6}$.

The number of contact sets grows slightly more slowly than the total number of structures with N; an exponential fit gives $\mu \approx 2.38$ in this case. This implies, in particular, that the computational gain brought by replacing structures with contact sets (see Sec. II B) gradually increases with N. For $N = 30$, the number of contact sets is ~ 200 times smaller than the number of structures.

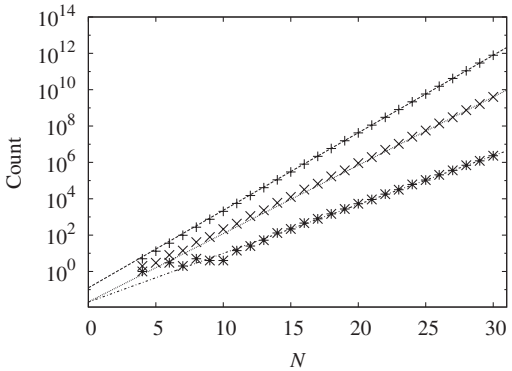

FIG. 1. The total number of structures (+), the number of contact sets (×), and the number of designable structures (*), D_N, against chain length, N. Lines are exponential fits to the data.

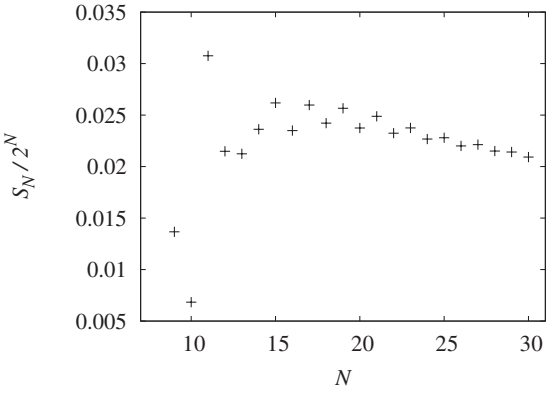

FIG. 2. The fraction of designing sequences, $S_N/2^N$, against chain length, N. The fact that this fraction is roughly constant implies that S_N grows slightly faster than D_N with N; as illustrated in Fig. 1, the D_N data scale as μ^N with $\mu \approx 1.86$.

Figure 2 shows the fraction of designing sequences, $S_N/2^N$, which, unlike the fraction of designable structures, is a slowly varying function of N. For $12 \leq N \leq 30$, $S_N/2^N$ lies in the range 2.1%–2.6%. However, the data show a decreasing trend, indicating that the same need not hold for larger N.

Among the designable structures, the spread in designability is large. The vast majority of the structures are designed by only a few sequences, whereas the maximum designability is 813 for $N \leq 30$. The average designability, S_N/D_N, increases slowly with N and is 9.63 for $N = 30$.

Figure 3 shows the maximally designable structures and the corresponding prototype sequences for $N = 26, \ldots, 30$. The division into a hydrophobic core and a polar surface is nearly perfect for these sequences. The structures are compact, but not maximally compact. For instance, the $N = 30$ structure contains 17 contacts, whereas the maximum number of contacts is 20 for this N. The structures contain, despite their high designability, 5–8 positions each that are strictly conserved. These positions are indicated by dashed circles. Most but not all of the conserved positions are of H type.

B. Highly designable structures and their neutral nets

Next, we study how highly designable structures and their neutral nets are distributed in structure and sequence space, respectively, focusing on chain length $N = 30$. We carried out the same analyses for $26 \leq N \leq 29$ as well, with qualitatively similar results which, for clarity, will not be discussed here. We consider all $N = 30$ structures with designability ≥ 389, which ensures that the prototype sequence is uniquely determined by our definition (see Sec. II A). This leaves us with 336 structures. Many of the associated neutral sets are fragmented, but they all have a single dominating component, the neutral net. The average size of the 336 neutral nets is 480.

To assess structural similarity, we use a contact-based Jaccard distance, J; two structures with contact sets A and B are assigned a distance of $J = 1 - |A \cap B|/|A \cup B|$. Figure 4(a) shows the distribution of J for pairs of highly designable $N = 30$ structures. The distribution has a tail extending to relatively small J, but the typical J is large. For comparison, Fig. 4(a) also shows the J distribution for general designable $N = 30$ structures, which is slightly shifted toward even higher J. This trend continues for pairs of random structures, which often do not share any contact at all. The J distribution is in this case dominated by its highest bin ($0.95 < J \leq 1$), which alone has a frequency of 0.69. For clarity, this distribution is not included in Fig. 4(a). Highly designable structures, thus, have an increased tendency to share contacts. Recurring substructure motifs[45] are a possible factor that could contribute to a shift of the J distribution in this direction.

We now turn to the neutral nets. As a measure of the separation between a pair of neutral nets, we determine the minimum Hamming distance between any two members of the two nets, called H. Figure 4(b) shows the distribution of H for pairs of highly designable $N = 30$ structures. The most common value is $H = 6$, but the distribution is broad. In particular, it can be seen that the minimal separation, $H = 1$, occurs with a non-negligible frequency (~0.8%).

Pairs of neutral nets with $H = 1$ are of special interest. For such a pair, it is possible to find a mutational path between the two prototype sequences that is entirely embedded within the two neutral nets. Such a path will be called *direct*. At some point along a direct path, a sudden change from one fold to the other occurs, caused by a single-point mutation.

Figure 5 illustrates how the highly designable $N = 30$ structures are interconnected by direct paths. In this graph, each vertex represents a neutral net and each edge indicates an $H = 1$ pair. The average degree of the network is $k_{\text{av}} = 2.62$ and, thus, well above 1. A random graph with this k_{av} is likely to contain a giant component. The network in Fig. 5 has indeed a single dominating component, comprising 282 of the 336 neutral nets. In addition, there are 4 clusters with 2 neutral nets each, and 46 isolated neutral nets (which are not shown in the figure).

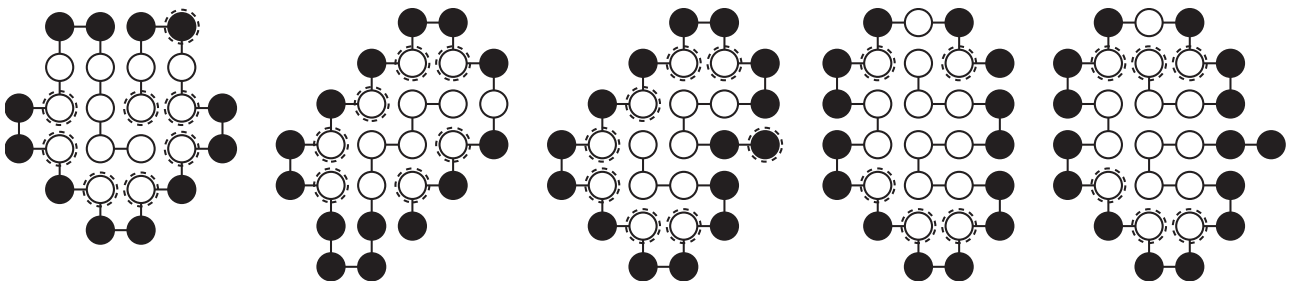

FIG. 3. Maximally designable structures for $N = 26$ (designability 341), 27 (430), 28 (498), 29 (804), and 30 (813). The sequences shown are prototype sequences; open and filled circles represent H and P beads, respectively. Dashed circles indicate strictly conserved positions.

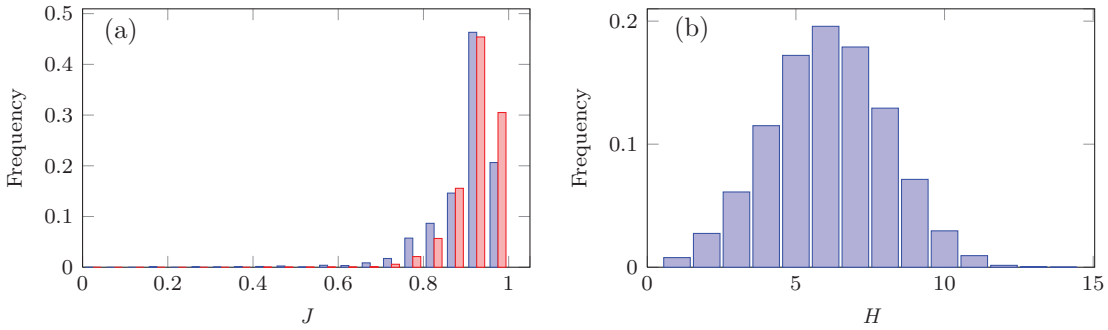

FIG. 4. (a) Histograms of the structural distance J for highly designable (blue) and general designable (red) $N = 30$ structures. (b) Histogram of the sequence-space separation H between neutral nets of highly designable $N = 30$ structures.

Figure 5 also shows the degree distribution of the graph, $P(k)$, along with the expected degree distribution for a (Bernoulli) random graph with the same mean, $k_{\text{av}} = 2.62$. The two distributions are roughly similar, although there are deviations at small k. A scale-free topology, as observed for protein-protein interaction networks,[46] is signaled by an enhanced tail of the degree distribution; the tail follows a power law, $k^{-\gamma}$, reflecting the presence of highly connected hub nodes. Our network, describing the interconnectedness of neutral nets, seems to be different in character. Over the range of k values observed in our finite network, there is no indication of an enhanced tail of this kind.

C. Minimal direct paths

The above analysis shows that for many pairs of highly designable $N = 30$ structures (~0.8%), it is possible to find a direct mutational path connecting the two prototype sequences. Whether or not such a path exists was here determined by an exhaustive search covering all possible paths. This option is not available for real proteins. A natural but drastic restriction is to search only among paths of the minimal length h, which is equal to the number of positions where the two given proteins differ. How likely is the set of all $h!$ minimal paths to contain a direct path, given that a direct path exists?

This problem can be studied for HP proteins. To this end, we consider again the highly designable $N = 30$ structures, this time focusing on those (>400) pairs of structures that are connected by direct paths and, thus, correspond to the edges in Fig. 5. For each such pair, we determine the length, L, of the shortest direct path between the prototype sequences, by using Dijkstra's algorithm.[47] We compare L with the shortest possible length of a path connecting these two sequences, which is given by their Hamming distance, h. The difference $\Delta L = L - h$ is always an even number, because any change beyond the minimal number needs to be undone.

Figure 6 shows our observed distribution of ΔL. It turns out that there exists a direct path of minimal length ($\Delta L = 0$) for 60% of these structure pairs. A search restricted to minimal paths, therefore, fails to find any of the direct paths known to exist in 40% of the cases ($\Delta L > 0$). Actually, in some cases, a quite elaborate search might be required in order to find any

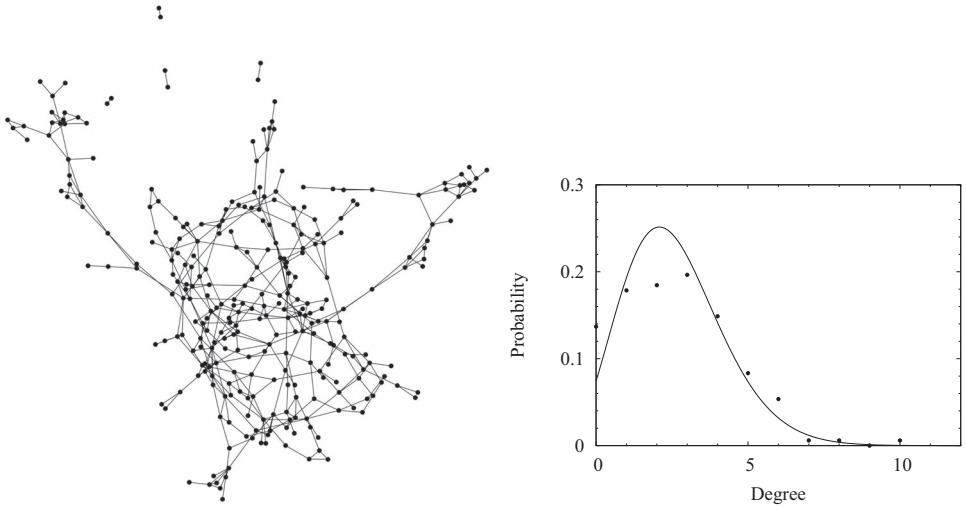

FIG. 5. (Left) Illustration of how the neutral nets associated with highly designable $N = 30$ structures are interconnected by direct paths. Each vertex represents a neutral net, and edges indicate the existence of direct paths ($H = 1$). Forty-six of a total of 336 neutral nets have a separation $H > 1$ to all the other neutral nets, and are omitted in the figure. (Right) The degree distribution, $P(k)$, of the graph to the left (including the isolated neutral nets). The line shows the expected Poissonian distribution for a random graph with the same mean, $k_{\text{av}} = 2.62$.

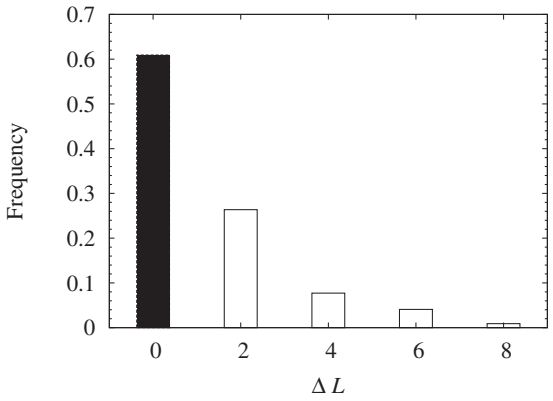

FIG. 6. Distribution of $\Delta L = L - h$ calculated over all pairs of highly designable $N = 30$ structures that are connected by direct paths (>400). A minimal-length direct path exists for 60% of the pairs ($\Delta L = 0$). L is the length of the shortest existing direct path between the prototype sequences, and h denotes their Hamming distance.

of the direct paths, as indicated by observed ΔL values of up to 8.

D. Stability

So far, we have classified sequences as either designing or not, thus ignoring differences in thermodynamic stability. Previous work on $N = 18$ HP chains found that among sequences designing the same structure, the thermodynamic stability is correlated with mutational stability.[26]

Here, we assess this correlation through a global analysis of all the $\sim 3.0 \times 10^6$ designing $N = 27$ sequences. The mutational stability of a (designing) sequence can be calculated as the number of single-point mutations that preserve the structure. As a measure of thermodynamic stability, we use the ground-state population $P_0(\beta)$ at a given inverse temperature $\beta = 1/k_B T$ (k_B is Boltzmann's constant). To compute $P_0(\beta)$, we determine the exact density of states, $g(E)$, for each sequence. Knowing $g(E)$, $P_0(\beta)$ can be obtained as

$$P_0(\beta) = \frac{g(E_0)e^{-\beta E_0}}{\sum_E g(E)e^{-\beta E}}, \tag{3}$$

where E_0 is the ground-state energy.

Figure 7 shows the mean and standard deviation of $P_0(\beta)$, at $\beta = 4$, as obtained at different fixed values of the mutational stability. Although the standard deviations are large, a very clear correlation can be seen between thermodynamic and mutational stability; the average $P_0(4)$ increases from ~ 0.2 at low mutational stability to >0.8 at high mutational stability. Figure 7 also shows the number of sequences with a given mutational stability. Consistent with the fact that most structures have low designability, we find that most sequences have low mutational stability. The number of sequences falls off approximately exponentially with increasing mutational stability in the range 5–20.

Consider now a direct mutational path between two prototype sequences. The mutational stability is, by definition, relatively high at the end points, and should be reduced near the switch point between the two folds. Being correlated with

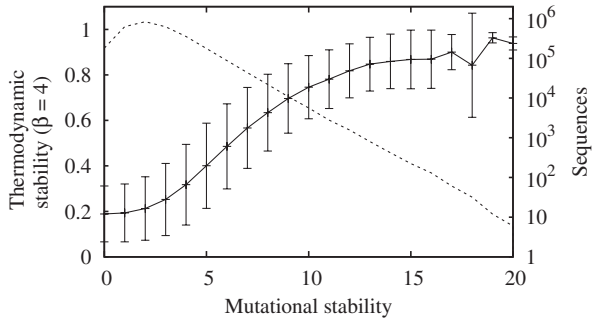

FIG. 7. Thermodynamic stability versus mutational stability, based on data for all designing $N = 27$ sequences. The figure shows the mean and standard deviation of the ground-state population $P_0(4)$ (see Eq. (3)), calculated at different fixed mutational stabilities. The dashed line indicates the number of sequences with a given mutational stability.

mutational stability, the thermodynamic stability should follow the same trend.

To test this picture, we examine how both stabilities vary along direct paths between highly designable $N = 30$ structures. For a given structure pair, we identify all direct paths that have the shortest length observed for that pair. In total, this gives us a set of 1392 direct paths, which we divide into subsets corresponding to different path lengths, L. For a given L, we compute the mean and standard deviation of both the ground-state population $P_0(4)$ and the mutational stability at all positions i along the paths, $i = 0, \ldots, L$. The results of this analysis are displayed in Fig. 8, where different curves correspond to different L. We limit ourselves to L values of 7–12, for each of which we have ≥ 100 paths. The calculated stabilities are plotted against the relative position i/L, a number between zero and one. When viewed as functions of i/L, the stability curves obtained for different L have a roughly

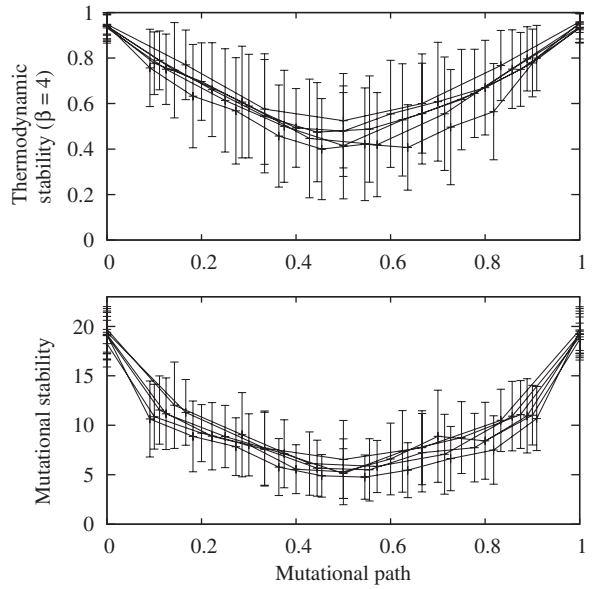

FIG. 8. Mean and standard deviation of the thermodynamic and mutational stabilities along direct mutational paths connecting highly designable $N = 30$ structures. Different curves correspond to different path lengths, $L = 7$–12. The x axis shows the relative position along the paths. The thermodynamic and mutational stabilities are defined as in Fig. 7.

similar shape. Although the path-to-path variation is large, as shown by the standard deviations, there is, on average, a clear drop in both thermodynamic and mutational stability in the central region between the prototype sequences. In particular, this strongly indicates that fold switching, indeed, is associated with a reduced thermodynamic stability.

To unambiguously demonstrate this point, we specifically locate all fold switches along these 1392 direct paths. The sequences involved in fold switches (a switch consists of two sequences) turn out to show a large variation in thermodynamic stability, from 0.024 to 0.949 (at $\beta = 4$). The average thermodynamic stability at the switch points is 0.438, which is similar to the values seen at the center of the paths in Fig. 8. This confirms that, statistically, the thermodynamic stability is markedly reduced at the switch points. The precise location of the switch points varies and need not be close to $i/L = 0.5$. Also worth noting is that the thermodynamic stability is not necessarily minimal at the switch point; we see many examples of direct paths ($\sim$400 of 1392) along which the minimum thermodynamic stability does not occur at the switch point.

IV. DISCUSSION

In this article, we have studied mutation-induced fold switching in the minimal HP model, where the sequence-to-structure mapping can be fully and exactly determined for short chains. This property opens the possibility to address questions out of reach in more detailed models about the sequence-to-structure relationship in general and fold switching in particular.

Our analysis mainly focused on the 336 most designable $N = 30$ structures, with designabilities in the range 389–813. The network describing the interconnectedness of the corresponding neutral nets was mapped out, and found to have $>$400 links ($\sim$0.8% of the pairs). The degree distribution of this network is roughly similar to what one would expect for a set of randomly connected nodes (Fig. 5).

The existence of a single-mutation link between two neutral nets implies that a direct mutational path can be found between their prototype sequences. For each of our $>$400 connected pairs of neutral nets, we determined the length L of the shortest path of this kind, which has to be greater than or equal to the Hamming distance h between the prototype sequences. The shortest direct path turned out to be of minimal length, $L = h$, for 60% of the pairs. But in 40% of the cases, one must, thus, search beyond the $h!$ minimal paths in order to find any of the direct paths known to exist.

Along a minimal mutational path, only one substitution can occur at each sequence position. For real proteins, this, in particular, means that, instead of 20, there are at most two possible amino acids at each position. The direct path recently discovered between the G_A and G_B proteins belongs to the class of minimal paths.[12–14] The path was found despite that even this restricted class of paths is far too large to be fully explored for real proteins.

The G_A/G_B experiments found that the thermodynamic stability is reduced at the switch point.[14] We computed the ground-state population $P_0(\beta = 4)$ at the switch points on 1392 direct paths between highly designable $N = 30$ structures. The obtained $P_0(4)$ values vary widely from path to path, between 0.024 and 0.949. Statistically, there is nevertheless a clear reduction in thermodynamic stability at the switch points; the average ground-state population $P_0(4)$ is 0.44 at the switch points, but $>$0.9 for the prototype sequences.

Bornberg-Bauer, Chan, and Wroe investigated the organization of sequences within a neutral net using $N = 18$ HP chains, and found the thermodynamic and mutational stabilities to be correlated.[26,32] A global analysis of all the $\sim 3 \times 10^6$ designing $N = 27$ sequences corroborates this finding; the average ground-state population, $P_0(4)$, increases from $\sim$0.2 at low mutational stability to $>$0.8 at high mutational stability. This analysis shows that there exists a clear correlation across different HP structures between thermodynamic and mutational stability.

A complete list of all designing $N \leq 30$ sequences and their structures, including the density of states of all the $N = 27$ sequences, will be made electronically available at http://cbbp.thep.lu.se/activities/hp/.

ACKNOWLEDGMENTS

This work was in part funded by the Swedish Research Council.

[1] A. V. Finkelstein and O. G. Ptitsyn, *Protein Physics: A Course of Lectures* (Academic, San Diego, 2002).
[2] H. J. Dyson and P. E. Wright, Nat. Rev. Mol. Cell Biol. **6**, 197 (2005).
[3] P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky, and A. K. Dunker, Biophys. J. **92**, 1439 (2007).
[4] A. G. Murzin, Science **320**, 1725 (2008).
[5] A. R. Davidson, Proc. Natl. Acad. Sci. U.S.A. **105**, 2759 (2008).
[6] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, and B. F. Volkman, Proc. Natl. Acad. Sci. U.S.A. **105**, 5057 (2008).
[7] C. G. Roessler, B. M. Hall, W. J. Anderson, W. M. Ingram, S. A. Roberts, W. R. Montfort, and M. H. J. Cordes, Proc. Natl. Acad. Sci. U.S.A. **105**, 2343 (2008).
[8] F. J. Blanco, I. Angrand, and L. Serrano, J. Mol. Biol. **285**, 741 (1999).
[9] S. Dalal and L. Regan, Protein Sci. **9**, 1651 (2000).
[10] T. A. Anderson, M. H. J. Cordes, and R. T. Sauer, Proc. Natl. Acad. Sci. U.S.A. **102**, 18344 (2005).
[11] X. I. Ambroggio and B. Kuhlman, Curr. Opin. Struct. Biol. **16**, 525 (2006).
[12] P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan, Proc. Natl. Acad. Sci. U.S.A. **104**, 11963 (2007).
[13] Y. He, Y. Chen, P. Alexander, P. N. Bryan, and J. Orban, Proc. Natl. Acad. Sci. U.S.A. **105**, 14412 (2008).
[14] P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan, Proc. Natl. Acad. Sci. U.S.A. **106**, 21149 (2009).
[15] S. Rackovsky, Phys. Rev. Lett. **106**, 248101 (2011).
[16] Y. Shen, P. N. Bryan, Y. He, J. Orban, D. Baker, and A. Bax, Protein Sci. **19**, 349 (2010).
[17] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[18] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).
[19] K. F. Lau and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **87**, 638 (1990).
[20] D. J. Lipman and W. J. Wilbur, Proc. R. Soc. London, Ser. B **245**, 7 (1991).
[21] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **93**, 839 (1996).
[22] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[23] E. Bornberg-Bauer, Biophys. J. **73**, 2393 (1997).
[24] S. Govindarajan and R. A. Goldstein, Proteins **29**, 461 (1997).
[25] U. Bastolla, H. E. Roman, and M. Vendruscolo, J. Theor. Biol. **200**, 49 (1999).
[26] E. Bornberg-Bauer and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **96**, 10689 (1999).
[27] B. Blackburne and J. D. Hirst, J. Chem. Phys. **115**, 1935 (2001).

[28]Y. Cui, W. H. Wong, E. Bornberg-Bauer, and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **99**, 809 (2002).
[29]Y. Xia and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **99**, 10382 (2002).
[30]Y. Xia and M. Levitt, Proteins **55**, 107 (2004).
[31]G. Tiana, B. E. Shakhnovich, N. V. Dokholyan, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **101**, 2846 (2004).
[32]R. Wroe, E. Bornberg-Bauer, and H. S. Chan, Biophys. J. **88**, 118 (2005).
[33]J. D. Bloom, J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami, and F. H. Arnold, Proc. Natl. Acad. Sci. U.S.A. **102**, 606 (2005).
[34]K. B. Zeldovich, P. Chen, B. E. Shakhnovich, and E. I. Shakhnovich, PLOS Comput. Biol. **3**, e139 (2007).
[35]R. Wroe, H. S. Chan, and E. Bornberg-Bauer, HFSP J. **1**, 79 (2007).
[36]T. Chen, D. Vernazobres, T. Yomo, E. Bornberg-Bauer, and H. S. Chan, Biophys. J. **98**, 2487 (2010).
[37]A. Irbäck and E. Sandelin, Biophys. J. **79**, 2252 (2000).
[38]E. N. Govorun, V. A. Ivanov, A. R. Khokhlov, P. G. Khalatur, A. L. Borovinsky, and A. Y. Grosberg, Phys. Rev. E **64**, 040903 (2001).
[39]E. Sandelin, Biophys. J. **86**, 23 (2004).
[40]A. Irbäck and C. Troein, J. Biol. Phys. **28**, 1 (2002).
[41]A. Irbäck and E. Sandelin, J. Chem. Phys. **108**, 2245 (1998).
[42]V. Shahrezaei, N. Hamedani, and M. R. Ejtehadi, Phys. Rev. E **60**, 4629 (1999).
[43]See supplementary material at http://dx.doi.org/10.1063/1.3660691 for a listing of D_N, S_N, the total number of structures, and the number of contact sets for $N \leq 30$.
[44]N. Madras and G. Slade, *The Self-Avoiding Walk* (Birkhäuser, Boston, 1993).
[45]T. Wang, J. Miller, N. S. Wingreen, C. Tang, and K. A. Dill, J. Chem. Phys. **113**, 8329 (2000).
[46]S. Maslov and K. Sneppen, Science **296**, 910 (2002).
[47]E. W. Dijkstra, Numer. Math. **1**, 269 (1959).