On Hydrophobicity Correlations in Protein Chains

Anders Irbäck and Erik Sandelin
Complex Systems Division, Department of Theoretical Physics, Lund University, Sölvegatan 14A, S-223 62 Lund, Sweden

ABSTRACT We study the statistical properties of hydrophobic/polar model sequences with unique native states on the square lattice. It is shown that this ensemble of sequences differs from random sequences in significant ways in terms of both the distribution of hydrophobicity along the chains and total hydrophobicity. Whenever statistically feasible, the analogous calculations are performed for a set of real enzymes, too.

INTRODUCTION

Functional protein sequences exhibit the ability to fold spontaneously into a unique native state (Creighton, 1993). A natural step in order to understand this crucial property is to compare good and bad folding sequences in simple models where conformational space can be properly explored. Most such studies have been directed toward identifying physical characteristics of good folders, and in this important area some progress has been made (Săli et al., 1994; Bryngelson et al., 1995; Klimov and Thirumalai, 1998; Nymeyer et al., 1998). In this paper we address the question of how good folders differ from random sequences in purely statistical terms. A related but different topic is how sequences that share the same (unique) native state are distributed in sequence space. This question and its evolutionary implications have recently attracted considerable attention (Li et al., 1996; Bornberg-Bauer, 1997; Govindarajan and Goldstein, 1997a,b; Bastolla et al., 1999; Broglia et al., 1999; Bornberg-Bauer and Chan, 1999; Tiana et al., 2000).

In a recent study of a hydrophobic/polar off-lattice model, it was found that good folders tend to show negative hydrophobicity correlations along the chains (Irbäck et al., 1997). The analogous calculations gave, moreover, qualitatively similar results for a major class of real proteins, corresponding to typical total hydrophobicities (Irbäck et al., 1996). On the other hand, the opposite behavior, positive hydrophobicity correlations, has been reported for a class of designed model sequences that display certain protein-like features (Khokhlov and Khalatur, 1998, 1999). These designed sequences are, for instance, not meant to have unique native states, so the different results do not represent a contradiction. However, it shows that sequence correlations in proteins is a delicate issue that requires a careful analysis.

The main goal of this paper is to test the robustness of the conclusion that good folding model sequences as well as functional proteins show negative hydrophobicity correlations. To this end we perform new calculations for both model and real sequences. The model we study is the minimal HP model on the square lattice (Lau and Dill, 1989; Dill et al., 1995). This choice makes it possible for us to improve significantly on the statistics in the previous study (Irbäck et al., 1997), which was based on an off-lattice model. The real sequences studied are single-domain enzymes taken from the CATH protein structure classification database (Orengo et al., 1997), which we hope displays statistical properties representative of functional (globular) folding units. With this restriction on protein type, it turns out that the previous, somewhat artificial, restriction on total hydrophobicity (Irbäck et al., 1996) can be lifted.

METHODS

Sequences

Let us first define the sequences studied. The real sequences studied are the 173 nonhomologous single domain enzymes found in the October 1998 release of the CATH database (Orengo et al., 1997). These sequences are transformed into binary hydrophobicity strings, by taking the six amino acids Leu, Ile, Val, Phe, Met, and Trp as hydrophobic ($\sigma_i = 1$) and the others as hydrophilic ($\sigma_i = -1$). This choice is somewhat arbitrary. Therefore, we also tried a 20-valued hydrophobicity scale, which did not affect any of the conclusions below. In CATH, the most general level of classification is denoted "class" and describes the relative content of α helices and β sheets. Below, the class dependence of our results is checked by separate calculations for each of the three major classes: mainly α, mainly β, and $\alpha\beta$. A fourth class, low secondary structure content, exists but it is not considered separately, as only 3 of the 173 sequences belong to it. In our calculations we also divide the sequences into extracellular and intracellular ones. Following Martin et al. (1998), we take the presence of a disulphide bridge as an indicator of extracellular location. The number of enzymes in the different subsets studied can be found in Table 3 below.

The model we use is the minimal two-dimensional HP model (Lau and Dill, 1989), whose behavior is known in quite some detail (Dill et al., 1995). It contains only two types of amino acids, H (hydrophobic, $\sigma_i = 1$) and P (polar, $\sigma_i = -1$), and the chain conformation is represented as a self-avoiding walk on a lattice. The formation of a hydrophobic core is favored by defining the energy as minus the number of HH pairs that are nearest neighbors on the lattice but not along the chain. On the square lattice, it turns out that this simple choice of energy function is sufficient in order to get a significant number of sequences with unique ground states (Chan and Dill, 1994; Irbäck and Sandelin, 1998); complete enumeration of all possible sequences and structures shows that the fraction of such sequences is roughly 2% for $N \leq 18$. Throughout this paper we consider all HP sequences that have unique ground states as good folding sequences.

Received for publication 18 January 2000 and in final form 23 May 2000.
Address reprint requests to Dr. Anders Irbäck, Lund University, Department of Theoretical Physics, Complex Systems Division, Sölvegatan 14A, S-22362 Lund, Sweden. Tel.: 46-46-222-3493; Fax: 46-46-222-9686; E-mail: irback@thep.lu.se.

0006-3495/00/11/2252/07 $2.00

Also central is that the sequences are able to fold fast into their native states, a requirement that we ignore. This is a reasonable simplification because the sequences are short and because almost all have the same energy gap between ground state and next lowest level.

Sequence correlations

Our statistical analysis of hydrophobicity strings can be divided into two parts. The first part deals with the distribution of hydrophobicity along the chains; how does a "good" sequence with length N and total hydrophobicity

$$M = \sum_{i=1}^{N} \sigma_i \tag{1}$$

differ from a typical sequence with the same N and M? This question can be addressed by monitoring variables such as the number of hydrophobic and hydrophilic clumps along the chain (White and Jacobs, 1990), Fourier amplitudes (Irbäck et al., 1996), or random walk (Brownian bridge) representations (Pande et al., 1994). In this paper we work with block variables, a widely used technique that has proven useful in studies of DNA sequences (Peng et al., 1992) as well as proteins (Irbäck et al., 1996).

In addition to the distribution of hydrophobicity along the chains, we also study the distribution of the total hydrophobicity M. This analysis relies entirely on comparisons between observed sequences, which makes it statistically more difficult, especially for the real sequences with varying N.

The blocking method

In this method, for a given size s, the sequence is divided into blocks each consisting of s consecutive σ_i along the chain. The block variable $\sigma_k^{(s)}$ is then defined as the sum of the s σ_i values in block k ($k = 1, \ldots, N/s$). A useful quantity is the mean-square fluctuation

$$\psi^{(s)} = \frac{s}{N} \sum_{k=1}^{N/s} \psi_k^{(s)} \qquad \psi_k^{(s)} = \frac{1}{K} (\sigma_k^{(s)} - sM/N)^2 \tag{2}$$

where we choose the normalization factor

$$K = \frac{N^2 - M^2}{N^2 - N}(1 - s/N). \tag{3}$$

With this choice, the average of $\psi^{(s)}$ over all possible sequences with given N and M takes the simple form (Irbäck et al., 1996)

$$\langle \psi^{(s)} \rangle_{N,M} = s, \tag{4}$$

independent of N and M.

The distribution of total hydrophobicity

We study the M distribution for different fixed N, focusing on the mean $\langle M \rangle_N$ (the subscript indicates fixed N) and the normalized variance

$$\chi = \frac{1}{N} \langle (M - \langle M \rangle_N)^2 \rangle_N . \tag{5}$$

It is easily verified that

$$\chi = \frac{4}{N} \sum_{i=1}^{N} h_i(1 - h_i) + \frac{1}{N} \sum_{i \neq j} c_{ij} , \tag{6}$$

where $h_i = (1 + \langle \sigma_i \rangle_N)/2$ denotes the fraction of sequences that have $\sigma_i = 1$, and $c_{ij} = \langle \sigma_i \sigma_j \rangle_N - \langle \sigma_i \rangle_N \langle \sigma_j \rangle_N$ is the σ_i, σ_j correlation. So, if the σ_i values are uncorrelated, then

$$\chi = \chi_1 \equiv \frac{4}{N} \sum_{i=1}^{N} h_i(1 - h_i), \tag{7}$$

which becomes

$$\chi = \chi_0 \equiv 4h(1 - h) \tag{8}$$

in case the hydrophobicity profile $\{h_i\}$ is flat with $h_i = h$ for all i. Below these two predictions are tested for the model sequences.

Unfortunately, our set of enzymes cannot be analyzed this way, due to limited statistics. However, as we will see, it turns out that the data for the mean $\langle M \rangle_N$ can be approximately described by a simple linear relation, $\langle M \rangle_N \approx \bar{M} = (2\bar{h} - 1)N$. As an effective measure of the fluctuations in M, we therefore consider

$$\bar{\chi} = \left\langle \left(\frac{M - \bar{M}}{N^{1/2}} \right)^2 \right\rangle, \tag{9}$$

where the average now is over all sequences, irrespective of N. If the σ_i values for each N were uncorrelated with identical $h_i = \bar{h}$, then we would have

$$\bar{\chi} = \bar{\chi}_0 \equiv 4\bar{h}(1 - \bar{h}). \tag{10}$$

Let us finally stress that $\psi^{(s)}$ and χ are fundamentally different measurements. In the blocking method individual sequences are compared to random sequences with the same N and M. Hence, $\psi^{(s)}$ provides direct information on the distribution of $\sigma_i = \pm 1$ along the chains. This is not true for χ and the correlation c_{ij}. This correlation is not necessarily physical. The behavior of the analogue of c_{ij} in the ordered phase of an Ising magnet provides an illustration of this. In this case, c_{ij} does not vanish at large distance, although the physical correlation length is finite.

Individual structures

As mentioned in the Introduction, several recent model studies have addressed the question of how sequences that fold to the same native state are related. In particular, using an HP-like model with compact structures only, Li et al. (1996) found that structure-preserving mutations tend to be largely independent for highly designable structures. To see whether this behavior is consistent with our analysis, we perform two measurements for different fixed structures, too.

Consider a given structure r, and let $\{h_i^{(r)}\}$ be the corresponding hydrophobicity profile ($h_i^{(r)}$ is the probability that $\sigma_i = 1$). The first quantity we calculate is

$$\Delta\chi^{(r)} = \chi^{(r)} - \frac{4}{N} \sum_{i=1}^{N} h_i^{(r)}(1 - h_i^{(r)}), \tag{11}$$

where $\chi^{(r)}$ is defined as χ in Eq. 5 but for fixed structure. $\Delta\chi^{(r)}$ measures the average σ_i, σ_j correlation for fixed structure (see Eq. 6). The second

quantity is the entropy

$$S = -\sum_{i=1}^{N} [h_i^{(r)} \ln h_i^{(r)} + (1 - h_i^{(r)}) \ln(1 - h_i^{(r)})] \quad (12)$$

for a system of independent σ_i with hydrophobicity profile $\{h_i^{(r)}\}$. If the σ_i values are approximately independent, then e^S provides an order-of-magnitude estimate of the actual number of sequences, N_r. If this is not the case, then e^S overestimates N_r.

RESULTS

In this section we present the results of our analyses of the mean-square block fluctuations $\psi^{(s)}$ and the distribution of total hydrophobicity, M, for model and real sequences. We end the section with some comments on our model results and related studies of similar models.

The blocking method

Model sequences

In our block variable analysis of HP sequences, we consider the 6349 $N = 18$ sequences that have unique native states, which can be obtained by exhaustive enumeration (Chan and Dill, 1994). The results are compared to expected values for random sequences, as described in Methods. This comparison makes sense only if the hydrophobicity profile $\{h_i\}$ is uniform. From Table 1 it can be seen that h_i is approximately constant in the midpart but increases towards the ends. As a check, we therefore calculate the mean-square block fluctuation $\psi^{(s)}$ in two ways for each sequence: first, for the full sequence; and second, after elimination of two amino acids at each end. Fig. 1 shows the results of both these calculations. We see that the average $\psi^{(s)}$ is smaller than for random sequences, irrespective of whether the endpoints are included or not. The conclusion that $\psi^{(s)}$, on average, is suppressed for good sequences is in perfect agreement with earlier results for a different model (Irbäck et al., 1996, 1997).

Enzymes

We now repeat essentially the same analysis for the enzymes. The only difference is that, because N is not fixed, the hydrophobicity profile $h(\xi)$ is taken to be a function of the relative position ξ along the chains. To calculate $h(\xi)$, we divide the interval in ξ from 0 (N end) to 1 (C end) into 100 bins. The results obtained are shown in Fig. 2 *a*. We see that $h(\xi)$ is approximately constant throughout the interval $0 \leq \xi \leq 1$.

In an earlier block analysis of functional protein sequences (Irbäck et al., 1996), in which there was no restriction on protein type, the ends were found to display a different behavior than the rest of the sequences, and therefore they were removed from the analysis. To check if this is true for the present data set, we calculate the average of $\psi_k^{(4)}$ (see Eq. 2) as a function of ξ, using 25 bins in ξ. The results are shown in Fig. 2 *b*. Although the uncertainties are somewhat large, there is no sign of the ends behaving differently.

Given these two findings, we calculate the block fluctuations using the full sequences, without any elimination of amino acids at the ends.

In Fig. 3 we show the average $\psi^{(s)}$ against block size s for the 173 enzymes. Also shown are the results obtained for five different subsets of these sequences (see Methods). We see that the results are similar in the different cases, and that $\psi^{(s)}$ is smaller than for random sequences. Qualitatively, the behavior is similar to that found for the model sequences.

In this analysis we have chosen to focus on $\psi^{(s)}$. Similar deviations from randomness are expected in other quantities such as the number of hydrophobic/hydrophilic clumps along the chain. The number of clumps tends to be large when $\psi^{(s)}$ is small (Irbäck et al., 1997).

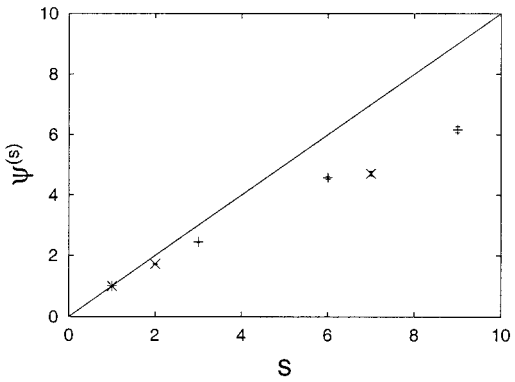

FIGURE 1 The mean-square block fluctuation $\psi^{(s)}$ against block size s for good $N = 18$ sequences in the HP model. Shown are results both for the full sequences (+) and for the subsequences consisting of the central 14 amino acids (×). The straight line represents random sequences; see Eq. 4.

TABLE 1 Hydrophobicity profile $\{h_i\}$ for good $N = 18$ sequences in the HP model

h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9
0.794	0.642	0.467	0.456	0.553	0.498	0.526	0.479	0.523

By symmetry, $h_i = h_{19-i}$.

The distribution of total hydrophobicity

Model sequences

We now turn to the distribution of the total hydrophobicity M. Table 2 shows $h = (1 + \langle M \rangle_N/N)/2$ and the normalized variance χ (see Eq. 5) for good HP sequences for $N = 12, \ldots, 18$. Also shown in this table are the two predictions χ_0 and χ_1 defined in Methods, and a prediction χ_2 that will

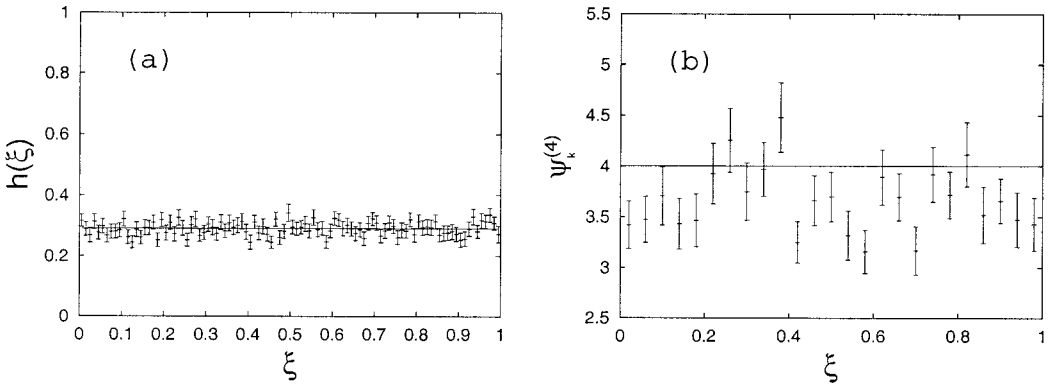

FIGURE 2 (*a*) Hydrophobicity profile $h(\xi)$ for the enzymes. The horizontal line indicates the mean $\bar{h} \approx 0.29$. (*b*) $\psi_k^{(4)}$ as a function of ξ for the enzymes. The horizontal line represents random sequences.

be explained below. Note that h depends quite weakly on N. This implies that the fraction of hydrophobic amino acids, unlike the core to surface ratio of compact chains, does not increase with N. Of course, it would be interesting to see whether this trend persists for much larger N.

From Table 2 we see that χ is smaller than χ_0, which implies that the σ_i values are not both uncorrelated and uniformly distributed. Comparing to χ_1 shows that the major part of this difference is due to correlations rather than non-uniformity. The fact that $\chi < \chi_1$ means that the average c_{ij} $(i \neq j)$ is negative.

The two measurements h and χ are, of course, not enough to fully characterize the distribution of good sequences. To get an idea of how much information they provide, we may compare to the one-dimensional Ising distribution

$$P(\sigma) \propto \exp\left(K_1 \sum_i \sigma_i\sigma_{i+1} + K_2 \sum_i \sigma_i\right). \qquad (13)$$

The measured values of h and χ for good $N = 18$ sequences can be reproduced by choosing $K_1 \approx -0.16$ and $K_2 \approx 0.13$. For these parameters it turns out that $e^S \approx 1.9 \times 10^5$, S being the entropy, which means that the effective number of sequences contained in $P(\sigma)$ is considerably larger than the number of good $N = 18$ sequences, 6349.

Enzymes

To study the N dependence of the total hydrophobicity M for the enzymes, we divide the data set into groups corresponding to different intervals in N. Fig. 4 shows the average M for these groups against N. We see that the N dependence is approximately linear. Although the uncertainties are difficult to estimate, it is interesting to note that the behavior is in perfect agreement with the model results.

Next we calculate $\bar{\chi}$ in Eq. 9, using $\bar{M} = N(2\bar{h} - 1)$ and $\bar{h} = 0.29$, as obtained from a fit to the data in Fig. 4. Table 3 shows $\bar{\chi}$ for all sequences and for the different subgroups described in Methods. We see that $\bar{\chi}$ for all sequences is larger than predicted by Eq. 10, which contrasts sharply with the model results above. We also note that there seems to be a strong dependence on group. In particular there appears to be a big difference between intra- and extracel-

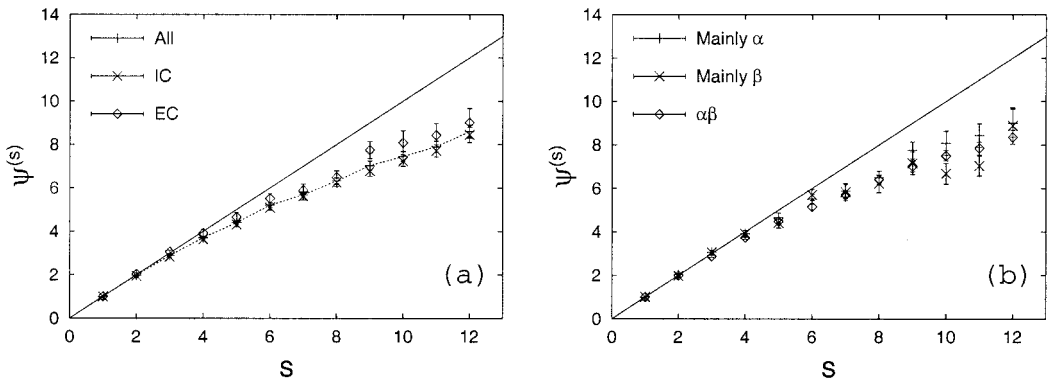

FIGURE 3 The mean-square block fluctuation $\psi^{(s)}$ against block size s for different groups of enzymes. (*a*) All sequences (data points connected by *dashed line*) and intracellular (IC)/extracellular (EC) sequences. (*b*) Division of the sequences into three structural classes: mainly α, mainly β, and $\alpha\beta$. The straight lines represent random sequences.

TABLE 2 $h = (1 + \langle M \rangle_N/N)/2$ **and the normalized variance** χ **of** M **for good HP sequences for different** N

N	h	χ	χ_0	χ_1	χ_2
12	0.527	0.577	0.997	0.913	0.589
13	0.507	0.550	1.000	0.937	0.553
14	0.519	0.684	0.999	0.924	0.688
15	0.556	0.594	0.987	0.959	0.593
16	0.542	0.687	0.993	0.936	0.663
17	0.555	0.695	0.988	0.961	0.639
18	0.548	0.718	0.991	0.949	0.646

Also shown are the three predictions χ_0 (see Eq. 8), χ_1 (Eq. 7), and χ_2 (see Comments).

lular enzymes. However, it must be stressed that the uncertainties are large. Improved statistics are definitely needed in order to draw any firm conclusion about the different groups and possible deviations from the model results.

Comments

Our study of HP sequences has been focused on structure-independent properties. The question of how sequences that share the same (unique) native structure are related has recently been examined using similar models (Li et al., 1996; Bornberg-Bauer, 1997; Bornberg-Bauer and Chan, 1999). From these studies, a simple picture seems to emerge for structures that are highly designable. For high-N_r structures (N_r is the number of sequences that fold to the structure r), it has been found that the sequences tend to form a single cluster connected by one-point mutations, called a "neutral net" (Bornberg-Bauer, 1997), and that structure-preserving mutations tend to be largely independent (Li et al., 1996). The latter property was observed in a model with compact structures only. We checked that it holds in the present model too, which is illustrated in Fig. 5. From this figure it can be seen that the quantities e^S/N_r and $|\Delta\chi^{(r)}|$, as defined in Methods, indeed tend to be small for high N_r. Also indicated in this figure is whether or not the sequences

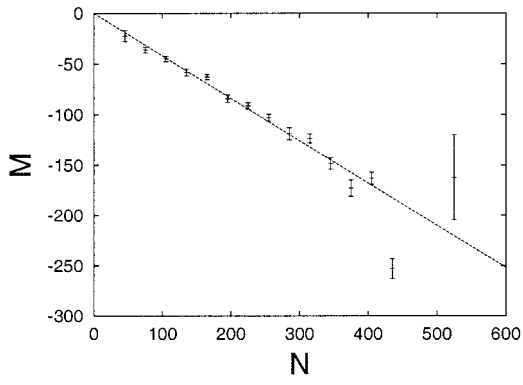

FIGURE 4 Total hydrophobicity M against N for the enzymes. The data points are averages over intervals of length 30 in N. The straight line is a least-square fit.

TABLE 3 Analysis of the fluctuations in M **for the enzymes**

Type of chain	No. sequences	$\bar{\chi}$	$\bar{\chi}_0$
All chains	173	1.50 ± 0.27	0.82
Intracellular	127	0.82 ± 0.13	0.83
Extracellular	46	2.92 ± 1.15	0.78
Mainly α	23	1.45 ± 0.25	0.81
Mainly β	39	1.63 ± 0.34	0.77
$\alpha\beta$	108	0.85 ± 0.14	0.83

The quantities $\bar{\chi}$ and $\bar{\chi}_0$ are defined by Eqs. 9 and 10, respectively.

form a neutral net, results first obtained by Bornberg-Bauer (1997).

The fact that structure-preserving mutations are largely independent for high N_r does not contradict our previous results. To verify this, we calculated χ from the known hydrophobicity profiles $\{h_i^{(r)}\}$ under the assumption that the σ_i values are independent for each structure. The value obtained this way, χ_2, can be found in Table 2 above, and is indeed a relatively good approximation to the observed χ.

Admittedly, the model used in this study is crude. In particular, Buchler and Goldstein (1999, 2000) have recently argued, based on a study of compact lattice chains, that the use of a two-letter alphabet leads to designability artifacts, which disappear with increasing alphabet size. Let us stress, therefore, that the analyses discussed in this paper can be tested on real proteins in a direct manner. Let us also comment on the stability of our results. First, we note that the dependence on chain length N is weak. This was explicitly shown for χ, and is true for $\psi^{(s)}$ too, although our discussion focused on one system size in this case. Second, we note that our results are in nice agreement with those obtained earlier using a simple hydrophobic/polar off-lattice model (Irbäck et al., 1997). To further explore the model dependence of our results, we also did calculations for a "solvation-like" two-letter model discussed by Ejtehadi et al. (1998a,b) and by Buchler and Goldstein (1999, 2000). This model differs from the HP model in that the interaction strength is additive [$\epsilon(H, H) = -2\epsilon$, $\epsilon(H, P) = -\epsilon$ and $\epsilon(P, P) = 0$], which means that the total energy can be expressed as a simple sum of monomer contributions. Buchler and Goldstein argued that HP-like models, unlike pair-contact models with larger alphabets, tend to have solvation-like designability properties. It is therefore interesting to note that when analyzing sequences with unique ground states in the solvation-like model defined above, we obtained results qualitatively different from those for the HP model. More precisely, it turns out that the block fluctuations are significantly larger, close to random, for the solvation-like model.

Summary and Discussion

Hydrophobicity plays a key role in the formation of protein structures, which makes it of utmost interest to understand

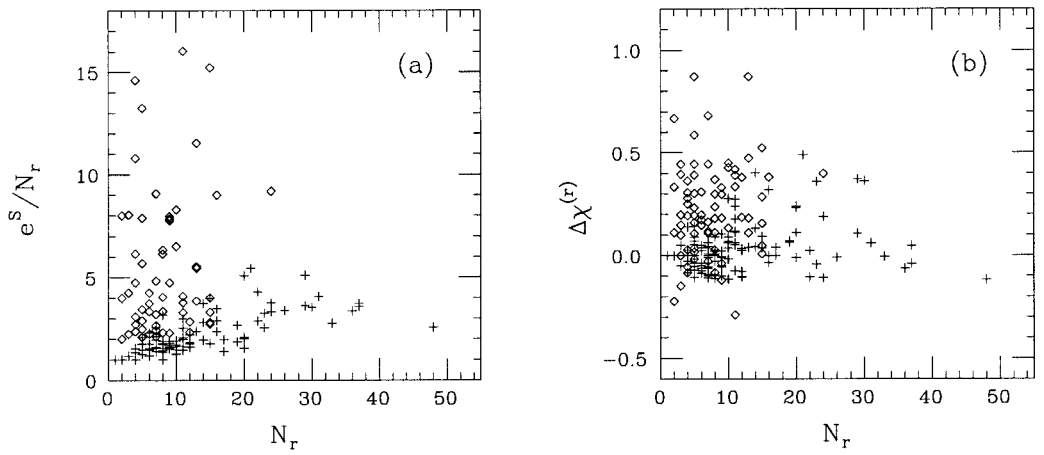

FIGURE 5 (*a*) e^S/N_r, N_r and (*b*) $\Delta\chi^{(r)}$, N_r scatter plots for the 1475 designable $N = 18$ structures in the HP model. The shape of the plot symbol indicates whether the sequences form a neutral net (+) or not (◇).

the statistical distribution of hydrophobicity along the chains. In this paper we have analyzed hydrophobic/polar sequences in the two-dimensional HP lattice model. Whenever statistically feasible, the analogous calculations were performed for a set of real enzymes, too. Our main findings are as follows.

1. Both model sequences and enzymes show mean-square block fluctuations $\psi^{(s)}$ that are smaller than for random sequences. In particular, this implies that the enzymes display the same behavior that had been found previously for general proteins with typical total hydrophobicities (Irbäck et al., 1996). The present analysis was performed without any restriction on total hydrophobicity.
2. The average total hydrophobicity M varies approximately linearly with chain length N over the range of N studied, both for model sequences and enzymes. This implies, contrary to what one naively might expect, that the fraction of hydrophobic amino acids does not grow with increasing N. The fluctuations in M are difficult to study for the enzymes, due to statistical uncertainties. For the model sequences it turns out that the normalized variance χ is significantly smaller than for random sequences.

We also divided the enzymes into different groups according to their structural content, and to whether they reside in an intra- or extracellular environment. The fluctuations in total hydrophobicity appeared to depend on group. However, whether this dependence is significant or not is difficult to say, due to statistical uncertainties. The mean-square block fluctuations are statistically much easier to measure, and show only a weak dependence on group. The conclusion that $\psi^{(s)}$ is suppressed is, in particular, the same for all the different groups.

A full explanation of the suppression of $\psi^{(s)}$ is probably hard to give. Let us note, however, that long hydrophobic or hydrophilic stretches in the amino acid sequence are likely to lead to degenerate structures, and the suppression of sequences containing such stretches should indeed tend to make $\psi^{(s)}$ smaller.

The nonrandomness of the block fluctuations provides an indirect confirmation of the important role played by hydrophobicity in the formation of protein structures. Furthermore, it is tempting to take the similarity with the model results as an indication that the ability to form a stable structure represents a significant selective advantage in the evolution of proteins. It would be interesting to check that the behavior remains the same in more realistic models.

This work was supported by the Swedish Foundation for Strategic Research.

REFERENCES

Bastolla, U., H. E. Roman, and M. Vendruscolo. 1999. Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J. Theor. Biol.* 200:49–64.

Bornberg-Bauer, E. 1997. How are model protein structures distributed in sequence space? *Biophys. J.* 73:2393–2403.

Bornberg-Bauer, E., and H. S. Chan. 1999. Modeling evolutionary landscapes: mutational stability, topology and superfunnels in sequence space. *Proc. Natl. Acad. Sci. USA.* 96:10689–10694.

Broglia, R. A., G. Tiana, H. E. Roman, E. Vigezzi, and E. Shakhnovich. 1999. Stability of designed proteins against mutations. *Phys. Rev. Lett.* 82:4727–4730.

Bryngelson, J. D., J. N. Onuchic, N. D. Socci, and P. G. Wolynes. 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins Struct. Funct. Genet.* 21:167–195.

Buchler, N. E. G., and R. A. Goldstein. 1999. Effect of alphabet size and foldability requirements on protein structure designability. *Protein Struct. Funct. Genet.* 34:113–124.

Buchler, N. E. G., and R. A. Goldstein. 2000. Surveying determinants of protein structure designability across different energy models and amino-acid alphabets: a consensus. *J. Chem. Phys.* 112:2533–2547.

Chan, H. S., and K. A. Dill. 1994. Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys.* 100:9238–9257.

Creighton, T. E. 1993. Proteins: Their Structure and Molecular Properties. Freeman, New York.

Dill, K. A., S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. 1995. Principles of protein folding: a perspective from simple exact models. *Protein Sci.* 4:561–602.

Ejtehadi, M. R., N. Hamedani, H. Seyed-Allaei, V. Shahrezaei, and M. Yahyanejad. 1998a. Stability of preferable structures for a hydrophobic-polar model of protein folding. *Phys. Rev. E.* 57:3298–3301.

Ejtehadi, M. R., N. Hamedani, H. Seyed-Allaei, V. Shahrezaei, and M. Yahyanejad. 1998b. Highly designable protein structures and inter-monomer interactions. *J. Phys. A.* 31:6141–6155.

Govindarajan, S., and R. A. Goldstein. 1997a. Evolution of model proteins on a foldability landscape. *Proteins: Struct. Funct. Genet.* 29:461–466.

Govindarajan, S., and R. A. Goldstein. 1997b. The foldability landscape of model proteins. *Biopolymers.* 42:427–438.

Irbäck, A., C. Peterson, and F. Potthast. 1996. Evidence for nonrandom hydrophobicity structures in protein chains. *Proc. Natl. Acad. Sci. USA.* 93:9533–9538.

Irbäck, A., C. Peterson, and F. Potthast. 1997. Identification of amino acid sequences with good folding properties in an off-lattice model. *Phys. Rev. E.* 55:860–867.

Irbäck, A., and E. Sandelin. 1998. Local interactions and protein folding: a model study on the square and triangular lattices. *J. Chem. Phys.* 108:2245–2250.

Khokhlov, A. R., and P. G. Khalatur. 1998. Protein-like copolymers: computer simulation. *Physica A.* 249:253–261.

Khokhlov, A. R., and P. G. Khalatur. 1999. Conformation-dependent sequence design (engineering) of *AB* copolymers. *Phys. Rev. Lett.* 82:3456–3459.

Klimov, D. K., and D. Thirumalai. 1998. Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics. *J. Chem. Phys.* 109:4119–4125.

Lau, K. F., and K. A. Dill. 1989. A lattice statistical model for the conformational and sequence spaces of proteins. *Macromolecules.* 22: 3986–3997.

Li, H., R. Helling, C. Tang, and N. Wingreen. 1996. Emergence of preferred structures in a simple model of protein folding. *Science.* 273:666–669.

Martin, A. C. R., C. A. Orengo, E. G. Hutchinson, S. Jones, M. Karamirantzou, R. A. Laskowski, J. B. O. Mitchell, C. Taroni, and J. M. Thornton. 1998. Protein folds and function. *Structure.* 6:875–884.

Nymeyer, H., A. E. García, and J. N. Onuchic. 1998. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci. USA.* 95:5921–5928.

Orengo, C. A., A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. 1997. CATH: a hierarchic classification of protein domain structures. *Structure.* 5:1093–1108.

Pande, V. S., A. Y. Grosberg, and T. Tanaka. 1994. Nonrandomness in protein sequences: evidence for a physically driven stage of evolution? *Proc. Natl. Acad. Sci. USA.* 91:12972–12975.

Peng, C.-K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley. 1992. Long-range correlations in nucleotide sequences. *Nature.* 356:168–170.

Săli, A., E. Shakhnovich, and M. Karplus. 1994. Kinetics of protein folding: a lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* 235:1614–1636.

Tiana, G., R. A. Broglia, and E. I. Shakhnovich. 2000. Hiking in the energy landscape in sequence space: a bumpy road to good folders. *Proteins Struct. Funct. Genet.* 39:244–251.

White, S. H., and R. E. Jacobs. 1990. Statistical distribution of hydrophobic residues along the length of protein chains: implications for protein folding and evolution. *Biophys. J.* 57:911–921.