LU TP 03-44

Coupled Folding-Binding versus Docking: A Lattice Model Study

Nitin Gupta*

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur, Kanpur-208016, India

Anders Irbäck†

Complex Systems Division, Department of Theoretical Physics

Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden

(Dated: November 24, 2003)

Abstract

Using a simple hydrophobic/polar protein model, we perform a Monte Carlo study of the thermodynamics and kinetics of binding to a target structure for two closely related sequences, one of which has a unique folded state while the other is unstructured. We obtain significant differences in their binding behavior. The stable sequence has rigid docking as its preferred binding mode, while the unstructured chain tends to first attach to the target and then fold. The free-energy profiles associated with these two binding modes are compared.

*Electronic address: nitingpt@iitk.ac.in

†Electronic address: anders@thep.lu.se

I. INTRODUCTION

Protein structures are often viewed as well-defined static entities. For many proteins, this simplified static picture is accurate enough to provide valuable information about the function.[1] However, there are proteins that are wholly unstructured or only partially structured and yet functional. In fact, recent studies suggest that such proteins are more common than previously thought, especially in eukaryotic cells;[2,3] for instance, it has been estimated that as much as 17% of the proteins in *Drosophila* are wholly unstructured.[4] In many cases, intrinsically unstructured proteins adopt specific structures upon binding to their biological targets.[2,3] Folding and binding are then coupled, thus establishing a direct link between folding and function.

It has been suggested that unstructured proteins offer several advantages in cellular regulation.[2,3] Being unstructured might, for instance, allow one protein to interact with several targets. It might also be useful for control purposes due to rapid turnover. Furthermore, it has been argued that being unstructured might facilitate the binding of the protein to a target,[5] by increasing the 'capture radius'. This mechanism, termed 'fly casting',[5] was analyzed using a low-dimensional representation of the binding process. More recently, the thermodynamics of coupled folding-binding were examined using a generalized random energy model.[6]

Here we study coupled folding-binding by computer simulations of a simple chain-based model. Simulating coupled folding-binding of a chain in a controlled manner requires proper sampling of the full conformational space and is harder than simulating folding of an isolated chain, since rigid-body translations and rotations must be taken into account. The complexity of the problem makes it highly desirable to study coarse-grained models before entering high-resolution modeling. For the present study, we use the minimal two-dimensional hydrophobic/polar HP lattice model of Lau and Dill.[7] This model has been widely used to investigate basics of protein folding.[8] This and similar lattice-based models have also been used to gain insights into topics such as protein evolution,[9–11] prion-like conformational propagation[12,13] and protein aggregation.[14,15]

Our study is inspired by recent experiments by Wahlberg *et al.*[16] on the *in vitro* evolved protein Z_{SPA-1}. This sequence was engineered[17] from the Z domain of staphylococcal protein A, a well characterized three-helix-bundle protein.[18] It was selected for binding to the

Z domain itself. In the Z:Z_{SPA-1} complex, Z_{SPA-1} adopts a structure similar to the solution structure of the Z domain.[16,19] However, in solution, Z_{SPA-1} turns out to be structurally disordered.[16] The engineered Z_{SPA-1} thus exhibits coupled folding-binding. In Ref. 20, computer simulations of the solution behaviors of Z_{SPA-1} and the Z domain were performed, using a relatively detailed off-lattice model with 5–6 atoms per amino acid.[20] Simulating the binding behavior of an unstructured protein like Z_{SPA-1} at this level of resolution remains, however, a challenge.

The present study consists of two parts. First, we study binding statistics for very short HP chains with up to $N = 14$ monomers, to get an idea of how likely binding is to occur for stable and unstable sequences, respectively, in this model. We then perform a more detailed study of two $N = 25$ sequences. In particular, this study allows us to compare the free energy of coupled folding-binding with free-energy profiles for docking and for folding of an isolated chain.

II. MODEL AND METHODS

In the HP model,[7] the protein chain is represented by a string of hydrophobic (H) or polar (P) beads on the square lattice. Adjacent beads along the chain are connected by links of unit length. It is forbidden for two beads to occupy the same lattice site. Two beads that are neighbors on the lattice but non-adjacent along the chain are said to be in contact. The energy that a configuration gets is determined by the number of HH contacts, each HH contact being assigned an energy $\epsilon < 0$. This defines the model for a single chain in isolation.

Here we study single chains interacting with some fixed target structure. Two types of targets are considered. In Sec. III A, the target is an immobilized HP chain, and the interchain interactions are taken to be the same as the intra-chain interactions; that is, each HH contact between the target and the flexible chain is assigned the energy ϵ. Any cross HH contact is given this energy whether or not it is present in the final bound state. In Sec. III B, the target is an extended structure with one particular binding site. In this case, there is only one specific bead of the flexible chain that can interact with the binding site of the target. A contact between these two sites is taken to be favorable by an energy $\epsilon_b = 3\epsilon$.

To study the thermodynamics of these systems, we use Monte Carlo methods. The

moving chain is allowed four types of moves: local one- and two-bead moves, non-local pivot moves, and one-step translation moves (left, right, up or down). All moves are subject to a Metropolis accept/reject step, to ensure that detailed balance is fulfilled. One round of simulation, called one 'sweep', consists of (at most) $N-1$ one-bead moves, $N-2$ two-bead moves, one pivot move and one translation move, N being the number of beads.

In Sec. III B, in addition to the thermodynamic simulations, we also study the binding process as a function of Monte Carlo time. These calculations follow exactly the same protocol, except that the non-local pivot update is omitted, to avoid large unphysical deformations of the chain.

With its simplified conformational space and its minimal two-letter alphabet, the HP model is not meant for studies of specific proteins, but rather to shed light on general questions for generic sequences. It is worth noting[8] that if the principle of minimal frustration[21] holds, then the native structures of functional proteins should be strongly favored by the hydrophobicity pattern alone. This suggests that the HP model, despite its simplicity, might be able to capture non-trivial features of the mapping from sequence to native structure. A statistical analysis of HP model sequences with unique ground states lends support to this view; it turns out that the hydrophobic beads are anticorrelated along the chains,[22,23] which is the same behavior that real (globular) protein sequences show.[22,24] A recent study of the distribution of hydrophobicity in protein sequences can be found in Ref. 25. For a study of designed hydrophobic/polar copolymers with positive hydrophobicity correlations, see Refs. 26,27.

The HP model has the advantage over high-resolution models that exact results can be obtained for short chains, by exhaustive enumeration of all possible states. On the square lattice, all possible sequences with unique ground states (in isolation) have been determined for $N \leq 25$,[23] along with the corresponding structures. Below, we make use of these results. Sequences having a unique ground state will be referred to as stable.

III. RESULTS AND DISCUSSION

A. Statistics for Short Chains

To get an idea of how likely different binding behaviors are to occur in this model, we start by studying a large number of short sequences with up to $N = 14$ beads. Each sequence interacts with some immobilized target chain with the same length. All target sequences considered are stable and held fixed in their unique ground-state conformations. The two-chain system is contained in a box of size $2N \times 2N$. The target is fixed in the middle of the box so that the moving chain can attack it from any side. We say that the moving protein binds to the target if there is a unique minimum-energy configuration for the two-chain system. Whether or not this criterion is met can be determined with high confidence by Monte Carlo methods for these chain lengths. As an operational criterion for binding we require that the minimum-energy configuration is unique and is visited at least 10 times during the course of the simulation. To ensure that these visits to the bound state are 'independent', a visit is counted only when the system comes to this state after going to some state which had at least 20% higher energy than the bound state.

We first test all possible sequences with a given N for binding to one particular target conformation with that N, using $N = 10$, 11 and 12. As target, we somewhat arbitrarily pick the first entry in the list of all stable ground states obtained in Ref. 23.[28] The results of these calculations are summarized in Table I. From this table it can be seen that the fraction of binding sequences is a few per cent, which is comparable to the number of stable sequences. A clear majority of the binding sequences are unstable, which means that docking cannot be the mode of binding. That most binding sequences are unstable does not mean that such sequences have an intrinsically higher propensity to bind than stable ones, but rather it merely reflects the fact that most sequences are unstable.

As mentioned in the introduction, one suggested use of being unstructured is that it might enable the chain to bind to different targets. In the model, there are sequences showing this behavior. An example of this can be found in Fig. 1, which shows an unstructured $N = 12$ sequence that is able to bind to two different targets (both of which are held fixed). The structure that the unstable sequence adopts upon binding is seen to depend on the target.

Finally, restricting ourselves to stable sequences, we test for self-binding, by using the

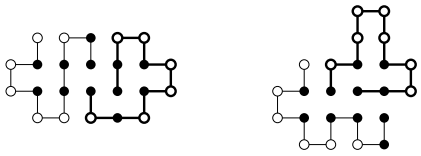

FIG. 1: Same sequence (HHPHHPPHPPHP) binding to two different targets. The moving sequence (thin lines) adopts different folds as it binds to the two fixed targets (thick lines).

same sequence as target and moving chain simultaneously. In obtaining Table I above, this calculation was done for those three sequences that served as targets. Using $N = 12$ and $N = 14$, we now consider all stable sequences. Table II shows the results of these calculations. We see that ∼10% of the stable sequences are self-binding both for $N = 12$ and $N = 14$. For $N = 12$, we observe that all the 9 self-binding stable sequences bind in their ground-state conformations. For $N = 14$, on the other hand, there exist 2 self-binding stable sequences for which the bound structure differs from the isolated one, while the remaining 40 keep the same structure after binding.

Let us briefly summarize the results presented so far. Our study of all possible sequences for $N = 10$, 11 and 12 shows that a significant fraction (a few per cent) can bind to the targets considered. Furthermore, most of these binding sequences are unstructured in isolation. Our study of all stable sequences for $N = 12$ and $N = 14$ shows that ∼10% of

TABLE I: Numbers of sequences for $N = 10$, 11 and 12 that can bind to the ground-state conformations of HHPPHPPHPH, HHPPHPPPPHP and HHPPHPPHPHPH, respectively. The stable sequences for these N are known from previous work.[23,29]

	$N = 10$	$N = 11$	$N = 12$
Total no. of sequences, 2^N	1024	2048	4096
Stable sequences[a]	6	62	87
Binding stable sequences[b]	0	6	18
Binding unstable sequences	20	44	140
Total no. of binding sequences	20	50	158

[a] A sequence is stable if it has a unique ground state in isolation.

[b] A sequence is binding if the system of moving chain plus target has a unique ground state.

them are self-binding. For most of these self-binding sequences, the complexed conformation is identical to the stable structure in isolation, indicating the possibility of a docking-like binding behavior.

B. Coupled Folding-Binding versus Docking

Having studied binding propensities of short chains, we now turn to a more detailed study of two longer chains with $N = 25$, whose binding behaviors are contrasted. This time the target is an extended structure located in the left-bottom end of an $N \times N$ box, in which the chain is confined (see Figs. 2 and 3). The target has only one binding site, and the binding is specific in that only one bead of the moving chain can interact with this binding site.

The two sequences considered are given in Table III. The first sequence, called S, was studied in isolation in Ref. 23. It was obtained by applying a sequence optimization algorithm[30] to 326 stable sequences all having the same unique ground-state conformation. It turns out that S has a ground-state energy of 13ϵ. The next most favorable conformations have only 11 HH contacts, so there is an energy gap of 2ϵ, making S unusually stable. The other sequence studied, called U, is unstable. Its degenerate ground state has an energy of 11ϵ. In our calculations, we observed five different conformations of U with this energy. Sequence U is a close analog of sequence S with just one hydrophobic bead mutated to a polar one. The two sequences were taken to be similar in order for their bound structure to be the same.

Our thermodynamic simulations of these two sequences were started from random configurations and contained 10^8 Monte Carlo sweeps each. The results show first of all that

TABLE II: Numbers of stable sequences that can bind to themselves for $N = 12$ and $N = 14$. The total numbers of stable sequences are from previous work.[23,29] The last row gives the numbers of self-binding stable sequences that bind in their ground-state conformations.

	$N = 12$	$N = 14$
Total no. of sequences	4096	16384
Stable sequences	87	386
Self-binding stable sequences	9	42
Self-binding stable sequences with unchanged structure	9	40

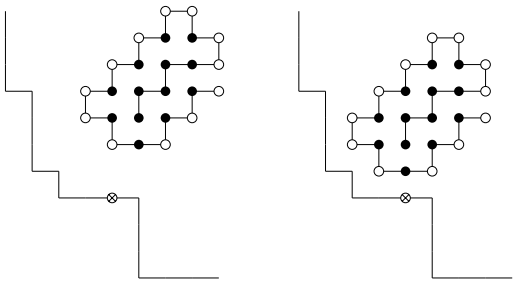

FIG. 2: Typical binding behavior of sequence S. The chain first folds and then just translates to bind to the target. Filled and open circles represent hydrophobic and polar beads, respectively, and the circle with a cross represents the binding site.

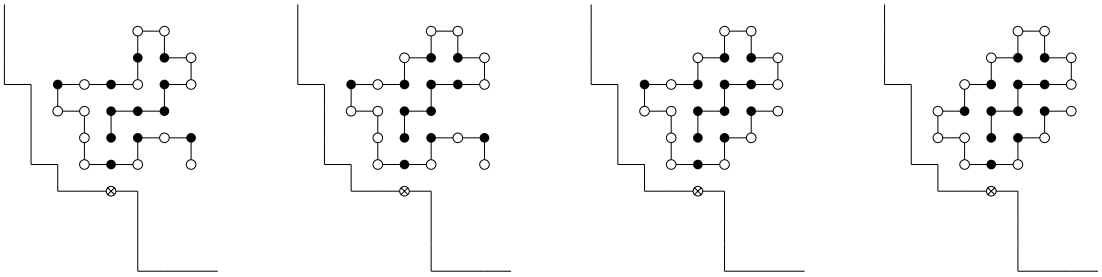

FIG. 3: Snapshots from a simulation of sequence U. After binding to the target, the chain rearranges itself into the minimum-energy state (right-most image). Symbols are as in Fig. 2.

both these sequences do bind to the chosen target, in the sense that the system of target plus moving chain has a unique minimum-energy state. This state, referred to as the bound state, was visited many independent times in the simulations. The bound structure is the same for both chains (see Figs. 2 and 3), and coincides with the stable structure of S in isolation. The energy of the bound state is $13\epsilon + \epsilon_{\rm b}$ for S and $11\epsilon + \epsilon_{\rm b}$ for U.

To characterize the binding behaviors of these sequences, we monitor the following two

TABLE III: The two $N = 25$ sequences studied, S and U. The position at which they differ is underlined. The bead that can interact with the target is in bold type.

S:	HHHHH PPHPP HPHPH PP<u>H</u>P**H** PHPHP
U:	HHHHH PPHPP HPHPH PP<u>P</u>P**H** PHPHP

quantities:

1. The binding parameter I. I is the geometric distance between the binding site of the target and that bead of the moving chain that can interact with the target. As binding progresses, the value of I reduces.

2. The folding parameter Q'. The chains studied form 14 internal contacts, called native contacts, in their bound states. The number of these native contacts being present, Q, provides a measure of the 'nativeness' of the chain. As the chain folds to its bound-state structure, the value of Q increases. It turns out that the value $Q = 13$ is impossible to attain for these sequences. In our free-energy calculations, we therefore use a folding parameter Q' defined by

$$Q' = \begin{cases} Q & \text{if } Q \leq 12 \\ 13 & \text{if } Q = 14 \end{cases} \tag{1}$$

Figures 4 and 5 show the free energy calculated as a function of these two variables, $F(I, Q')$, for the sequences S and U, respectively. The free energy is defined by $P(I, Q') = \exp[-F(I, Q')/kT]$, where $P(I, Q')$ is the joint I, Q' probability distribution. The temperature is taken as $T = \epsilon/2.6k$ for S and $T = \epsilon/3.2k$ for U (k is Boltzmann's constant). These temperatures are chosen such that the binding probabilities for the two chains are significant and close to each other ($\approx 35\,\%$). Note that the systems are in their minimum-energy states if and only if $(I, Q') = (0, 13)$.

For sequence S (Fig. 4), we find that $F(I, Q')$ has a simple shape with two narrow valleys along the lines $Q' = 13$ and $I = 0$, respectively, and a broad and shallow minimum centered at $(I, Q') \approx (11, 5)$, where the chain is unbound and unfolded. This suggests that there are two very different major binding modes for this sequence. One way for the chain to reach its bound state is along the line $Q' = 13$, which corresponds to rigid docking; the chain first folds and then moves towards the binding site. The other major binding mode is along the $I = 0$ valley. Here the chain first attaches to the binding site and then folds to its final shape. Following Ref. 5, we refer to this behavior as the fly-casting mechanism.

Let P_{d} and P_{f} denote the probabilities of finding the system in the $Q' = 13$ and $I = 0$ corridors, respectively (not counting the bound state); that is,

$$P_{\mathrm{d}} = \sum_{I>0} P(I, Q' = 13) \qquad P_{\mathrm{f}} = \sum_{Q'=0}^{12} P(I = 0, Q') \tag{2}$$

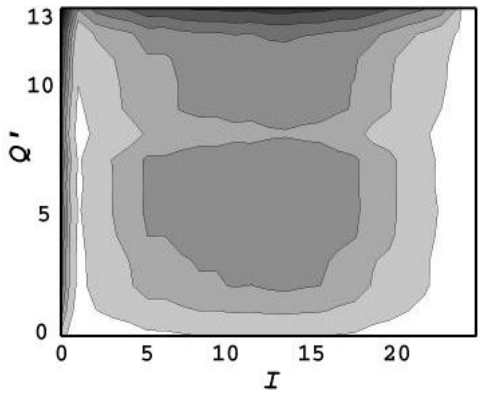

FIG. 4: Free energy $F(I, Q')$ for sequence S at $T = \epsilon/2.6k$. The bound state, in which the chain is both folded and attached to the target, corresponds to the upper left corner of the figure. A chain that is folded but not attached to the target moves along the upper edge of the figure (docking), whereas a chain that is attached to the target but incompletely folded moves along the left-hand edge (fly casting). The contours are spaced at intervals of $1\,kT$ and dark tone corresponds to low free energy. Contours more than $8\,kT$ above the minimum free energy are not shown.

where the subscripts d and f refer to docking and fly casting. For sequence S, we find that $P_{\rm d}$ is about twice as large as $P_{\rm f}$ (see Table IV), suggesting that docking is the preferred binding mode for this sequence.

For the unstructured sequence U (Fig. 5), there is no free-energy valley corresponding to rigid docking, so $P_{\rm d}$ is small (see Table IV). There is, by contrast, an $I = 0$ valley corresponding to fly casting for this sequence, too. The population $P_{\rm f}$ is, in fact, higher for U than for S (see Table IV). The $I = 0$ valley is, for both sequences, separated from rest of the conformational space by a free-energy ridge. For sequence U, there are some

TABLE IV: The probabilities $P_{\rm d}$ and $P_{\rm f}$ as defined by Eq. (2)[a] along with the binding probability $P_{\rm b} = P(I = 0, Q' = 13)$, for the sequences S and U.

	S	U
$P_{\rm b}$	0.35	0.35
$P_{\rm d}$	0.32	0.05
$P_{\rm f}$	0.17	0.23

[a] The subscripts d and f refer to docking and fly casting.

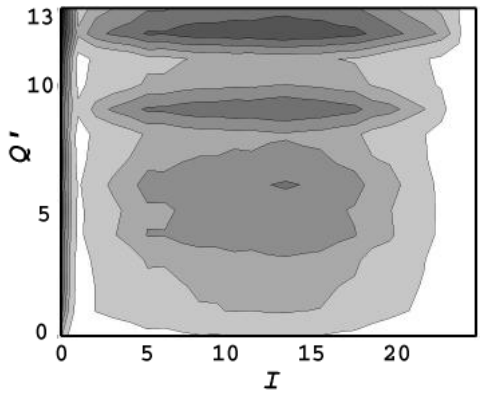

FIG. 5: Same as Fig. 4 for sequence U at $T = \epsilon/3.2k$.

narrow troughs in these hills towards the $I = 0$ corridor, which should make it easier for this sequence to reach this corridor. In addition to the $I = 0$ valley, sequence U exhibits two somewhat less pronounced valleys along the lines $Q' = 9$ and $Q' = 12$, respectively, as well as a broad unbound and unfolded minimum at lower Q'. Note that there exist $Q' = 12$ conformations with minimal intra-chain energy for this sequence. The presence of the $Q' = 9$ and $Q' = 12$ valleys suggests that this chain can follow several different paths to its bound state. If the chain follows one of the fixed-Q' valleys, it reaches the binding site without having its full bound-state structure; folding is completed after the chain has attached itself to the target. When increasing the box size to $3N \times 3N$, we observed valleys at the same values of Q' which were stretched along I.

It is instructive to take a closer look at the free energy in the docking and fly-casting corridors, respectively. Figure 6a shows the free energy along the docking corridor, $F(I, Q' = 13)$, for sequence S. We see that in order to reach the bound state, the chain has to pass a free-energy barrier with a height of $\approx 3\,kT$. This barrier is entropic. Figure 6b shows the free energy along the fly-casting corridor, $F(I = 0, Q')$, for both sequences. Two features are worth noting. First, the shape of the curve is roughly the same for the two sequences, although the curve is shifted downwards for sequence U which spends more time in this corridor. Second, there is no major free-energy barrier in this corridor (the highest barrier is about $1\,kT$), and the last part towards the bound state (from $Q' = 8$–9 to $Q' = 13$) is downhill in free energy.

It is also interesting to compare the free energy in the fly-casting corridor (Fig. 6b) with the free energy of folding for sequence S in isolation. For this purpose, we performed a sim-

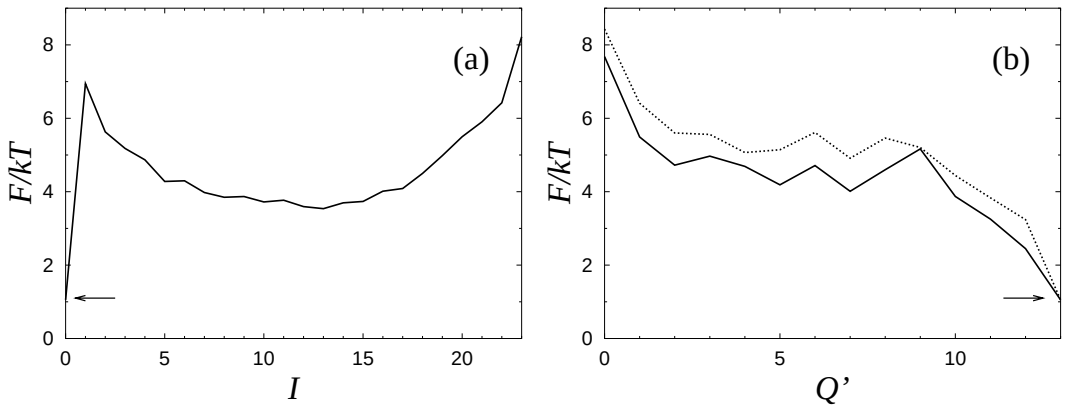

FIG. 6: Free-energy profiles for the docking and fly-casting corridors. (a) $F(I, Q' = 13)$ for sequence S. (b) $F(I = 0, Q')$ for sequence S (dotted line) and sequence U (solid line). The arrows indicate the location of the bound state, in which the chains are both folded and attached to the target. The temperature is $\epsilon/2.6k$ for S and $\epsilon/3.2k$ for U (same as in Figs. 4 and 5, respectively).

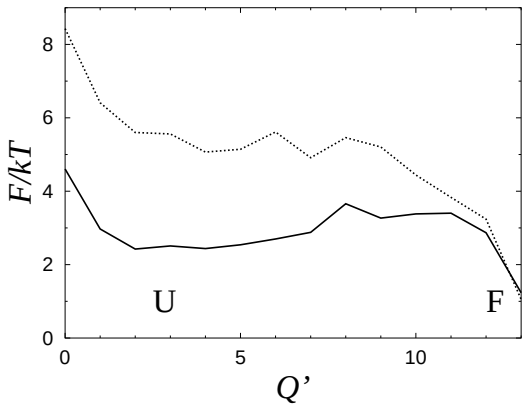

FIG. 7: Two free-energy profiles for sequence S. The dashed line shows the free energy along the fly-casting corridor, $F(I = 0, Q')$, at $T = \epsilon/2.6k$ (same as in Fig. 6b). The full line shows $F(Q')$ for the isolated chain at $T = \epsilon/2.268k$. U and F refer to unfolded and folded conformations, respectively.

ulation of S in isolation at $T = \epsilon/2.268k$, where the native population (29%) is comparable to the binding probability at $T = \epsilon/2.6k$ (35%). From Fig. 7, it can be seen that the free energy for the isolated chain is markedly different from that for the fly-casting corridor. In particular, we see that the bound-state minimum of the interacting chain is broader than the native minimum of the isolated chain; the isolated chain must reach $Q' = 11$ before the free energy starts to decrease.

We also performed Monte Carlo-based kinetic simulations (see Sec. II) of the sequences S and U, using the same temperatures as in the thermodynamic runs. For each sequence, we carried out a set of 500 simulations, each containing 5×10^6 Monte Carlo sweeps. The simulations were started from configurations obtained by short runs at a higher temperature ($T = \epsilon/2k$). The resulting configurations from these preparatory runs (the starting configurations for the kinetic runs) showed a wide variation in both location and conformation for the moving chain. The averages of I and Q over this ensemble were about 12 and 4–5, respectively.

Figure 8 shows the evolution of this ensemble of 500 systems with Monte Carlo time in the (I, Q) plane (not Q'). This parametric plot, with Monte Carlo time as a parameter, has the advantage of being independent of the difference in time scale between the two different temperatures used. To reduce noise, each data point represents an average over 10^4 Monte Carlo sweeps, thus giving 500 data points in total. In Fig. 8, we also indicate the equilibrium values of I and Q, which were obtained by a separate, very long simulation. The equilibrium values are given by $(\langle I\rangle, \langle Q\rangle) = (6.1, 12.1)$ for S and $(\langle I\rangle, \langle Q\rangle) = (4.9, 11.1)$ for U. Note that, as expected, S has a higher value of $\langle Q\rangle$ and U has a lower value of $\langle I\rangle$.

For sequence S, we see that the relaxation towards the equilibrium point has a clear two-step character, where the first step corresponds to folding (increasing Q) and the second step to binding (decreasing I). This behavior is indeed what one expects if docking is the preferred binding mode. For sequence U, there is no such clear distinction between folding and binding; I and Q evolve in a more correlated manner. Hence, the results of the kinetic simulations support the conclusions from the free-energy analysis.

IV. SUMMARY

We have studied different binding mechanisms of polypeptide chains using the minimal HP model. The aim of this study is to understand the nature of an important aspect of protein-protein recognition and it can serve as a guideline for further, detailed studies of this complex phenomenon.

For small chain lengths, we found that the fraction of sequences that can bind to a given (fixed) target is comparable to the fraction of sequences that are stable in isolation. The overlap between these two sets of sequences is small; most binding sequences are unstable

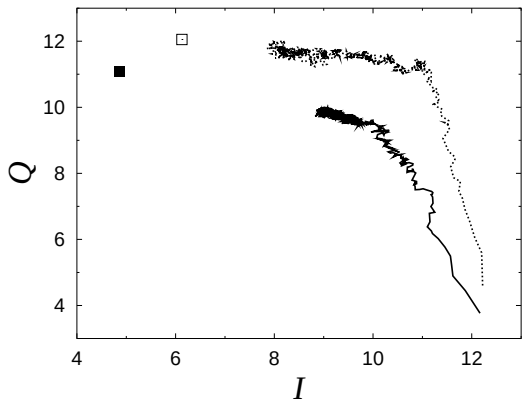

FIG. 8: Parametric plot of Q against I with Monte Carlo time as a parameter for sequence S (dotted line) and sequence U (solid line). The curves represent averages over 500 simulations and the bottom-right ends correspond to time zero. The unfilled and filled squares represent the equilibrium points ($\langle Q\rangle$,$\langle I\rangle$) for S and U, respectively. Temperatures are as in Figs. 4 and 5.

in isolation and fold upon binding to the target.

We then compared the binding behaviors of two related $N = 25$ sequences. One of the sequences, S, had been optimized for high stability in isolation,[23] and was found to have rigid docking as its preferred binding mode. The other, slightly mutated sequence, U, was found to have a bound structure identical to that of S, but to be unstructured in isolation. In these respects, sequence U is reminiscent of the engineered protein Z_{SPA-1} mentioned in the introduction. It turned out that sequence U tends to first attach to the target and then fold, which supports the conclusion[5] that unstructured chains prefer binding through a fly-casting mechanism. The free-energy profile associated with this binding mode was found to lack high barriers and to exhibit a broad bound-state minimum, compared with the native minimum of sequence S in isolation. The observed difference in shape between these free energies suggests that by attaching itself to the target, the chain becomes able to fold more efficiently than it does in isolation. The free energy of docking obtained for sequence S contains, as expected, a significant entropic barrier, which the chain has to overcome in order to reach the bound state.

These free-energy profiles were obtained using two particular sequences in this model. Therefore, it should be stressed that the shape of the free energy was not considered at all when selecting these two sequences and the target structure; the goal was just to have two binding sequences with the same bound structure but different stability properties in

isolation. As a result, we believe that the trends seen have some generality. Nevertheless, it is clear that it would be highly desirable to extend these calculations to other and longer sequences and to more realistic off-lattice models. Furthermore, it would be interesting to study the kinetics over a larger time interval; this requires a larger number of systems than 500 which we used, in order to keep the statistical errors under control.

Acknowledgments

We thank Björn Samuelsson for valuable discussions and help with figures. This work was in part supported by the Swedish Research Council.

[1] C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland Publishing, New York, 1991).

[2] P.E. Wright and H.J. Dyson, *J. Mol. Biol.* **293** 321 (1999).

[3] H.J. Dyson and P.E. Wright, *Curr. Opin. Struct. Biol.* **12**, 54 (2002).

[4] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner and Z. Obradovic, *J. Mol. Graphics Modell.* **19**, 26 (2001).

[5] B.A. Shoemaker, J.J. Portman and P.G. Wolynes, *Proc. Natl. Acad. Sci. USA* **97**, 8868 (2000).

[6] G.A. Papoian and P.G. Wolynes, *Biopolymers* **68**, 333 (2003).

[7] K.F. Lau and K.A. Dill, *Macromolecules* **22**, 3986 (1989).

[8] For a recent review, see H.S. Chan, H. Kaya and S. Shimizu, in *Current topics in Computational Biology*, edited by T. Jiang, Y. Xu and M.Q. Zhang (MIT Press, Cambridge, Massachusetts, 2002), pp. 403–447.

[9] E. Bornberg-Bauer, *Biophys. J.* **73**, 2393 (1997).

[10] Y. Cui, W.H. Wong, E. Bornberg-Bauer and H.S. Chan, *Proc. Natl. Acad. Sci. USA* **99**, 809 (2002).

[11] B.P. Blackburne and J.D. Hirst, *J. Chem. Phys.* **119**, 3453 (2003).

[12] P.M. Harrison, H.S. Chan, S.B. Prusiner and F.E. Cohen, *J. Mol. Biol.* **286**, 593 (1999).

[13] P.M. Harrison, H.S. Chan, S.B. Prusiner and F.E. Cohen, *Protein Sci.* **10**, 819 (2001).

[14] R.A. Broglia, G. Tiana, S. Pasquali, H.E. Roman and E. Vigezzi, *Proc. Natl. Acad. Sci. USA* **95**, 12930 (1998). Erratum: **96**, 10943 (1999).

[15] S. Istrail, R. Schwartz and J. King, *J. Comp. Biol.* **6**, 143 (1999).

[16] E. Wahlberg, C. Lendel, M. Helgstrand, P. Allard, V. Dincbas-Renqvist, A. Hedqvist, H. Berglund, P.-Å. Nygren and T. Härd, *Proc. Natl. Acad. Sci. USA* **100**, 3185 (2003).

[17] M. Eklund, L. Axelsson, M. Uhlén and P.-Å. Nygren, *Proteins* **48**, 454 (2002).

[18] M. Tashiro, R. Tejero, D.E. Zimmerman, B. Celda, B. Nilsson and G.T. Montelione, *J. Mol. Biol.* **272**, 573 (1997).

[19] M. Högbom, M. Eklund, P.-Å. Nygren and P. Nordlund, *Proc. Natl. Acad. Sci. USA* **100**, 3191

(2003).

[20] G. Favrin, A. Irbäck and S. Wallin, preprint, in press: *Proteins* (2003).

[21] J.D. Bryngelson and P.G. Wolynes, *Proc. Natl. Acad. Sci. USA* **84**, 7524 (1987).

[22] A. Irbäck and E. Sandelin, *Biophys. J.* **79**, 2252 (2000).

[23] A. Irbäck and C. Troein, *J. Biol. Phys.* **28**, 1 (2002).

[24] A. Irbäck, C. Peterson and F. Potthast, *Proc. Natl. Acad. Sci. USA* **93**, 9533 (1996).

[25] E. Sandelin, preprint, in press: *Biophys. J.* (2003).

[26] A.R. Khokhlov and P.G. Khalatur, *Phys. Rev. Lett.* **82**, 3456 (1998).

[27] E.N. Govorun, V.A. Ivanov, A.R. Khokhlov, P.G. Khalatur, A.L. Borovinsky and A. Yu. Grosberg, *Phys. Rev. E* **64**, 040903(R) (2001).

[28] The list of all stable sequences with $N \leq 25$ and their ground states obtained in Ref. 23 is electronically available at http://www.thep.lu.se/complex/wwwserver.html

[29] H.S. Chan and K.A. Dill, *J. Chem. Phys.* **95**, 3775 (1991).

[30] A. Irbäck, C. Peterson, F. Potthast and E. Sandelin, *Struct. Fold. Des.* **7**, 347 (1998).