

# Molecular recognition in a lattice model: An enumeration study

Thorsten Bogner, Andreas Degenhard, Friederike Schmid

Condensed Matter Theory Group, Fakultät für Physik, Universität Bielefeld

(Dated: July 30, 2004)

We investigate the mechanisms underlying selective molecular recognition of single heteropolymers at chemically structured planar surfaces. To this end, we study systems with two-letter (HP) lattice heteropolymers by exact enumeration techniques. Selectivity for a particular surface is defined by an adsorption energy criterium. We analyze the distributions of selective sequences and the role of mutations. A particularly important factor for molecular recognition is the small-scale structure on the polymers.

PACS numbers: 87.15.Aa, 87.14.Ee, 46.65.+g, 68.35.Md

Selective molecular recognition governs many biological processes such as DNA-protein binding<sup>1</sup> or cell-mediated recognition<sup>2</sup>. Biotechnological applications range from the development of biosensoric materials<sup>3</sup> to cell-specific drug-targeting<sup>4</sup>. The specificity in these processes results from the interplay of a few unspecific interactions (van der Waals forces, electrostatic forces, hydrogen bonds, and the hydrophobic force)<sup>5</sup> and a heterogeneous composition of the polymer chain. Selectivity is a genuinely cooperative effect. The question how it emerges in a complex system is therefore very interesting from the point of view of statistical physics, and the study of idealized models can provide insight into general principles<sup>6-9</sup>.

Previous theoretical studies have mostly considered heteropolymer adsorption in either regular<sup>10</sup> or random<sup>6,8,11,12</sup> systems. The interplay of cluster sizes on random heteropolymers and random surfaces and its influence on the adsorption thermodynamics and kinetics was studied analytically and with computer simulations<sup>8,12,13</sup>. Concepts from the statistical physics of spin glasses were used to study the adsorption of polymers on a “native” surface compared with that on an arbitrary random surface<sup>6,8,14</sup>.

In the present paper, we focus on a different question: We investigate mechanisms by which specific heteropolymers *distinguish* between given surfaces. To this end, we adopt an approach which has proven highly rewarding in the context of the closely related problem of protein folding<sup>15-18</sup>: We enumerate exactly all compact polymer conformations within a lattice model. The protein is described as a heteropolymer chain consisting of two types of monomers, hydrophobic (H) and polar (P), which occupy each one site on the lattice. Sites surrounding the polymer are assumed to contain solvent. The protein is exposed to an impenetrable flat surface covered with sites of either type H or type P, which form a particular surface pattern. It may adsorb there and change its conformation during the adsorption process. However, we require that both the free and the adsorbed chain are compactly folded in a given shape (cubic or rectangular)<sup>17,18</sup>. Nearest neighbor particles interact with fixed, type dependent interaction energies. Surface sites H and P are considered to be equivalent to monomer sites H

and P. The total energy is then given by:

$$E_{\text{tot}} = \sum_{\langle i,j \rangle} \sum_{\alpha,\beta} \tau_i^\alpha \tau_j^\beta E_{\alpha\beta} \quad (1)$$

Here the sum  $\langle i,j \rangle$  runs over nearest neighbor pairs, the sums  $\alpha$  and  $\beta$  run over the types hydrophobic (H), polar (P), or solvent (S), and  $\tau_i^\gamma$  is an occupation number which takes the value one if the site  $i$  is occupied with type  $\gamma$ , and zero otherwise. For compact chains with a fixed sequence, the energy spectrum as defined by Eq. (1) is (except for a fixed offset) fully characterized by only two parameters: One which describes the relative incompatibility of H and P inside the globule,  $V = 2E_{HP} - E_{HH} - E_{PP}$ , and one which accounts for the difference between the affinities of H and P to the solvent,  $W = 2(E_{HS} - E_{PS}) + (E_{PP} - E_{HH})$ . Since one of these parameters sets the energy scale, the model has only one dimensionless free parameter,  $V/W$ . Motivated by Ref.<sup>17</sup>, where  $V/W = 0.13$ , we chose  $V/W = 0.1$ .

We consider two-dimensional and three-dimensional systems with system sizes up to  $6 \times 6$  (in 2D) and  $3 \times 3 \times 3$  (in 3D), respectively. For each system, a set of sequences was picked randomly (uncorrelated monomers, equal probability for H and P). For each sequence, we then evaluated the energies for all possible compact chain conformations in contact with all possible surfaces. This allowed us to determine exactly the ground-state adsorption energy on every surface. We call a sequence *selective*, if there exists one unique surface with highest adsorption energy, *i.e.*, if the difference

$$E_{\text{gap}} = E_{\text{ad}}^{1st} - E_{\text{ad}}^{2nd} \quad (2)$$

between the adsorption energies on the two most favorable surfaces is nonzero. The lowest-energy *structure* of the chain on its favorite surface (the “selected” surface) is not necessarily unique.

We note that this selectivity criterion is a “zero-temperature” criterion. Entropic contributions to the adsorption free energy are not accounted for. Furthermore, we disregard dynamic and kinetic factors<sup>19</sup>, which presumably also play a role in molecular recognition processes.

In all systems, more than 90% of all sequences were selective. The distribution on the different surfaces was

highly inhomogeneous, see Fig. 1. A closer inspection reveals that two main factors contribute to the frequency with which sequences select a particular surface pattern: A high number of hydrophobic sites inside the pattern is beneficial, whereas hydrophobic sites at the border are unfavorable. This is due to the fact that bound proteins preferring the latter surface patterns must have hydrophobic monomers at the edges. The resulting unfavorable contacts to the solvent have to be compensated to achieve an energetic minimum. This reduces the number of suited sequences. The frequency distribution could be fitted remarkably well by the simple formula

$$N = \frac{A n_{core}^m + B}{n_{border} + 1}, \quad (3)$$

where  $n_{core}$  denotes the number of hydrophobic core sites,  $n_{border}$  the number of hydrophobic border sites,  $m$  the total number of core sites, and  $A$  and  $B$  are fit parameters. For the  $5 \times 5$  system, such a fit is illustrated in Fig. 1. The fitting is also successful for other systems, even for the 3D case, if one identifies sites at the corner of the surface with border sites. The functional form of Eq. (3) was guessed empirically, with no underlying theory, and should not be over-interpreted. Nevertheless, we can conclude that the relative frequency of surface patterns is mostly determined by a few, unspecific surface characteristics.

The previous analysis raises the question how sequences which are selective for different surfaces differ from one another, or, conversely, which features sequences belonging to the same surface have in common. We have used different approaches to address this problem.

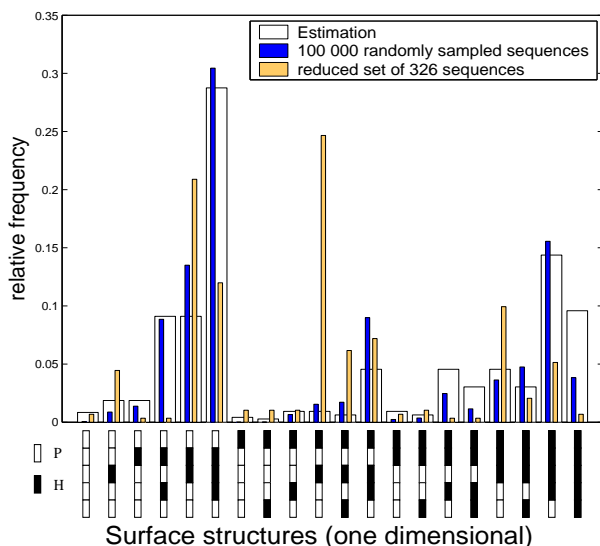


FIG. 1: Relative frequency of sequences selective for different surfaces on the  $5 \times 5$  lattice. The black bars show the result for a random sample, the gray bars for a sample based on a “master sequence”. Also shown are the values obtained with a least-square fit to Eq. (3) (see text for explanation).

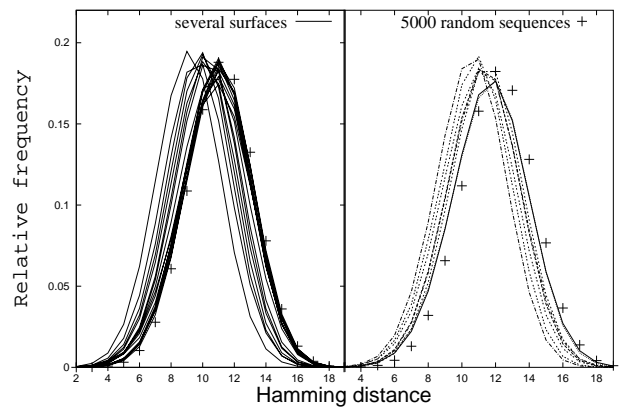


FIG. 2: Histograms showing the distribution of Hamming distances for the  $5 \times 5$  (left) and the  $3 \times 3 \times 3$  (right) lattice in set of sequences belonging to several surfaces (lines). Also shown for comparison are the results for a set of 5000 random sequences (crosses).

The first approach was motivated by the biological principle of mutation. A similarity measure between two chain sequences can be defined by counting the minimum number of point mutations required to construct one sequence, starting from the other. For our two-letter sequences, this is quantified using the *Hamming distance*

$$d(s, s') := \frac{1}{2} \sum_i |s_i - s'_i|, \quad (4)$$

between sequences  $s$  and  $s'$ . The sum  $i$  runs over all monomers along the chain, and the variables  $s_i, s'_i$  are taken to be  $s_i, s'_i = 1$  if the  $i$ th monomer of the sequence  $s$  is hydrophobic, and  $s_i, s'_i = -1$  otherwise. Two sequences that have a Hamming distance of  $n$  are thus separated by  $n$  point mutations. Since sequences can be read in both directions, Eq. (4) usually yields two values for a pair of sequences. We have always used the smaller one.

Based on this definition, we can now study whether sequences belonging to the same surface are “close” in sequence space. Examples of distributions of Hamming distances for different surfaces are shown in Fig. 2. The distributions for different surfaces, and even for different system sizes, are very similar. The number of mutations with the highest occurrence is nearly half the total number of monomers in the polymer chain. Moreover, the distribution is not very different from that of a totally random set of sequences, which is also shown in Fig. 2 for comparison. Hence we conclude that the sequences selective for a particular surface are widely distributed over the sequence space, and that proximity in sequence space is not a relevant factor for molecular recognition.

This result has interesting practical consequences. An important issue for many cell-surface recognition processes is the question how efficiently nature distinguishes between different surfaces<sup>20</sup>, *i.e.*, how many mutations are required to change a polymer sequence that is selec-

tive for a particular surface to make it selective to another surface or a whole class of different surfaces. In our model, the observation that sequences selective to the *same* surface appear to be widely spread in sequence space suggests that one might find sequences which are selective to very *different* surfaces at close vicinity in sequence space.

In order to test this idea, we have attempted to compute subsets of sequences, which are close in sequence space and nevertheless “recognize” all surfaces, *i.e.*, which contain at least one selective sequence for each surface. Such sets were constructed following a two-step procedure. First, we identified a *center* or *master* sequence, which was a suitable initial point for the mutation process. This was done mainly by trial and error, starting from the sequences belonging to the least favorable surfaces. Second, we evaluated the number of mutations necessary to provide a subset of sequences recognizing all surfaces. This analysis was carried out for different two-dimensional systems. The results are shown in table I. In spite of the exponential growth in the number of possible polymer chain conformations and possible sequence realizations, the number of necessary mutations  $r$  in table I increases only slightly with the surface size. The distribution of the sequences on the surfaces is shown for one of these reduced subsets in Fig. 1, and can be compared with the full distribution. The general features are comparable.

We note that the values  $r$  for the minimum number of mutations required to *recognize* all surfaces, as given in table I, are upper limits and can possibly be reduced further with more efficient master sequences. Even so,  $r$  is in some cases smaller than the minimum number of mutations necessary to *generate* all surfaces (starting from a common master surface). Hence only a few point mutations can alter the adsorption characteristics profoundly. This result matches with experimental results obtained from binding force measurements on antibodies<sup>21</sup>. Experimentally, it was observed that the wild-type antibody and a mutant in which an amino acid at one position in the chain has been exchanged differ in the measured affinity by roughly one order of magnitude.

We return to the problem of determining common features of sequences which are selective for the same surface. To clarify the question whether there exist any such features, we have applied an artificial neural net-

TABLE I: Number of mutations  $r$  necessary to generate a subset of sequences which recognize all surfaces, together with the corresponding subset size for various lattice sizes. In the case of rectangular folding ( $5 \times 4$  and  $6 \times 5$ ), the largest side forms the interface to the surface.

Lattice	surface size	r	size of set
$5 \times 4$	5	2	209
$5 \times 5$	5	2	326
$6 \times 5$	6	3	466
$6 \times 6$	6	4	7807

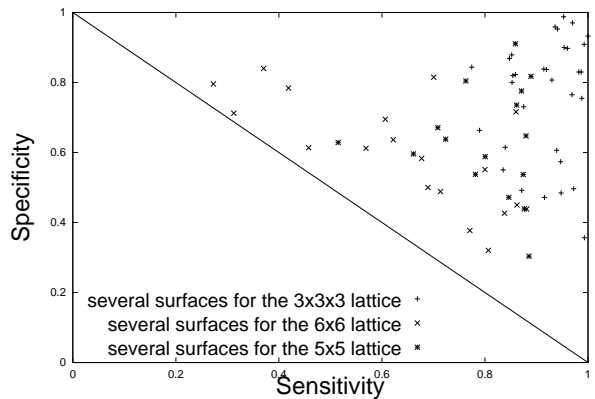


FIG. 3: Performance of the fully-connected two-layer perceptron trained for several surface structures on the  $5 \times 5$  and  $3 \times 3 \times 3$  lattice displayed in a sensitivity (true positive) versus specificity (true negative) plot. The diagonal line represents results with a 50% correct classification rate corresponding to random guessing. For the  $6 \times 6$  system the results were obtained by a fully-connected three-layer perceptron with 16 hidden units. In all cases the data have been transformed to Fourier space and the perceptron was optimized via a back-propagation algorithm, see Ref.<sup>22</sup>

work (ANN). After training the ANN with a set, composed equally from selective as well as non selective sequences for a given surface, the performance of the ANN was tested with a second, disjoint set. This analysis was performed for all surfaces with at least 100 selective sequences. The results of the testing, Fig. 3, show that there do exist relevant features for the recognition process that can be learned by the ANN.

The next question is: What does the ANN learn? In the case of a two-layer perceptron, the answer is relatively simple<sup>22</sup>: The ANN classifies by dividing the sequence space of dimension  $N$  into two parts by a  $N - 1$  dimensional hyper-plane. The fact that this classification is successful suggests that insight might be gained by a more general characterization than the mere mutual (Hamming) distances. In order to achieve this we applied the “Principal Component Analysis” (PCA)<sup>23</sup>. In this approach, the data, *i.e.* in the present paper the discrete Fourier transform of the sequences, are treated as a random vector. In general the modes are correlated, in particular if common features within a set of sequences exist. This is characterized by the variances and covariances given in the covariance matrix. Diagonalization of this matrix yields a description by uncorrelated components, the eigenvectors. The eigenvalues are a measure for the squared variances of these components. Low eigenvalues correspond to characteristic components within the set.

We have carried out PCAs for various surfaces in the  $5 \times 5$ -, the  $6 \times 6$ -, and the  $3 \times 3 \times 3$  system. The results revealed an unexpected common feature: For all surfaces in the  $5 \times 5$ - and the  $3 \times 3 \times 3$  system, two components turned out to be especially meaningful, namely almost

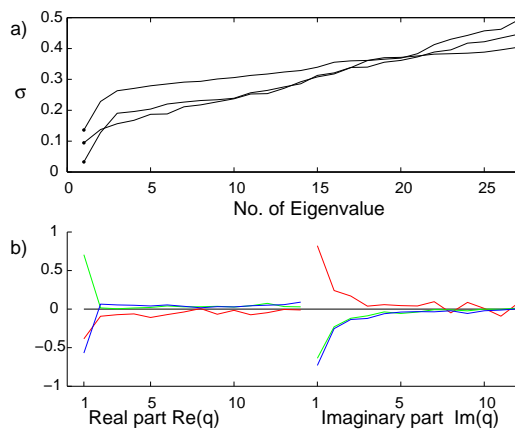


FIG. 4: (a) Square roots of the eigenvalues of the covariance matrix and (b) coordinates in Fourier space ( $q$ -space) of the eigenvector corresponding to the smallest variance (circles in a) for three surface patterns of the  $3 \times 3 \times 3$  system.

exactly the highest frequency modes (real and imaginary part). The corresponding variances were considerably smaller than those of all other components, see Fig. 4. In the  $6 \times 6$  system, the result was not as simple, yet the high-frequency components were still among the significant components.

These results can be visualized by projecting the sequence space onto the highest-frequency plane. Fig. 5 illustrates for the  $3 \times 3 \times 3$ -system that sets of sequences belonging to different surfaces often occupy different regions in this plane.

To summarize, we have studied the recognition of chemically structured surfaces by single polymer chains

comprising hydrophilic and hydrophobic monomer units. Starting from already folded conformations, we investigated distributions of selective sequences and the role of point mutations. We found that sequences recognizing the same surface are widely distributed in sequence space, *i.e.*, they are separated by many mutations. Conversely,

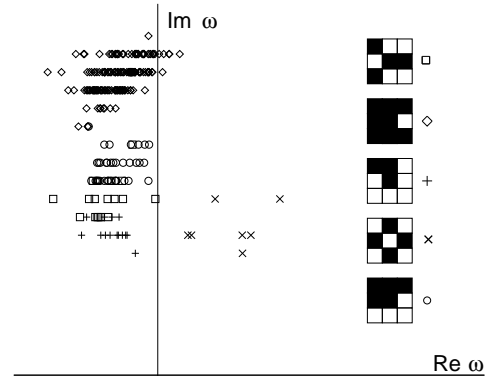


FIG. 5: Projection of sequences on the highest frequency ( $\omega$ ) plane for the  $3 \times 3 \times 3$  lattice and various surface structures. Some of the sets are completely separated in this plane.

it was in many cases possible to construct a subset of sequences which recognize all surfaces and nevertheless differ from one another by only a few mutations. Despite their wide distribution, sequences recognizing the same surface have features in common, which can be learned by a neural network. One factor which turned out to be particularly important in this recognition process is the local, small-scale structure on the polymers.

We thank Alexey Polotsky for useful discussions and the german science foundation (DFG) for partial support.

- <sup>1</sup> K. Zakrzewska, *Biopolymers* **70**, 414 (2003).
- <sup>2</sup> B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson, *Molecular biology of the cell* (Garland Publishing, Inc. New York & London, 1994), 3rd ed.
- <sup>3</sup> E. Nakata, T. Nagase, S. Shinkai, and I. Hamachi, *J. Am. Chem. Soc.* **126**, 490 (2004).
- <sup>4</sup> Y. Christi, *Biotechn. Adv.* **22**, 313 (2004).
- <sup>5</sup> J. Israelachvili, *Intermolecular and surface forces* (Academic Press, 1991).
- <sup>6</sup> J. Janin, *Proteins* **25**, 438 (1996).
- <sup>7</sup> M. Muthukumar, *Proc. Nat. Ac. Sci: USA* **96**, 11690 (1999).
- <sup>8</sup> A. K. Chakraborty, *Phys. Rep.* **342**, 1 (2001).
- <sup>9</sup> N.-K. Lee and T. A. Vilgis, *Phys. Rev. E* **67**, 050901 (2003).
- <sup>10</sup> M. Muthukumar, *J. Chem. Phys.* **103**, 4723 (1995).
- <sup>11</sup> A. Polotsky, F. Schmid, and A. Degenhard, *J. Chem. Phys.* **120**, 6246 (2004).
- <sup>12</sup> A. Polotsky, F. Schmid, and A. Degenhard, *J. Chem. Phys.* (to appear in Vol.121 no.9).
- <sup>13</sup> A. J. Golumbskie, V. S. Pande, and A. K. Chakraborty, *Proc. Nat. Ac. Sci: USA* **96**, 11707 (1999).
- <sup>14</sup> S. Srebnik, A. K. Chakraborty, and E. L. Shakhovich, *Phys. Rev. Lett.* **77**, 3157 (1996).
- <sup>15</sup> K. A. Dill, *Biochemistry* **24**, 1501 (1985).
- <sup>16</sup> K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
- <sup>17</sup> H. Li, R. Helling, C. Tang, and N. S. Wingreen, *Science* **273**, 666 (1996).
- <sup>18</sup> H. Li, C. Tang, and N. S. Wingreen, *Phys. Rev. Lett.* **79**, 765 (1997).
- <sup>19</sup> J.-U. Sommer, *Eur. Phys. J. E* **9**, 417 (2002).
- <sup>20</sup> M. Davis, M. Krogsgaard, J. Huppa, C. Sumen, M. Purbhoo, D. Irvine, L. Wu, and L. Ehrlich, *Ann. Rev. Bioch.* **72**, 717 (2003).
- <sup>21</sup> R. Ros, F. Schwesinger, D. Anselmetti, M. Kubon, R. Schäfer, and A. Plückthum, *Proc. Nat. Ac. Sci: USA* **95**, 7402 (1998).
- <sup>22</sup> C. M. Bishop, *Neural Networks for Pattern Recognition* (Clarendon Press; Oxford University Press, 1995).
- <sup>23</sup> I. T. Jolliffe, *Principal component analysis* (Springer, 1986).