Designing specificity of protein-substrate interactions

Ivan Coluzza and Daan Frenkel
FOM Institute for Atomic and Molecular Physics, Kruislaan 407 1098 SJ Amsterdam, The Netherlands
(Received 11 June 2004; revised manuscript received 30 August 2004; published 30 November 2004)

One of the key properties of biological molecules is that they can bind strongly to certain substrates yet interact only weakly with the very large number of other molecules that they encounter. Using a simple lattice model, we test several methods to design molecule-substrate binding specificity. We characterize the binding free energy and binding energy as a function of the size of the interacting units. Our simulations indicate that there exists a temperature window where specific binding is possible. Binding sites that have been designed to interact quite strongly with specific substrates are unlikely to bind nonspecifically to other substrates. In other words, the conflict between specific interactions between small numbers of biomolecules and weak, nonspecific interaction with the rest need not be a very serious design constraint.

DOI: 10.1103/PhysRevE.70.051917 PACS number(s): 87.15.Rn, 82.20.Wt

I. INTRODUCTION

Biomolecules, such as proteins, tend to bind strongly to specific binding sites in target molecules. In addition, the binding needs to be selective: the molecules should bind strongly to one, or a few, partners and weakly, if at all, with all other biomolecules. The requirement that the binding should be strong and specific imposes constraints on the design of the binding sites. In particular, it suggests that binding sites should have a shape that is complementary to that of the substrate binding site and that its surface is patterned. Often, the total interaction (free) energy can be approximated as the sum of local intermolecular interactions that add coherently. In what follows, we focus on the role of the energetic patterning of binding sites.

It is important to recall that, even if the local intermolecular interactions are effectively random, binding is still possible. To see this, consider a nonspecific interaction with an associated binding energy that is the sum of N terms. We assume that the individual contributions are Gaussian distributed with a zero mean and variance σ^2 [1–4]. The probability P of having a binding energy E is given by

$$P(E) = (2\pi N\sigma^2)^{-1/2} e^{-[E^2/2N\sigma^2]}, \tag{1}$$

where N is the size (the number of interaction sites) of the binding region. The probability to form a bond its determined by the Boltzmann factor $\exp(-\beta E)$ corresponding to the interaction energy E. Even if the average interaction energy is zero, two sufficiently large binding regions are still likely to bind, as the average Boltzmann factor is given by

$$\langle \exp(-\beta E) \rangle = \exp(N\sigma^2\beta^2/2).$$

This implies that for large N, a truly random binding site is not inert. The effect of a nonspecific ("random") interaction has been discussed in detail by Pande *et al.* in the context of a study of the freezing transition in heteropolymers [5]. Note that the effective interaction strength due to random interactions scales with N, just as is the case for the interaction strength of specific (designed) interactions. However, the average strength per monomer is larger for designed specific interactions and hence one might expect that for any N one can always find conditions where specific binding dominates. But this argument ignores the fact that the spread in the binding free energy for random sequences is proportional to $\sqrt{N}$. Hence, for small enough N there are, most likely, specific random sequences that bind at least as strongly as the "designed" sequence. As N increases, ($\sqrt{N}/N$ decreases) this becomes less of a problem.

The above discussion suggests that binding sites should contain a sufficiently large number of monomeric units in order to guarantee that a designed binding site binds significantly stronger to a given template than a random binding site. Yet the site should be sufficiently small that nonspecific bonds can easily be disrupted by thermal fluctuations. One might think that this could be achieved by designing the individual site-site interactions to be small compared to the thermal energy k_BT. However, the same site-site interactions are responsible for the stability of the native state of the protein. Hence, weakening these interactions (or, equivalently, increasing the temperature) may result in denaturing of the protein, rather than in more specific binding.

There is a distinction between the specificity and selectivity of binding [6]. In order to quantify selectivity, it would be necessary to count the number of the substrate to which the protein can bind. In the present paper, we do not attempt such an exhaustive search (as this would be prohibitively expensive for the model systems that we consider). However, Gutin and Shakhnovich [7] showed, for a discrete version of the random energy model (REM), that the probability of degeneracy of the lowest-energy state decreases exponentially as its energy is lowered. This suggests that the specificity that we discuss below will, in most cases, also imply considerable selectivity.

In what follows, we consider under what conditions we can "design" a model substrate-binding site pair that binds significantly stronger than the corresponding "random energy" pair, while maintaining the structural integrity of the native state of the protein in solution. Hence, binding and folding are both the consequence of the heterogeneous interactions between monomeric units. To this end, we explore the role of system size and temperature on the binding specificity in a model that mimics a general protein-substrate sys-

tem. We consider two molecules, one of which (the "protein") is free to move, while the other is kept fixed and acts as the binding site of a substrate. We model the protein backbone as a linear, polypeptidelike heteropolymer living on a lattice. We then design ("evolve") the monomer sequence of the molecules according to three different scenarios, that we will refer to as OO, OR, and RR. First we consider the case of cooperative design, where the sequence of both the substrate and the ligand are evolved to increase the binding affinity. The second scenario is the model for a ligand that evolves to bind a substrate with a sequence that has been fixed *a priori*. The difference between model OO and OR lies in the role of the substrate. In scenario OO, the binding information is distributed over both protein and substrate: this approach should result in a protein and substrate that bind exclusively to each other. In the second approach, the protein is designed to bind to a specific substrate which, in its turn, can have multiple binding partners (low selectivity). Case RR represents the case of a protein-substrate pair that does not bind. This is the reference state that allows us to define the specificity of the other two systems as a function of the substrate size and temperature.

In the first part of this paper we describe the simulation techniques that we used to design and study protein-substrate interaction. We then discuss the binding of the two different molecules on the same substrate. We conclude with a discussion of the potential implications of this work.

II. MODELS AND METHODS

The system that we consider is a protein that is free to move in a box with hard walls in the presence of a substrate that is made of the same building blocks The box has a cubic shape and a lateral size of 3 times the length of the protein. The substrate is in the middle of the box. We model the chain as a linear, polypeptidelike heteropolymer, living on a lattice, with nearest-neighbor interactions. The conformational energy of the system is given by the following expression

$$E = E_{\text{intra}} + E_{\text{inter}} = \sum_{i}^{N_C} \left[\sum_{j \neq i}^{N_C} C_{ij} S_{ij} + \sum_{j' \neq i}^{N_S} C_{ij'} S_{ij'} \right], \quad (2)$$

where the indices i and j run over the residues of the protein, while j' runs over the elements of the substrate, and C is the contact matrix, defined as

$$C = \begin{cases} 1 & \text{if } i \text{ is neighbor of } j, \\ 0 & \text{otherwise}, \end{cases} \quad (3)$$

while S is the interaction matrix. For S we use the 20×20 matrix determined by Miyazawa and Jernigan [8] on the basis of the observed frequency of contacts between each pair of amino acids. It is important to notice that in Eq. (2) we do not include the interactions between the amino acids in the substrate. Although these interaction energies are, strictly speaking, neither energies nor free energies, they do provide a useful representation of the heterogeneity in the interactions between different amino acids.

A. Design of the folding and binding properties

A given lattice polymer can form a large number of compact conformations [2–4]. Obviously, every conformation is characterized by a different contact map. Hence, the energy of the polymer depends on its conformation. The mean-field approximation for its entropy is [1,9]

$$S(E) = \begin{cases} N \ln \gamma - \dfrac{E^2}{2N\sigma_B^2} & \text{if } E > E_c, \\ 0 & \text{if } E \leqslant E_c, \end{cases} \quad (4)$$

where N is the number of elements in the chain, σ_B is the standard deviation of the interaction matrix, and γ is the coordination number for fully compact structures on the lattice. E_c is the (lower) crossing point of the parabola with the abscissa, $E_c = -N\sigma_B (2 \ln \gamma)^{1/2}$. When the sequence of an heteropolymer is designed in a target configuration, a low-energy state is generated. If the energy E_N of this state is lower than E_c, then the system can fold in the target configuration. In the following we refer to this lowest-energy state as the native state of the heteropolymer. An important condition that must be satisfied for a successful design is that the homopolymers must be discarded. These particular sequences have highly degenerate ground states, which is not compatible to the definition of the folded state of a protein.

In Ref. [10] we presented a strategy to design a lattice protein in such a way that it will fold into a specific conformation. The basic design moves are single point mutations. As in the conventional Metropolis scheme, the acceptance of trial moves depends on the ratio of the Boltzmann weights of the new and old states. However, if this were the only criterion, there would be a tendency to generate homopolymer chains with a low energy, rather than chains that fold selectively into the desired target structure. To ensure the necessary heterogeneity, we impose the additional acceptance criterion

$$P_{\text{acc}} = \min\left\{ 1, \left(\frac{N_P^{\text{new}}}{N_P^{\text{old}}} \right)^{kT_p} \right\},$$

where T_p is an arbitrary parameter that plays the role of a temperature and N_P is the number of permutations that are possible for a given set of amino acids. N_P is given by the multinomial expression

$$N_p = \frac{N!}{n_1! n_2! n_3! \ldots}, \quad (5)$$

where N is the total number of monomers and n_1, n_2, etc., are the number of amino acids of type 1,2,… . While sampling the sequence space with a Monte Carlo scheme, we keep the temperature (T_{P}) associated with this quantity high. In doing so we generate a heterogeneous composition of amino acids. The importance of sequence heterogeneity for the design of specific structures is confirmed in our simulation, as it allows us to design heteropolymer sequences that have a nondegenerate native state. There is another, subtler, meaning of the "temperature" associated with the structural heterogeneity: it also is meant to represent the constraint that a protein lives in the presence of many other molecules to which it should not

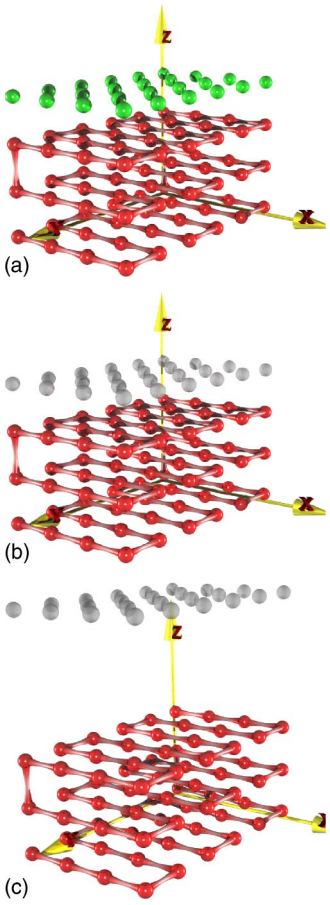

FIG. 1. (Color online) Spatial arrangement of the 72 amino acid chain with its 24 residues sub-strate, for scenario OO (a), scenario OR (b), and scenario RR (c).

bind unspecifically. By increasing this temperature we make it less likely that the protein will form an undesired, specific bond to any of the other proteins in the system. During a Monte Carlo run of several million cycles, a large number of distinct sequences are generated. The sequence S^* with the lowest energy is assumed to be the best candidate to fold into the native state. The energy of a given lattice polymer depends on its conformation.

$$E_{\text{native}} = \sum C_{ij} S_{ij}^{*}. \tag{6}$$

In this work we use this scheme to design a protein-substrate system with different binding properties. We start by imposing the template configuration, which should give information on the structure of the protein and on the desired bound state (e.g., Fig. 1). From the mean-field expression for the entropy in Eq. (4) we expect a wider distribution for the protein-substrate system, compared to the one of an isolated molecule. However, if the gap is still present, then the folded-bound state should be the equilibrium configuration. This condition does not exclude the case in which the interaction with the substrate is essential to keep the protein in the native state. Because we want to focus only on the binding properties regardless of the effect on the folding, we consider only a system with more intramolecular than intermolecular contacts—in other words, we use only compact proteins with a large fraction of intermolecular interactions.

In order to design the monomer sequence for the three different scenarios OO, OR, and RR we performed Monte Carlo sampling on a range of monomeric sequences. For each different scenario we applied the design process on a different subset of residues. In particular for case OO we include all the residues of both the protein and substrate, while for the others the sampling is limited to the amino acids of the protein, while the structure of the substrate is fixed.

Of course, once we have generated candidate sequences for the protein and substrate for the different cases, we still need to test if they do indeed have the desired binding properties.

B. Folding

To explore the possible conformations of the lattice polymer, we use three basic Monte Carlo moves: corner flip, crankshaft, and branch rotation. The corner flip involves a rotation of 180° of a given particle about the line joining its neighbors along the chain. The crankshaft move is a rotation by 90° of two consecutive particles. A branch rotation is a turn around a randomly chosen pivot particle of the whole section starting from the pivot particle and going to the end of the chain.

We explore the equilibrium properties of the system by sampling the free energy as a function of two order parameters. The first is the number of native contacts (both intramolecular and intermolecular) of the protein in a given conformation

$$Q(C) = \sum_{i<j}^{N} C_{ij}^{(1)} C_{ij}, \tag{7}$$

where $C_{ij}^{(1)}$ is the contact map of the reference structure and C_{ij} is the contact map of the instantaneous configuration. Only those contacts that belong to the reference structure contribute a value +1 to the order parameter. Because the number of native contacts includes the contacts with the substrate of the reference state, this order parameter can be used to compute the free-energy difference between the desired bound state and unbound state. A second order parameter Q_s allows us to study nonspecific binding. It is defined as the total number of contacts, native or non-native, between the chain and substrate. This order parameter is defined as

$$Q_s = \sum_{i}^{N_C} \sum_{j'}^{N_S} C_{ij'}. \tag{8}$$

Q_s allows us to characterize the interaction between the protein and substrate, irrespective of binding geometry. The free energy function the order parameter Q [Eq. (7)] is defined by

$$F(Q) = -kT \ln[P(Q)], \tag{9}$$

where $F(Q)$ is the free energy of the state with order parameter Q and $P(Q)$ is the histogram that measures the frequency of occurrence of conformations with order parameter

TABLE I. Sequences designed in the three different evolutionary scenarios and for the different protein-substrate sizes. The parameters used where, the design temperature $\beta_D = 20$ and the permutation temperature $\beta_P = 24$ in the range. Each letter represents a different amino acid (Ref. [5]). The letters in bold are the amino acids of the substrate.

Size	Scenario	Sequence	T_F
27	OO	YDCFRPIDGWRLQEMCKPNECWKNVEM **GSLYQFCTH**	0.2-0.5
27	OR	RQGCRDMDHIKWRELFKQSEVIKTMEL **YHYNGCNFP**	0.2-0.5
27	RR	MDSCRWLDCQKIMEFGKWMENQKWAER **HVPWYFKTP**	0.2-0.5
72		NDCALCKNREFIDMKDPEWRVMRGYDWVQMKQREWRL	
	OO	FKDNECIACKNPECTLCKYHEFIQMKDPEWPVMKH **GTFVTYHYSDWSLGHQNTGIACSS**	0.2-0.5
72		CNQSLRECMKDIFREWWHQGARNPFNDVGREMMKDG	
	OR	LREWCKQISPECAKQSLPESMKQIGREWFKDTAHNF **YCTWTYHMPVPLFHDVYKVITYNC**	0.2-0.5
72		GEQGDRKFLEQRNFKIIEMNSWHAIDMSNWKLLEMN	
	RR	DPKICEQRGPRFCDQADPKCLEMHQWKVIEMNSWRL **YCTWTYHMPVPLFHDVYKVITYNC**	0.2-0.5
75		NDMRPCDWKNIEMRCIDFKLAEGRLFQFKGIEMRLC	
	OO	DWKLNEMRCYQWKNSDMPPCQWKSIEMRCVQFKLG EFPV **VQGSTVTGSAHTWHAYDAHCYTWHY**	0.2-0.5
75		NDGWSHMGRDREFWHCQFKDAELPCCQVKAREIPCY	
	OR	MLKQTEFWHSMFRGADVWSYMLKAPEIWPCMLKQV EVPC **CYIYQHGGSNEMIKDKTTFTDRNNN**	0.2-0.5
75		NMQESAKRWNIMDEACKRFLHGQDHCRPGYIFQECT	
	RR	KRWLNMDEASKRWNAMDESTKRWSIMQEGCKPFLH GQDC **WTYHMPVPLFHDVYKVITYNCVIFE**	0.2-0.5
98		YCMRDQFIRREWCHLCMKDDLGRKEWCINCMKEDG IRKEWFNIGMREDLVSKEWFLNFMKEDAGRKEWCN	
	OO	VCMKEDTIRREWCVYCMKDQLGPSQWCP **PCPYTPGLTTSVYYIAFQSHIGTYHPHANFPQHS ALTQSMVNATFQHNV**	0.2-0.5
98		IWSKICDQCLEDMLNWRHFCFPCFEEMNAWKKGDY VRGEDMTHWRHSPVAQSDDMYAWKKGDAPSGEEM	
	OR	ANWKKFCQHCLEEMNIWRKICYSCLEQMA **FNGTLTRRQYVTVIQYPFMCLRGYKVPCIFNQTT PHDTRSIRYHPWHWV**	0.2-0.5
98		GMSIHQAYPELDWGNMKIKQHGREFEWNVMKCKD FASECEFLAMNCRSSASDCDWAVMKCKDAGRECE	
	RR	WVNMKCKQTYPELEWNGMRIKQHIPDLDWF **FNGTLTRRQYVTVIQYPFMCLRGYKVPCIFNQT TPHDTRSIRYHPWHWV**	0.2-0.5

Q. In practice, a direct (brute force) calculation of this histogram is not efficient, as the system tends to be trapped in local minima, especially at low temperatures. To solve this problem, we combined our sampling of chain conformations with a parallel-tempering routine [11–15]. Using this approach (with 14 different temperature windows 2, 1, 0.5, 0.2, 0.175, 0.15, 0.125, 0.1, 0.08, 0.061) we can get efficient sampling of the accessible free-energy landscape for the individual sequences.

However, we could not get the complete sampling of the free energy for all possible value of Q. To achieve this, we combined the normal parallel tempering with umbrella sampling of the polymer free-energy landscape [16–18]. In these simulations, we bias the sampling with respect to the order parameter such that all relevant conformations occur with approximately equal frequencies. We then bias the sampling of a particular value of the order parameter by imposing a bias potential that is opposite and (approximately) equal to

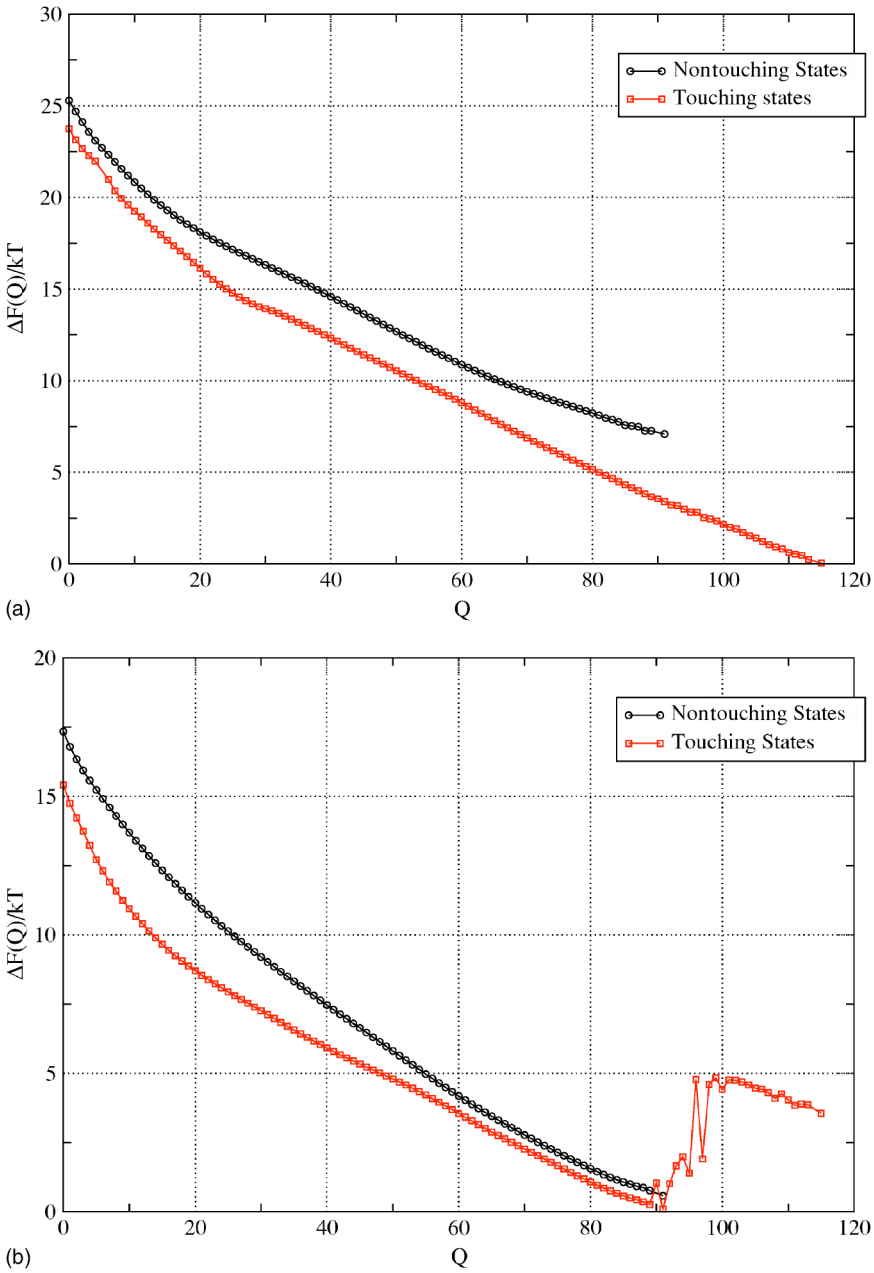

FIG. 2. (Color online) Plots of the free energy $F(Q)$ of the sequences OO (cooperative evolution) (a) and RR (independent evolution) (b), as a function of the number of native contacts Q [Eq. (7)], at $T=0.15$. States that touch the substrate (squares) have been plotted separately from those that do not (circles). The curve corresponding to the touching states is longer, because in teh definition of the order parameter we take into account also the native contacts with the substrate. All data were obtained with a combined parallel tempering and sampling simulation.

the free energy associated with that order parameter. As this free energy is not known *a priori*, the biasing potential is constructed iteratively. A more detailed description of this scheme is given in the Appendix.

III. RESULTS

To study the dependence of the binding specificity on system size and temperature, we consider a set of four different proteins with corresponding substrates. Each system was designed to reproduce the conditions of the three scenarios OO, OR, and RR. In order to design the first case we compute sequences of amino acids for the protein bound to the substrate, as shown in Fig. 1(a) for a protein with 72 residues and a substrate with 24 amino acids. In this case the design program will optimize the sequence to minimize the energy of the contacts within the chain and between chain and substrate. For the case OR, we impose the same target configuration as before, but we limit the optimization to the amino acids of the protein, and we assign a random sequence to the substrate. The final scenario is for nonspecific binding; this is achieved in two ways. First, we design a protein simply to fold into a given native structure, with no optimization of the substrate-binding energy. Second, we expose the protein form the OO and OR scenario to a random substrate without further design. It is important to stress that in the design of the OO and OR, the intramolecular bonds are optimized together with the intermolecular ones. In this way, we are able to construct model proteins that have the same internal structure both in the bound and unbound states. However, it is also possible to design structures that change upon binding. In Table I, we list the amino-acid sequences that were the result of the design procedure described above.

Free-energy calculations

As a first check, we verified that the generated sequences do indeed fold into the respective target structure. We show only the calculation of the binding free energy for a proteins consisting of 72 monomers (Fig. 1) as an example. In particular we consider the sequence OO (both protein and substrate optimized) and the sequence RR, where the protein sequence has been optimized to fold, but not to bind to a substrate, which has a random sequence. In Fig. 2 we plot the free energy of the sequences OO and RR, as function of the number of native contacts defined in Eq. (7). In each plot we distinguish between conformations that do and do not touch the substrate. As is to be expected (see Fig. 2), the binding free energy is much larger in the case where both the binding site and substrate have been optimized (OO), compared to the RR scenario. Moreover, in the case of the random interactions (RR), the free-energy minimum is reached before all contacts with the substrate are satisfied. To characterize the system in this regime, we computed the free energy $F(Q,Q_s)$ as a function of both the number of native contacts and the number of nonspecific contacts with the substrate [see Eq. (8)]. This should allow us to discriminate between configurations that are specifically and nonspecifically bound to the substrate. In Figs. 3(a) and 3(b) we plot $F(Q,Q_s)$ for OO and RR, respectively. The "funnel" shape of the surface in Fig. 3(a) demonstrates that the sequence OO does fold and sticks to the substrate in the designed way. In contrast, the free-energy surface for the sequence RR is flat at the bottom of the slope. This indicates that, in this case, the folded protein does not have a unique bound state with significant binding free energy. So much so that the presumed target state is not even favorable from a free-energy point of view. For the other sequences that we studied we found that, in every case, the design process (OO, OR, and RR) determined a similar free-energy landscape. The OR scenario (not shown in the figures) resulted in a free-energy landscape similar to that obtained in the OO case, but the binding strength was less. It is important to notice that in all scenarios the free chain retains the native intramolecular contacts, even in the unbound state. In [19] is presented a different situation where the substrate is able induce conformational changes.

Next, we consider the dependence of the binding strength on the size of the binding site. In Fig. 4(a) we plot the binding energy as a function of the size of the substrate for the three scenarios (OO, OR, and RR). The error bars represent the spread of the random interactions given in Eq. (1) around the mean value (calculated at two sigma). From the interaction matrix that we used, we get a mean interaction energy of around zero [8]. The figure shows that there is a significant gap (more than 2σ) between the binding energy in the case of designed binding sites compared to that of the purely random case of the designed energies and boundaries of the distribution. The gap is large enough to guarantee that the designed binding is energetically favorable compared to the random case, even for the smallest substrate. As expected, the binding specificity increases with the substrate size.

As mentioned in the Introduction, the presence of an energy gap between specific and nonspecific binding is not a sufficient condition to guarantee specific binding at any given temperature. To ensure specific binding of a given protein, there should exist a range of temperatures that are low enough to ensure that the designed protein structure is stable, yet high enough to guarantee that random (nonspecific) interactions are not strong enough to cause spurious bindings. As discussed in the Introduction, it is not *a priori* obvious that such a temperature window always exists. However, in the present case, it appears possible to satisfy this condition. Figure 4(b) shows the free-energy difference between the bound and unbound states of the chain in the native conformation for the cases OO, OR, and RR. As can be seen from the figure, the binding free energies behave more or less as the binding energies. In particular, a significant gap between specific and nonspecific bonding is maintained. This holds both for the case where both protein and substrate have been optimized and even for the case where only the protein has been optimized.

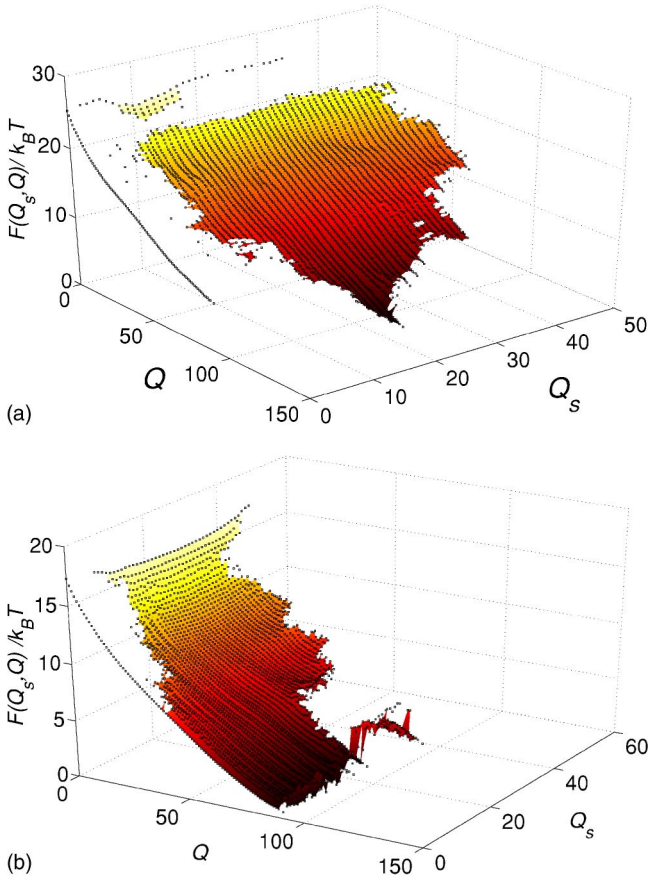

FIG. 3. (Color online) Plots of the free-energy landscape $F(Q,Q_s)$ of the sequences OO (evolution for binding) (a) and RR (random interaction) (b), as a function of the number of native contacts Q [Eq. (7)] and the number of contact with the substrate Q_s, 8, at $T=0.15$. The flat end of the slope in the second plot indicates that each bound state is equivalent in free energy to the unbound states, while in the first plot the funnel shape demonstrates that the cooperative evolved sequence has a clear free-energy advantage in the specific bind. The line separated from the surface represents the states that are not touching the substrate ($Q_s=0$), and the gap is caused by the poor sampling of the intermediate states.

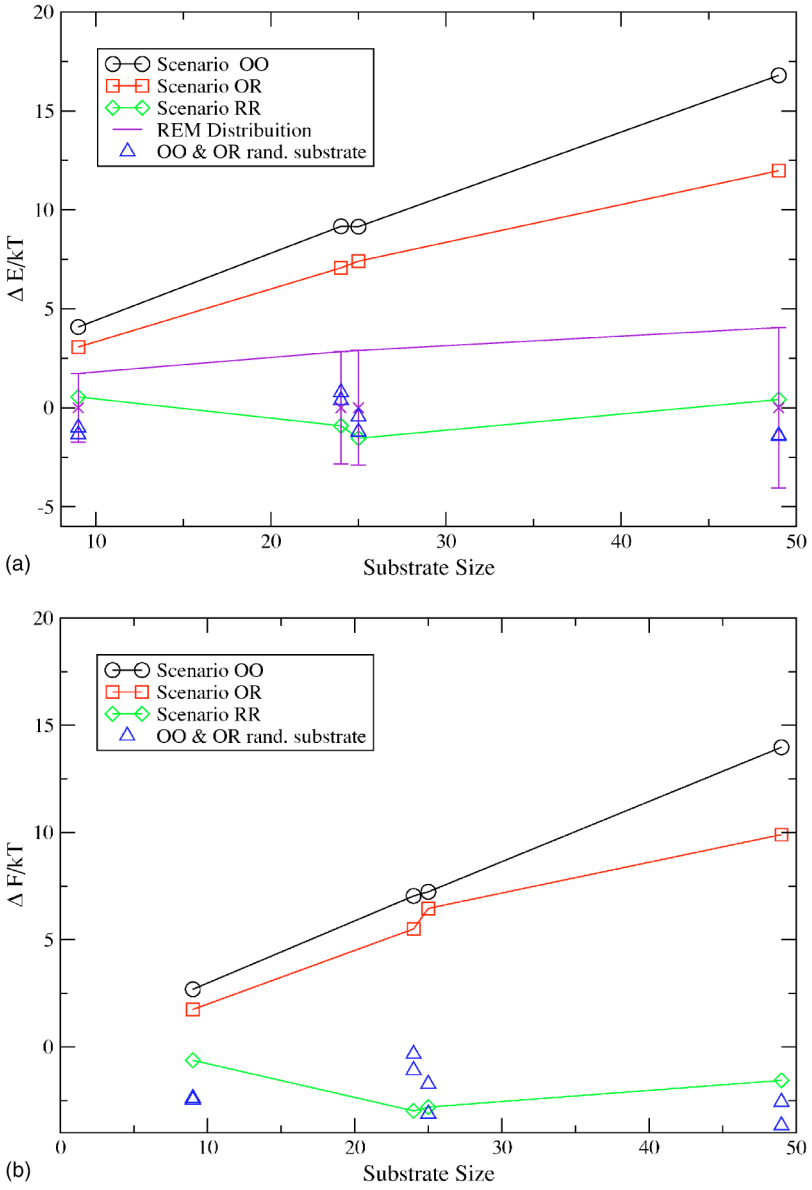

FIG. 4. (Color online) Plot of the size dependence of the binding energy (a) and of the binding free energy (b). The error bars in (a) represent the 2σ width of the distribution of interaction given by the random energy model. The triangles represent the interaction of the proteins designed for the OO and OR scenario, with a random substrate [1].

Clearly, the model used in the present study is highly simplified. Apart from the fact that we used a rather crude lattice model for the protein, we only considered the effect of binding energy on binding specificity. In reality, steric effects are at least as important and should be taken into account in any more realistic study. It would therefore be unwise to try to apply design calculations of the type described above to real protein systems. Nevertheless, some of the conclusions that we reach are likely to survive the transition to a more realistic model. First of all, the existence of a temperature window where specific binding is possible is also expected in models that take steric repulsion into account. Second (and interestingly), the present calculations suggest that binding sites that interact quite strongly with specific substrates are unlikely to bind nonspecifically to other substrates. In other words, the conflict between specific interactions between small numbers of biomolecules and a weak, nonspecific interaction with all the rest need not be a serious design constraint. This latter statement should be qualified: as the number of distinct species increases, so does the probability that at least one pair of molecules will, by accident, have a strong, nonspecific interaction. This will then result in an additional evolutionary pressure to keep nonspecific protein-protein interactions weak.

We note that the design of specific binding sites also plays a role in experimental schemes to detect specific proteins [20]. In this case a clear differentiation of the binding affinity between a substrate and proteins in solution is essential to isolate a particular molecule. As before, this implies a temperature window in which nonspecific bonds can be disrupted by thermal fluctuations, while the proteins themselves and the specific bonds that they form are still stable.

ACKNOWLEDGMENTS

We would like to thank Richard Sear for insightful comments. I.C. thanks F. Capuani and M. Cosentino-Lagomarsino for helpful discussions. This work was part of the research program of the Stichting voor Fundamenteel Onderzoek der Materie (FOM), which is financially sup-

ported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). An NCF grant for computer time on the TERAS supercomputer is gratefully acknowledged.

APPENDIX

Umbrella sampling and parallel tempering. Umbrella sampling is a method that speeds up the sampling of a rugged free-energy landscape by effectively flattening it. A simple way to flatten the landscape is to add a biasing potential to the normal Hamiltonian. To estimate this biasing potential we use an iterative method. During the simulation we sample the probability $P(Q)$ of finding a conformation with order parameter Q [Eq. (7)]. After a specified number of steps we calculate the new biasing potential W with the recursive equation

$$W_i(Q,T) = W_{i-1}(Q,T) - K \ln P(Q,T), \quad W_0(Q,T) = 0, \tag{A1}$$

where the index i indicates the iteration and K is a constant which we set to 0.5. Once we have the new biasing potential we add it to the energy in the acceptance criterion of every move. The potential W depends on the instantaneous structure of the system via the order parameter Q, but it also depends on the temperature. This temperature dependence is important when we combine umbrella sampling with parallel tempering. Each temperature has its own biasing potential. The acceptance rule for a temperature swapping move in the parallel tempering algorithm is then

$$P_{\mathrm{acc}} = \min\{e^{\Delta\beta\Delta E + \Delta W}, 1\},$$

$$\Delta W = W(Q_i,T_j) - W(Q_j,T_j) + W(Q_j,T_i) - W(Q_i,T_i), \tag{A2}$$

where i and j are replica indices. A similar procedure has recently been used in a paper by Faller *et al.* [21].

[1] B. Derrida, Phys. Rev. B **24**, 2613 (1981).
[2] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[3] T. Garel and H. Orland, Europhys. Lett. **6**, 307 (1988).
[4] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
[5] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Biophys. J. **73**, 3192 (1997).
[6] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. U.S.A. **91**, 12976 (1994).
[7] A. M. Gutin and E. I. Shakhnovich, J. Chem. Phys. **98**, 8174 (1993).
[8] S. Miyazawa and R. Jernigan, Macromolecules **18**, 534 (1985), Table VI.
[9] In the definition of the entropy the constant $\sqrt{\pi\sigma_{\mathrm{B}}^2}$ is ignored, as explained by Derrida [1].
[10] I. Coluzza, H. G. Muller, and D. Frenkel, Phys. Rev. E **68**, 046703 (2003).
[11] D. Frenkel and B. Smit, *Understanding Molecular Simulations* (Academic, New York, 2002), p. 389.
[12] R. H. Swendsen and J. S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[13] T. Geyer (unpublished).
[14] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).
[15] M. C. Tesi, E. J. J. vanRensburg, E. Orlandini, and S. G. Whittington, J. Stat. Phys. **82**, 155 (1996).
[16] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **329**, 261 (2000).
[17] N. B. Wilding, Phys. Rev. E **52**, 602 (1995).
[18] N. Tsunekawa, H. Miyagawa, K. Kitamura, and Y. Hiwatari, J. Chem. Phys. **116**, 6725 (2002).
[19] I. Coluzza and D. Frenkel (unpublished).
[20] J. Nam, C. S. Thaxton, and C. A. Mirkin, Science **301**, 1884 (2003).
[21] R. Faller, Q. Yan, and J. J. de Pablo, J. Chem. Phys. **13**, 5419 (2002).