

# How Are Model Protein Structures Distributed in Sequence Space?

Erich Bornberg-Bauer

Abteilung 0815 Theoretische Bioinformatik, Deutsches Krebsforschungszentrum Im Neuenheimer Feld 280, Heidelberg, D-69120, Germany

**ABSTRACT** The figure-to-structure maps for all uniquely folding sequences of short hydrophobic polar (**HP**) model proteins on a square lattice is analyzed to investigate aspects considered relevant to evolution. By ranking structures by their frequencies, few very frequent and many rare structures are found. The distribution can be empirically described by a generalized *Zipf's law*. All structures are relatively compact, yet the most compact ones are rare. Most sequences falling to the same structure belong to "neutral nets." These graphs in sequence space are connected by point mutations and centered around *prototype sequences*, which tolerate the largest number (up to 55%) of neutral mutations. Profiles have been derived from these homologous sequences. Frequent structures conserve hydrophobic cores only while rare ones are sensitive to surface mutations as well. *Shape space covering*, i.e., the ability to transform any structure into most others with few point mutations, is very unlikely. It is concluded that many characteristic features of the sequence-to-structure map of real proteins, such as the dominance of few folds, can be explained by the simple **HP** model. In analogy to protein families, nets are dense and well separated in sequence space. Potential implications in better understanding the evolution of proteins and applications to improving database searches are discussed.

## INTRODUCTION

Understanding how evolution has shaped today's biopolymers is of fundamental interest for both characterizing the biophysical processes during early evolution and developing strategies to design functional molecules with desired properties. Evolving entities must in principle "accomplish" two tasks: to *conserve* acquired features in their genotype and to *adapt* to new requirements on the phenotype level. Since there is a tradeoff between these tasks it is crucial to understand the principles of the genotype phenotype relation. Theoretical concepts, developed in the 1930s by S. Wright (1932) and others, proposed the concept of *fitness landscapes*. In this picture, evolution is viewed as a walk over the set of genotypes preferring "fitter" offspring. This is done by selecting for some functional criterion which is a phenotype property. Later considerations emphasized the importance of phenotypically neutral mutations (Kimura, 1968; King and Jukes, 1969). Maynard-Smith (1970) assumed that, since only few mutations can be advantageous for the phenotype, a continuous gradient of fitness must be maintained so that mutated offspring survive. Applied to biopolymers, this implies that residues that are essential for function will be conserved, and others replaced in evolutionary diverse sequences. Unfortunately, it is difficult to define fitness a priori. It is, however, generally assumed that function largely depends on structure. Consequently, it can be more intriguing to study appropriate details of the se-

quence-to-structure map. In principle, the structure prediction problem is of comparable complexity for proteins and RNA. Therefore we were motivated by recent success in characterizing the sequence-to-structure map for RNA by statistical analysis of a simplified model, the secondary structure model (Fontana et al., 1993; Schuster et al., 1994; Tacker et al., 1996; Huynen et al., 1996; Bornberg-Bauer, 1996). However, the sequence structure relation is more involved for proteins: first, in contrast to RNA, real proteins do not unify genotype and phenotype in one molecule; second, as a consequence, there are two sources of neutrality: the redundancy of the genetic code (i.e., at the *genotype level*), and some structural robustness of folding (i.e., at the *phenotype level*); finally, there is a serious computational problem: simplified structure representations allowing for folding strategies with polynomial complexity are not available. Consequently investigations remain restricted to chains that are short compared to those possible when analyzing RNA secondary structures.

For proteins, several simplified approaches such as lattice models (Lau and Dill, 1989; Shakhnovich and Gutin, 1990; Skolnick and Kolinski, 1990), spin glass analogies (Bryngelson and Wolynes, 1987), and others (Dill et al., 1995) were developed during the last decades to investigate basic principles that govern protein folding. Most of these strategies address the question of how *single* molecules fold (Skolnick and Kolinski, 1990; Chan and Dill, 1994; Sali et al., 1994; Onuchic et al., 1995; Abkevich et al., 1995). Often they are based on mean field models or stochastic optimization. They are computationally intensive and therefore not applicable to investigate ensembles that are large enough to characterize the sequence-to-structure map. Recent studies on the *foldability* landscape of a cubic lattice model predicted a considerable amount of neutrality in *interaction space*, and that similar structures and structures with similar

Received for publication 18 March 1997 and in final form 18 August 1997.

Address reprint requests to Dr. Erich Bornberg-Bauer, Abteilung 0815 Theoretische Bioinformatik, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, Heidelberg, D-69120, Germany. Tel.: 49-6221-42-2723; Fax: 49-6221-42-2849; E-mail: bornberg@dkfz-heidelberg.de, erich@santafe.edu.

© 1997 by the Biophysical Society

0006-3495/97/11/2393/11 \$2.00

optimal foldabilities to cluster together (Govindarajan and Goldstein, 1997b). Though not an explicit fitness criterion, the ability to fold fast to a unique and thermodynamically stable state provides, to some extent, a reasonable prerequisite to yield a viable biopolymer under *in vitro* folding conditions. Such an evolutionary process has been studied as a random walk through the foldability landscape (Govindarajan and Goldstein, 1997a).

Since we are interested in the properties of the sequence-to-structure map and features that are important to understand evolutionary adaptation of structures, this study is based on short chains of a simple lattice model, the **HP** model.

## HP-LATTICE PROTEINS

The **HP** model (Lau and Dill, 1989) is one of the best investigated (see Dill et al., (1995) and refs. therein) and assumes the hydrophobic effect to be the major structure determining effect. It can be viewed as a very coarse-grained model that, since it depends on very few parameters, provides us with a simple and nonambiguous framework. All residues have the same size and the peptide chain is constructed by placing residues sequentially on the non-occupied beads of a regular lattice. The resulting chain has identical bond lengths and discrete bond angles. We use relative moves for handling structures: the structure is represented as a non-self-intersecting *self-avoiding walk* on a regular lattice and the movement of the chain as a sequence of moves where each is encoded relative to the prior. The method is well known [see, e.g., Lau and Dill (1989)]; this version has been adapted to apply to any regular lattice [an example is given in Fig. 1; for a detailed description see Bornberg-Bauer (1997)]. It is versatile and provides convenient computational techniques for handling, comparing,

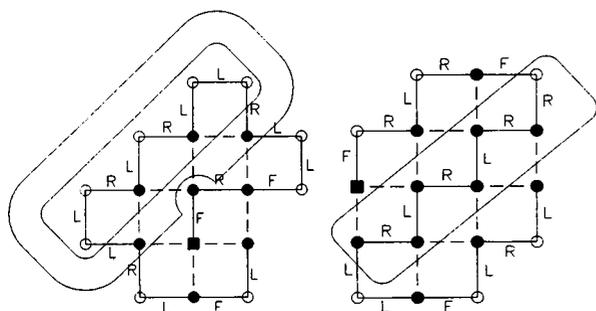


FIGURE 1 Examples of frequent structures. *Left*: The most frequent structure as formed by a typical sequence. Closed circles denote **H**'s, open circles **P**'s, solid lines correspond to peptide bonds connecting two subsequent residues, dashed lines are energy-contributing contacts between 2 **H**'s. Letters along the bonds [F (*forward*), L (*left*), and R (*right*)] denote the corresponding relative moves. Squares symbolize the first residue since the structure is not considered identical as the results from reverse sequences. The first move is F by definition, the first non-F move R. The structure can be encoded as FRLLRLLRLRLLRLFL. Frequent motifs are boxed (see text). *Right*: The most frequent maximum compact structure. It can be encoded as FRLRFRRLRLRLLFLRL.

and storing data. Structures will be referred to as identical if and only if they are represented by the same walk on the lattice, i.e., if their "sequences" of relative moves are identical. The *shape space*  $\mathcal{X}$  is represented by the set of all possible self-avoiding walks. Our notation for encoding moves is insensitive to translational and rotational symmetries. Mirror symmetries are disallowed by choosing the first non-F move to be an R (see Fig. 1). Mirror symmetric sequences, corresponding to reverse sequences, refer to molecules with different chemical properties and are therefore distinguished.

The Hamming distance,  $h_{1,2} = h[S_1(n), S_2(n)]$ , is defined as the minimum number of point mutations required to convert one sequence  $S_1$  into another  $S_2$  of equal length  $n$  while insertions and deletions are ignored. While certainly not sufficient to describe all kinds of evolutionary relationships of *real* protein sequences of variable length, it is well suited for lattice proteins (where, e.g., L (leucine) and W (tryptophan) are not distinguished by size) and, in particular, for the **HP** model (where both L and W are represented by an **H**). It thus provides us with a metric in *sequence space*  $\mathcal{S}$ .

A key assumption in the **HP** model is the dominance of the *hydrophobic force* for the overall stability that, to a large extent, determines the spatial structure of the backbone. The importance of the hydrophobic-polar pattern in the sequence for the structure is supported by theoretical studies evaluating potentials (Huang et al., 1995), derivation of empirical potentials (Casari and Sippl, 1992), and the properties of mutation matrices (Koshi and Goldstein, 1997). Much experimental evidence also supports this view, such as rational design based on the careful generation of **HP** patterns (Kamtekar et al., 1993), the finding that certain proteins fold noncooperatively to a natively like state in the absence of packing interactions (Schulman and Kim, 1996), and investigations of folding random sequences composed from small alphabets (Davidson et al., 1995; Cordes et al., 1996). Side chain packing in the **HP** model may then allow for selected structures within this relatively small set of possible coarse-grained structures. Heteropolymers are composed from a two-letter alphabet,  $\mathcal{A} = \{\mathbf{H}, \mathbf{P}\}$ , with only one stabilizing interaction if and only if hydrophobic residues (**H**) are neighbors on the lattice but not along the chain. For contacts (**HH**, **HP**, **PP**) the potential then takes the form  $\mathcal{E} = (-1, 0, 0)$ . The energy function (e.g., on a square lattice) is simply the negative sum of all **HH** contacts. It has been termed a *correlated* folding code (Chan and Dill, 1996), which describes "the physical requirement that all instances of an interaction between residue types  $i$  and  $j$  must have the same energy"; i.e., values of interactions depend exclusively and completely on the nature of the residues.

**HP** sequences  $S$  in general show a large structural degeneracy [ $g_0(S) \gg 1$ ] i.e., the number of configurations that correspond to one lowest energy state  $e_0(S)$ . For  $n = 18$  on a square lattice, only 2.4% of all chains fall into a unique ground state  $g_0(S) = 1$ , all others to a gemisch of multiple lowest-energy structures (Dill et al., 1995). Clearly enough

the model is only a crude abstraction of a realistic protein since hydrogen bonds, disulfide bonds, and electrostatic interactions such as salt bridges are neglected. Yet the most salient features of real protein structures are retained: the hydrophobic effect comprises solvent-driven collapse to a native state, chains have much conformational freedom, and the self-avoiding walk constraint accounts for steric restrictions (*excluded volume effect*).

It is often assumed that only the sequence determines the native state of proteins, which is either the minimum free energy state or a structurally related very low-energy state. Therefore, mostly ground states of uniquely folding short sequences on a square lattice are studied.

Following the ideas of Maynard-Smith (1970), Lipman and Wilbur (1991) studied the influence of neutral mutations in the HP model to investigate evolutionary pathways in sequence space. They used a structure notion that is based on contact maps and confined their study to very compact states. They found that many sequences are linked by neutral neighbors and belong to large connected networks (see next section). They concluded that neutral mutations play an important role for the transition from one structure to another and for efficiently exploring shape space.

## RESULTS

### Convergence

First we investigate the ensembles for the occurrence of *convergent* sequences (Chan and Dill, 1991b), i.e., different sequences  $S_i$  that fold uniquely into the same structure  $X$ . We encoded structures by relative moves and computed the frequencies of occurrence for individual structures. For the  $2^{18}$  possible sequences in the complete sequence space  $\mathcal{S}(18)$  for length  $n = 18$  and a binary alphabet there are 6349 (2.4%) sequences  $S_i$  (29 symmetric ones and 3160 nonsymmetric ones) that fold to a unique structure. They assume 1475 different structures  $X$ . We sort structures, count their occurrences, and rank them with decreasing frequencies. The most frequent structure  $X^1$  is assigned rank 1, the next frequent one  $X^2$  rank 2, and so forth. Structures of equal frequencies  $F$  are sorted and ranked lexicographically. In the following let  $X^r$  denote the structure with rank  $r$ . The set of all  $F^r$  sequences  $S$  converging into one  $X^r$  will be termed a *neutral set*  $\mathcal{X}^r = (X^r; S_1^r, S_2^r, \dots, S_{F^r}^r)$  hereafter. Following earlier work on RNA secondary structures (Schuster et al., 1994; Bornberg-Bauer, 1996; Tacker et al., 1996) we represent this distribution in a  $\log_{10} - \log_{10}$  plot. Results for  $\log(\tilde{F}^r)$  vs.  $\log(r)$ , where  $\tilde{F}^r = F^r/|\mathcal{X}^r| = F^r/6349$  denotes the relative frequency of a structure, are shown in Fig. 2.

We find frequencies following a characteristic distribution: there are few very frequent and many rare structures. The distribution can be *empirically* approximated by a generalized Zipf's law (Gonnet and Baeza-Yates, 1991) as

$$F^r(r) = a(r + b)^{-c}. \quad (1)$$

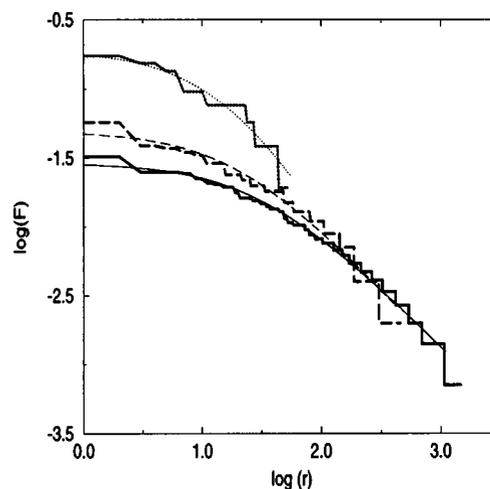


FIGURE 2 The frequency distribution of structures. Zipf plot showing the log of the frequency distribution of structures versus the log of their rank  $r$ . Results for ground states of uniquely folding sequences are shown for chain lengths  $n = 13$  (dotted), 16 (dashed), and 18 (solid line). Corresponding fits to a generalized Zipf's law are drawn in thin lines.

Here  $r$  is the rank and  $F^r(r)$  is the frequency of the corresponding structure,  $a$  is a suitable normalization constant, and  $b$  can be interpreted as the number of "very frequent" structures. For comparison we also show the distribution for lengths  $n = 13, 16$ , which are qualitatively the same. Parameters for the best fits are given in Table 1.

While  $b$  naturally increases with  $n$ , the exponent  $c$  decreases with  $n$ . This means that, in the limit of longer sequences the distribution of structures may become more even. This type of distribution has been reported for a large number of natural systems although a generally applicable causal theory has not been given to date. It was reported as an empirical law for the gap size distribution when comparing sequences (Benner et al., 1993). In a recent work Zipf analysis of tuple frequencies was used to demonstrate the "statistical linguistic qualities" to estimate the information contents in protein sequences and to show that they have a significantly different behavior from random sequences (Strait and Dewey, 1996). Furthermore, it can be found in words distribution of text (a typical value for  $c$  in most natural languages is close to 1) and there has been much attention to the observation of such behavior in DNA. However, expressions of this form can also be derived from certain Markov processes that show a similar curvature [see Czirik et al. (1995) and Strait and Dewey (1996) and references therein for a detailed analysis].

TABLE 1 Parameters for best fits to Zipf's law for data from Fig. 2

$n$	$a$	$b$	$c$
13	17.4	18.40	1.5
16	0.83	19.77	0.94
18	0.52	28.95	0.86

In the remainder of this work sequence space  $\mathcal{S} := \mathcal{S}(18)$  of uniquely folding sequences except when indicated otherwise.

An obvious distinction between rare and frequent structures can be defined by the average size of neutral sets  $\bar{F} = |\mathcal{S}|/|\mathcal{X}| = 6349/1475 = 4.3$ . Consequently, sets with  $F^r \geq 5$  [i.e., with rank  $r < 418$  (28%)] are termed frequent. Nets with  $F^r \geq F^1/\bar{F} = 48/4.3 = 11.2$  (i.e., with  $r \leq 102$ ) will be defined as very frequent and all others defined as rare (where  $F^1 = 48$ ).

It is commonly believed that proteins assume very compact shapes. However, the degree of compactness is not always maximal, neither in real nor in model proteins (Yue et al., 1995; Dill et al., 1995; Goodsell and Olson, 1993). It is, therefore, remarkable to note that the most frequent maximum compact structure (MCS) with 10 contacts is not among the top frequent structures: it has rank  $r = 50$ , occurring only 17 times (i.e.,  $F^{50} = 17$ ). From 1475 structures only 331 (22%) assumed by 1142 (18%) sequences are MCSs, with 930 having 9 contacts and 214 having 8. This means that 20% of all 1673 possible MCSs (Chan and Dill, 1991a), can be found in the structure ensemble, compared to 6% of all structures with 9 contacts and 0.5% of those with 8. Still, MCSs can be realized by only a relatively small number of sequences in the "simple" HP model (Lau and Dill, 1989).

## Examples

In Fig. 1 the most frequent structure is shown. It is uniquely coded by 48 sequences. The structure has a well defined hydrophobic core and a regular motif with hydrophobic residues oriented to the interior and polar ones directed to the outside. It is obvious that in such a simple model, where the length corresponds to the lower limit of stable chains of real proteins, it is difficult to define and identify secondary structure elements. The motif LLRLRLL appears in 988 (16%) of all structures. The subfragment LRLRL appears in 1254 (11%) structures (see Fig. 1). These motifs show some regularity in the "sequence" of relative moves. This notion is different from earlier classifications of secondary structure elements that were solely based on patterns in the contact map (Chan and Dill, 1991b; Li et al., 1996), yet can be viewed as an appropriate analog to an  $\alpha$ -helix providing a hydrophobic, stabilizing bulk oriented to the inside and polar residues at the outside.

## The topology of neutral sets

Pondering the origin of this biased distribution, the relationship between sequences within some neutral sets in more detail is investigated.

The first investigation is the connectedness of neutral sets, i.e., to what extent sequences in the  $\mathcal{X}^r$  are linked by single point mutations. Each entirely linked subset is called a neutral component  $\mathcal{C}^r$  of  $\mathcal{X}^r$ . If two sequences  $S_i^r$  and  $S_j^r$  that

converge have  $h(S_i^r, S_j^r) = 1$ , they are called *neutral neighbors* and said to differ by a *neutral mutation*. The overwhelming majority of 511 sets (80% of all 684 sets with  $F^r \geq 3$ ) are represented by a single  $\mathcal{C}^r$ . If a neutral set consists of exactly one connected component, such a component will be termed a *neutral net*  $\mathcal{N}^r$  hereafter. Its sequences are all related and termed *homologous*. From all 684 networks with  $48 \geq F^r > 2$  there are 4  $\mathcal{X}^r$ s with 4  $\mathcal{C}^r$ s, 13 with 3 and 156 with 2. The most frequent structure with more than one component has rank 27. No other significant correspondence between the frequency of a structure and the number of components was found. As was already stated by Lipman and Wilbur (1991), the case of different components  $\mathcal{C}^r$ s in one set  $\mathcal{X}^r$  "is tantamount to convergent evolution in that members from one [set  $\mathcal{X}^r$ ] but different [components  $\mathcal{C}^r$ ] must have converged to an identical structure from different ancestries." They also claimed that the length of connected nets increases with the sequence length.

In the following, some characteristic properties of the topology of selected neutral sets are investigated. Results are summarized in Table 2.

All selected sets  $\mathcal{X}^r$  are nets  $\mathcal{N}^r$ . They belong to compact shapes, yet only  $\mathcal{X}^{50}$  is an MCS. Corresponding energies  $e_j$  vary in a range between lowest ground state energy  $e_0(S^r)$  and  $e_0(S^r) + 2$ . The distribution of numbers of neutral neighbors for all homologous sequences within each  $\mathcal{N}^r$  is investigated. It is found that  $h^+$ , the maximum  $h_{ij}$  between any two sequences in  $\mathcal{N}^r$  and  $\bar{h}$ , the average  $h$  between all pairs of sequences within  $\mathcal{N}^r$  increase with the frequency of the structure (i.e., decrease with  $r$ ). The overall maximum  $h^+$  of all nets  $\mathcal{N}$  corresponds to  $r = 1$  (i.e., the most frequent structure) and is equal to 7. This means there are two sequences in  $\mathcal{N}^1$  that are distinct at 7 (of 18 possible) positions.

Next, the number of neutral mutations within a  $\mathcal{N}^r$  and the average number of neutral neighbors per sequence in the net are counted. Again these values increase with lower  $r$ . This means that nets corresponding to more frequent structures are more extended as well as more clustered (denser). An interesting phenomenon is the apperency of a single sequence  $S_i^r$  with an extraordinarily large number of neutral neighbors in  $\mathcal{X}^r$ . We will call this a *prototype sequence*  $\hat{S}^r$  of a neutral set  $\mathcal{X}^r$ , since it is extremely stable toward muta-

**TABLE 2** Some characteristic properties for selected networks  $\mathcal{N}^r$

Rank of Neutral Set	1	14	50	120	910
Size of neutral set	48	30	17	11	2
Compactness (contacts)	9	9	10	9	9
Energies (from, to)	7-9	7-9	8-10	7-9	7
Maximum $h$	7	7	6	4	1
Average $h$	3.19	3.28	2.50	2.20	1.00
Neutral neighbors (overall)	99	53	28	14	1
Neutral neighbors (average)	2.1	1.8	1.6	1.3	1
Number of neutral components	1	1	1	1	1
Neutral neighbors (of $\hat{S}^r$ )	10	6	6	5	1
Number of $\hat{S}^r$	1	2	1	1	1

tions. Prototype sequences are, as can be seen in the following, in general identical to the consensus sequence of the homologous sequences, which is defined as  $\bar{S} = (\bar{s}_1, \bar{s}_2, \dots, \bar{s}_n)$ , where for each position  $i$  the residue with the highest probability  $\bar{s}_i$  is chosen. It is certainly remarkable that for prototype sequence  $\hat{S}^1$ , more than half of all the residues may be mutated and up to four mutations at the same time can be applied without altering the structure under the condition that the sequence folds uniquely. This is most striking, if one considers that real proteins are composed of a 20-letter alphabet with  $\sim 10$  letters in either class, **H** and **P**. In a natural alphabet it requires therefore two random mutations to exchange, on average, a hydrophobic into a polar residue, compared to only one mutation in a simple **HP** alphabet, four mutations compared to two, etc. Most neutral sets of very frequent structures and all from rank one to 12 can be characterized by a single prototype sequence  $\hat{S}^r$ , most others by two or three. One,  $X^{30}$  with  $F^{32} = 29$ , has 13 prototypes with three neutral neighbors.

In the following, neutral nets are investigated by inspecting properties of  $\hat{S}^r$  and their neighboring sequences in detail, which was done for the selected structures from before. Examples for the  $\hat{S}^r$ s and a sketch of some corresponding nets  $\mathcal{N}^r$  are shown in Fig. 3.

The prototype sequences  $\hat{S}^r$  are shown in their native structure. Arrows denote possible neutral mutations to  $\hat{S}^r$  and the regions that are completely conserved in the whole net are boxed. Selected corresponding nets are shown below with dots symbolizing single sequences, the  $\hat{S}^r$  as a larger dot in the center, 1-error neutral mutants are arranged in the next "shell" around it, 2-error mutants in the second "shell," and so forth. The complete  $\mathcal{N}^r$ s are visualized by the com-

binations of paths that correspond to neutral mutations and start from the innermost shell.

For  $\hat{S}^1$ , 8 of 10 neutral mutations are **P**→**H** substitutions on the surface, yet there are 2 **H**'s, at positions 3 and 16, respectively, that can be subjected to neutral mutations. As a general rule we find that, for very frequent structures to prevail, the hydrophobic core must remain largely unaltered.

The most frequent minimum free energy MCS,  $\mathcal{N}^{50}$ , shows a remarkable symmetry in both sequence and structure. The core motif **RLRLR** appears in the center and requires all residues to be **H**'s. The energies are in the range of  $-8$  to  $-10$ . The corresponding network is shown in Fig. 3. Again, only surface-exposed positions are affected and the network consists of paths representing combinations of permutations of the six mutable positions.

### Profiles

Next, the mutational flexibility at given positions among the homologous sequences within a  $\mathcal{N}^r$  is characterized. We generate profiles, i.e., we calculate the probability  $p_i(H)$  to find an **H** at sequence position  $i$  [note that  $p_i(P) = 1 - p_i(H)$ ].

While very frequent structures require highly conserved residues at given positions for **H**'s only, less frequent structures are, in general, also sensitive to mutations at positions occupied with **P**'s in the prototype sequence. This can be attributed to the property of "designing out" (i.e., avoiding) alternate conformations. This becomes more difficult for certain geometries and thus allows for a smaller number of possible sequences (Yue et al., 1995; Dill et al., 1995). Examples are shown in Fig. 4 for  $X^1$  and for  $X^{120}$ .

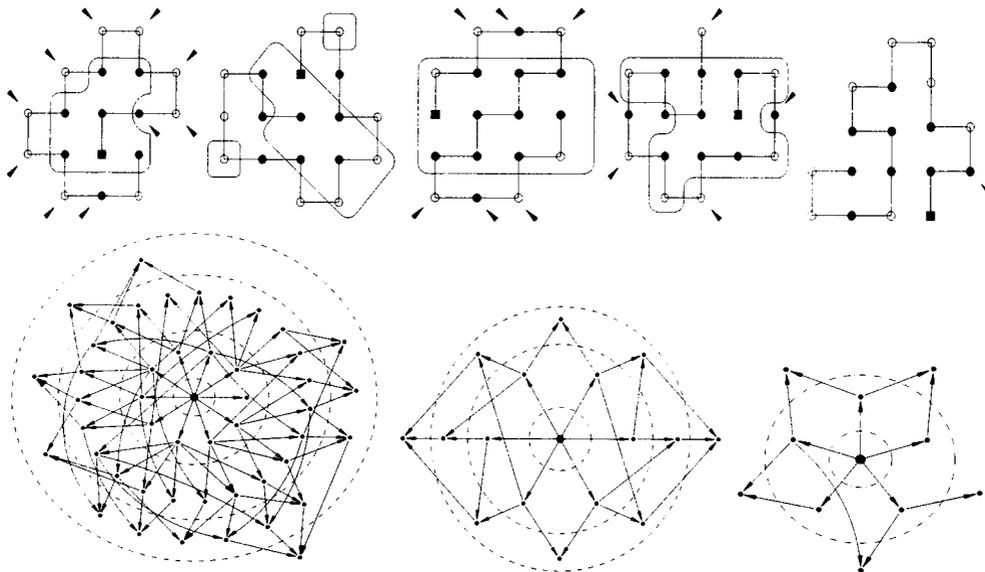


FIGURE 3 Mutational stability, prototype sequences, and neutral nets. *Upper line:* prototype sequences  $\hat{S}^r$ ,  $r = 1, 14, 50, 120, 910$  in their native structure. Arrows point to positions where neutral mutations are possible. Regions that are completely conserved in all homologous sequences are boxed. (For explanation of additional symbols, see Fig. 1.) *Lower line:* corresponding neutral networks  $\mathcal{N}^r$ ,  $r = 1, 50, 120$ . The "shells" of sequences with Hamming distances  $h = 1, 2$ , and  $3$  to the  $\hat{S}^r$  are indicated by three circles (*dot-dashed*). Each arrow corresponds to a single point mutation.

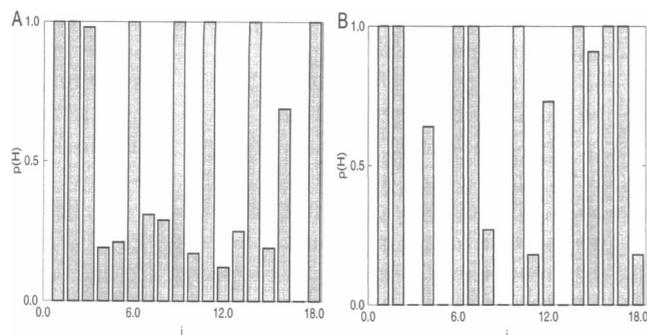


FIGURE 4 Profiles of homologous sequences in selected neutral sets. (A) Profile of  $X^1$ . (B) profile of  $X^{120}$ .

This holds true for most other nets. In all examples with a unique prototype sequence this prototype sequence is identical to the consensus sequence.

### Distribution of neutral sets in sequence space

A crucial issue for evolutionary adaptation is the possibility of transforming one structure into another. It is therefore important to know how far apart in sequence space sequences belonging to different neutral nets  $\mathcal{N}^r$  are. We therefore inspect the distribution of pairwise sequence distances  $h$  in between the  $\mathcal{N}^r$ 's. Fig. 5 shows the probability that two sequences that fold uniquely into two different structures have a given  $h$ .

For comparison, the distribution of Hamming distances that is obtained between two randomly chosen sequences (which is given by the binomial distribution) is shown. Compared to random a slight deviation toward a smaller pair distance and a slightly broader distribution is found. The former is a result of the fact that H-rich sequences are overrepresented. The distribution is shown in the same plot versus a horizontal axis that represents the number of H residues in a sequence. The broadening hints to a slight degree of clustering. Roughly speaking, however, distances

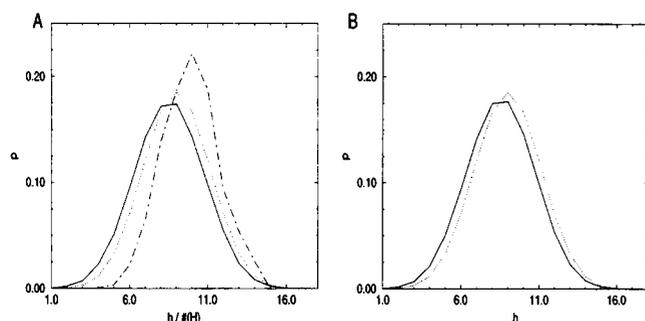


FIGURE 5 Pair correlation of Hamming distances  $h$  between structures in sequence space. The random distribution, following the binomial distribution (dotted), is shown for comparison. (A) Probability values  $p$  between full neutral nets (solid line). The probabilities  $p$  that sequences have a given number of H-residues (hydrophobicity, (H)) are shown in the same plot (dashed line). (B) Same as A but for prototype sequences.

are randomly distributed. This means that, although only a small fraction of sequence space yields uniquely folding sequences, sequence space is occupied nearly uniformly. No "higher order" clustering (i.e., except the trivial case of the homologous sequences) is visible.

The next area of interest was the number of mutations that are needed to transform a sequence that folds uniquely in a given "reference" structure into another "target" sequence which folds uniquely into a different structure. It is meaningful to understand how likely it is that new structures evolve through series of point mutations from existing precursor structures. Fontana, Schuster, and co-workers (Fontana et al., 1993; Schuster et al., 1994) introduced the notion of *shape space covering* for the RNA secondary structure case. It is assumed that, to enable fast adaptation of biopolymers and starting from any given initial (reference) structure, it should be advantageous to reach any typical (i.e., frequent) structure within a relatively small number of mutations. Together with the extension of neutral nets this can be viewed as an important measure of how fast evolutionary optimization may search sequence space. For this concept it is crucial to recall that evolution acts on populations. Individuals may have the same phenotype but genotypes are, depending on the error rate during reproduction, more or less scattered around a consensus sequence. Populations may evolve along neutral nets. When they come "close" in  $\mathcal{S}$  to another net with higher fitness, single individuals may "jump" to this net and reproduce there more efficiently. This was shown to be the case for RNA secondary structures in a series of computer experiments (Huynen et al., 1996).

Results so far suggest that shape space covering is very unlikely, since networks are localized and, on average, well separated. Still this does not exclude the possibility that nets are "interwoven," i.e., that single sequences of most nets are close in sequence space to some exposed sequences of many other nets. We therefore compute the Hamming distances between any two sequences that fold into different structures. The minimum Hamming distances  $h$  that are found between two neutral sets are remembered and the number of these instances summed up for each  $h$ .

In Fig. 6 the fraction of cumulatively covered targets for selected reference nets is shown. Covering distances are reported for the same nets as were shown in the former sections. Results are shown for computations when prototype sequences only are considered or when the complete neutral set was used to obtain reference sequences. For comparison, the covering distances when allowing for target sequences with degeneracy  $g_0(S) \leq 6$  (i.e., sequences that fold to the target structure but not uniquely) is reported. It is interesting to note that covering efficiency is primarily enhanced when considering full neutral nets instead of prototype sequences and not so much between the different ranks (except  $X^{910}$ ). Also, admitting degenerately folding sequences enhances the covering ability only marginally. Obviously prototype sequences of frequent structures are well "shielded" from their environment by the large number of neutral (or nearly neutral) neighbors. For the most fre-

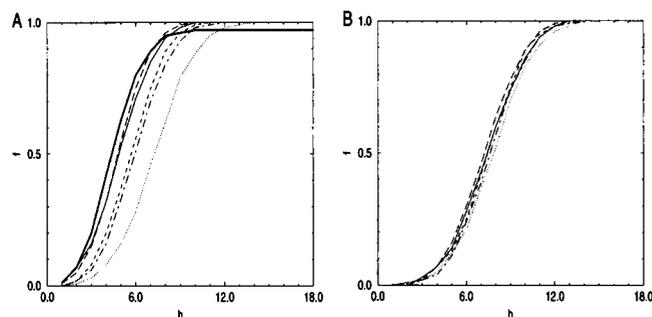


FIGURE 6 Shape space covering: fraction of cumulatively observed structures versus  $h$  for the ranked structures (see text for description). (A) For complete nets  $X^1$  (solid line),  $X^{14}$  (long dashed line),  $X^{50}$  (short dashed line),  $X^{120}$  (dot-dashed line), and  $X^{910}$  (dotted line). The leftmost (thick) line shows the covering fraction when degenerately folding sequences ( $g = 6$ ) are considered. (B) Same as A but for prototype sequences only.

quent structure it takes only two mutations to cover 10% of all structures, but it takes approximately five mutations to cover 50%. It requires at least nine selected mutations in nine different positions each (i.e.,  $\mathbf{H} \rightarrow \mathbf{P}$  or  $\mathbf{P} \rightarrow \mathbf{H}$ ) to find every structure. In spite of the great mutability within a  $\mathcal{N}^r$  it is very difficult to transform a structure into another.

### The correspondence with foldability

Foldability, i.e., the ability to rapidly reach a native state, was repeatedly claimed to be an important feature of proteins. It was shown to correspond directly to an energy gap between the ground state with  $e_0(S_i)$  and other “excited” states  $e_1(S_i) = e_0(S_i) + 1$ ,  $e_2(S_i) = \dots$  (Bryngelson and Wolynes, 1987; Chan and Dill, 1994; Sali et al., 1994; Goldstein et al., 1992; Abkevich et al., 1995). Since the  $\mathbf{HP}$  potential is very coarse-grained, every unique folder (i.e., sequences with  $g_0(S_i) = 1$ ) may be assumed to correspond to a reasonably fast folding sequence under biological folding conditions. Six of 6349 uniquely folding sequences, however, have a “real” energy gap. This means that no structure corresponds to the first excited state  $g_1(S)$  with one energy unit ( $\mathbf{HH}$ -bond) worse (less) than the minimum  $e_0(S)$ , i.e.,  $g_1(S) = 0$ . These instances were shown to fold significantly faster in simulations similar to Monte Carlo folding (Chan and Dill, 1994) and a chain growth procedure (Bornberg-Bauer, 1997). While *folding* is not at the core of this work, it is certainly notable that *all* six sequences with a gap are prototype sequences  $\hat{S}^r$  of very frequent structures. In fact, they represent 6 of the 10 most frequent structures. This is clearly beyond coincidence because only 170 (3%) of all sequences are well defined unique prototype sequences. Also, all these corresponding nets contain no sequence with a number of neutral neighbors that is one less than the number of neutral neighbors of the prototype sequence. There are several sequences (“next most”) that have two less neutral neighbors. The corresponding values are listed in Table 3.

TABLE 3 All six observed sequences with an energy gap are prototype sequences  $\hat{S}^r$  of frequent structures

Rank	1, 2	3, 6	9, 10
Size of neutral net	48	37	11
Energy gap of $\hat{S}^r$	2	2	2
Neutral neighbors (of $\hat{S}^r$ )	10	10	9
Neutral neighbors (next-most)	8	8	6

This is an interesting link between foldability, thermodynamic and mutational stability. Obviously uniqueness can be inherited with foldability.

### Switches between neutral sets

The overall number of direct transitions between neutral sets, (i.e., the minimal Hamming distance between any two members of two nets  $h(S_i^r, S_j^r) = 1$ ) is only 3428. The total of neutral mutations is 12912 in  $(\mathcal{S}, \mathcal{X})$  (Chan and Dill, 1994). This makes an average of 2.3 “connections” for each  $\mathcal{N}$ . This is certainly not enough to play an important role in exploring all sequence space through a continuous path of unique structures.

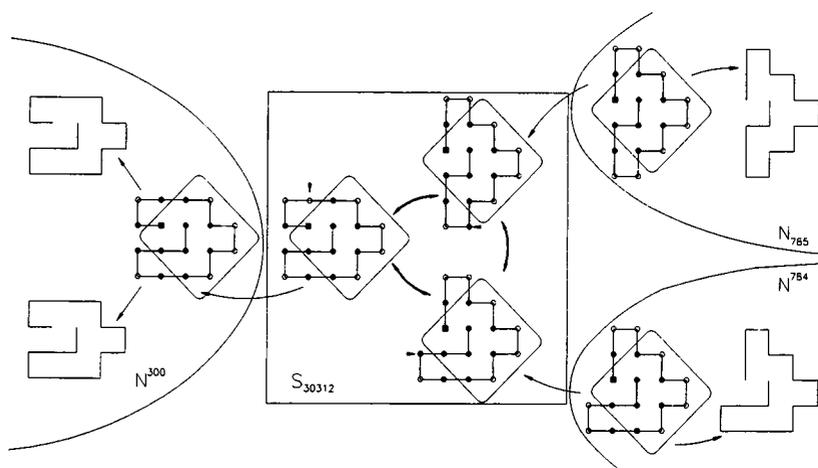
While most work on current-day biological protein structures emphasizes the uniqueness of the ground state, it is for several reasons interesting to speculate about the potential role of multiple ground states. Switching between two distinct structures is sometimes of fundamental importance for biological activity, e.g., for binding ligands. More dramatic changes have been reported with [e.g., for hemagglutinin (Lupas, 1996)] and without changes of solvent conditions [e.g., the prion molecule has been suggested to exist in two structurally distinct isoforms (Prusiner, 1995)].

Also, from an evolutionary point of view, it may well be important for a biopolymer *not* to fold to a unique ground state (but to two or more states with different functions) and not to lose flexibility of adaptation by freezing in a “mutational trap.” In a gedanken experiment, one can easily imagine such sequences play a role in bridging the gaps in sequence space between neutral sets. This might be especially meaningful when only certain structural regions are important for selection.

In Fig. 6 it was shown that switches may slightly enhance the probability of transition to other structures. An example for a 3-fold switch is given in Fig. 7.

Sequence  $S_{30312} = (\mathbf{HHPPHPHPHPHPHHH})$  has three equivalent ground state configurations with energy  $e = -9$ . Each of these configurations can be stabilized by a point mutation leading to a uniquely folding sequence of the corresponding neutral net. Indeed a significant region (*boxed residues*) remains unchanged, while two looplike regions are flexible. It is certainly interesting to note that the larger net  $\mathcal{N}^{300}$  with  $F^{300} = 10$  is destabilized by a  $\mathbf{H} \rightarrow \mathbf{P}$  mutation, while the slightly rarer nets ( $\mathcal{N}^{785}$ ) and ( $\mathcal{N}^{784}$ ) with  $F^{784} = F^{785} = 9$  are destabilized by  $\mathbf{P} \rightarrow \mathbf{H}$ , more surface-exposed mutations.

FIGURE 7 An example for a 3-fold switch. A single sequence ( $S_{30312}$ , boxed region) has three ground states that may refold into each other (double arrows) and are adjacent (in terms of single point mutations from  $P \rightarrow H$ , thin arrows) to the corresponding neutral net ( $N^{785}$ ,  $N^{784}$ ,  $N^{300}$  within the large ellipses).



## DISCUSSION

The complete sequence-to-structure map of a popular and well investigated protein model has been analyzed. The relevance to structural features and folding behavior has been repeatedly demonstrated [see Dill et al. (1995) and references therein]. I have combined the biophysical perspectives of structure formation and stability with the evolutionary aspects of sequence diversity. Unfortunately, due to computational limitations, the results are restricted to short chains on a square lattice. While some of the ensemble properties presented here may well change in detail when investigating other alphabets and energy sets, I think that the major conclusions will hold since they are compatible with natural observations and results from similar models. Recent progress in applying constraint programming techniques enables extension of the approach to more sophisticated models, including three-dimensional models and larger alphabets. Hence the influence of choice of the potential and alphabet size will be the subject of future work.

## Summary

A main result is that there are *few frequent, and many rare structures*. This is consistent with considerations about the limited number of natural protein folds (Chothia, 1992; Orengo, 1994) a mean field model (Govindarajan and Goldstein, 1996), the work from Li et al. (1996) on a similar lattice model, and RNA secondary structures (Schuster et al., 1994; Tacker et al., 1996; Bornberg-Bauer, 1996). While exact formulas for the size of shape space are available for RNA secondary structures (Schuster et al., 1994), this is not the case for lattice proteins and the influence of the potential on the dominance of compact structures.

MCSs are very likely to code for minimum free energy states (Camacho and Thirumalai, 1993). Since they are strongly overrepresented in ensembles that are derived from models with a stronger overall attraction force, they can be more easily attained by *any* sequence. In the studies by Li et al. (1996) *only* MCSs were considered and the average size

of neutral sets for length 27 and 4.75% uniquely folding sequences on the  $3^3$  cube is calculated as  $|\bar{F}| = (134, 217, 728 \times 4.75)/(103, 346 \times 100) = 61.7$ . In their model  $F^1 = 3749$ , and the most frequent structure is 61 times as frequent as average, since  $F^1/\bar{F} = 3749/61.7 = 60.7$ . (Values for two-dimensional models were not reported.) It can be assumed that size of neutral sets and regularities in frequent structures may, to some extent, be a consequence of confining a study to MCSs. In this model  $\bar{F}^* = 4.3$  is a consequence of potential and folding *only*. My studies also comprised different chain lengths and, as I have shown in a recent work, these principles can as well be observed when using a kinetically motivated algorithm (Bornberg-Bauer, 1997). While the properties of single structures are obviously strongly model-dependent, it is certainly interesting to note that ensemble properties are qualitatively independent (see also Comparison to RNA). Finally, it was found that Zipf's law, which is well known to describe a number of similar natural phenomena, provides a suitable *empirical* description of the structure distribution.

Similar conclusions hold for the emergence of regular structures. When not confined to MCSs, they are found as zig-zag patterns, but not in the contact map. Frequent structures show regular motifs with  $H$ 's on the inside and  $P$ 's to the outside. In contrast to earlier definitions (Chan and Dill, 1991a) I think these motifs are an alternative definition for secondary structure elements in small lattice models. The MCSs are, similar to the work from Li et al. (1996), highly regular, but do not represent very frequent structures. I assume that high designability does not depend on symmetry or potential, but the ability to design out alternative configurations. Regular elements would then be mostly a consequence of compactness, which in turn can be enforced by stronger attractive potentials. This is in agreement with off-lattice simulations that support the view that regularities are enforced by attractive potentials, but to observe a large amount an extended definition of secondary structures is required (Yee et al., 1994). It also complies with recent simulations suggesting that compactness only slightly en-

hances secondary structure formation (Hunt et al., 1994). Forcing structures on a lattice, however, in this context mimics hydrogen bonds in forming specific regularities, e.g., the lattice analogy to  $\alpha$ -helices.

Most converging sequences belong to single, connected *networks* that are *clustered in sequence space* and can be characterized by a single and stable *prototype sequence*. This sequence is extraordinarily resistant to mutations and in general is identical to the consensus sequence of the homologous sequences of the neutral set. Exchanging 50% of the **HP**-pattern means a very dramatic change when applied to a natural protein. Within the limits of the **HP** model I conclude that consensus sequences may code for structures that are extraordinarily stable thermodynamically and toward mutations.

The analysis of *profiles* and the role of degenerately folding sequences show that for all structures the mutability of **H**'s remains rather constant (and low), indicating that the hydrophobic residues require a constantly strong conservation for structures of all frequencies. Rarer structures require lower mutability for surface-exposed residues, which can be seen as the reason why they are, in general, poor in designing out. This is also reflected by entropy measures derived from profiles of homologous sequences. It will be investigated in more detail together with the role of compactness and correlated mutations in a forthcoming study.

Studies on *shape space covering* show that it is very difficult to convert one structure into another by a few point mutations. The number of direct connections (i.e.,  $h = 1$ ) between members of neutral sets is very small, such that an evolutionary strategy as proposed by Lipman and Wilbur (1991) seems very unlikely. Degenerately folding sequences provide transition regions with dual function in sequence space that may correspond to reduced, but still viable, activity. Still, however, the number of such *switches* may be too small for biological significance.

### Comparison to real proteins

If one is willing to accept the **HP** model as a reasonable approximation for *real* proteins, then we obtain a suitable conceptual framework that is in remarkable accordance with a number of properties of evolutionary relevance.

The relative scarcity of MCSs may well change with different potentials. However, this observation complies with observations that protein shapes tend to be compact, yet very often not maximally compact (Dill et al., 1995; Goodsell and Olson, 1993).

Although this model simulates only one force, the hydrophobic one, a number of structures are surprisingly stable toward mutations as long as the hydrophobic core is mostly conserved. This complies with the fact that the overall folds of real proteins are very stable toward mutations and mostly dependent on the binary **HP** pattern (Reidhaar-Olson and Sauer, 1988; Orengo et al., 1994; Kamtekar et al., 1993).

The small number of uniquely folding sequences is similar to the fraction (1–5%) of stable, well defined structures

that can be found in samples of random sequences that were assembled from a ternary hydrophobic-polar alphabet (**QLR**) (Davidson et al., 1995). In this and from similar studies it was also concluded that a significant fraction of random sequences will fold uniquely and most probably to a frequent structure. This is also very likely in this model, since frequent structures are represented by neutral nets that span regions in sequence space that are up to more than one-third of the diameter.

Real proteins preferably fold into a small number of "folds" (Chothia, 1992; Cordes et al., 1996; Orengo et al., 1994). These may correspond to the frequent structures in this model. From there and from considerations about the small fraction of sequence space that was explored during evolution (Eigen, 1987), the assumption was drawn that many structures were simply not found during evolution. This corresponds to the obvious conclusions that rare structures are hard to find in this model and have little capacity to adapt.

It is often assumed that protein structures ("folds") evolved independently (Orengo et al., 1994). This appears reasonable from the presented perspective when we consider the strong separation of nets in sequence space. Because point mutations maybe are not sufficient for fast adaptation, it may have become necessary to develop evolutionary alternatives that combine established units. This is probably reflected by the modular nature of many proteins.

From the knowledge of structure distribution in sequence space we may also infer some conclusions that are relevant to improve data base searches. This is of paramount importance if one considers efforts in the fields of bioinformatics that accompany current genome projects. Explicitly applying different rules for core residues and surface exposed residues might prove helpful to estimate the potential mutational tolerance of a given structure. Applied to sets of homologous sequences in a database (of real proteins) this might prove useful to obtain rough estimates for the number of potential (i.e., to date undetected) members of a protein family given some representative members. According to the strong separation in sequence space we expect it is thereby possible to delimit families more precisely than with current methods. This could possibly be achieved by procedures that infer information from all represented members of a family instead of using averaged representatives such as profiles or regular expressions. This may help to circumvent the difficulties of the twilight zone that frequently emerge when comparing a query sequence to many families.

Finally, the concept of prototype sequences that are very stable invites one to consider new possibilities to design stable and fast folding sequences that can be used for folding experiments.

### Comparison to RNA

It is also interesting to compare these results with recent computer experiments on RNA secondary structures. In

both cases we find a large number of neutral neighbors, a structure distribution following Zipf's law and *rugged landscapes*, i.e., a few random mutations randomize structure ensembles (Fontana et al., 1993; Schuster et al., 1994; Bornberg-Bauer, 1996; Tacker et al., 1996; Renner and Bornberg-Bauer, 1997). These features seem to be *generic properties of biopolymers* sequence-to-structure maps. However, there are most remarkable differences in the two maps regarding their possibilities to explore sequence spaces.

First, neutral nets for RNA are percolating through sequence space (Schuster et al., 1994; Tacker et al., 1996), which is a consequence of the boolean pairing logic of the nucleotides and the fine-grained energy spectrum in the density of states. This is clearly not the case for the **HP** model. This may be less effective for a larger alphabet and more sophisticated potentials, yet it should prevail as long as one noncomplementary force, such as the hydrophobic effect, dominates the spectrum of interactions. Recent investigations on average ensemble properties using the concept of *landscapes* (Renner and Bornberg-Bauer, 1997) have shown that larger alphabets (and therefore a larger sequence space) with a finer potential slightly smoothen the map. This is, roughly speaking, tantamount to a larger degree of neutrality and complies to the influence of different alphabet size in RNA.

Second, in the RNA secondary structure case virtually all structures can be observed next to a neutral net (Fontana et al., 1993; Schuster et al., 1994). Within a number of mutations small compared to the length of the sequence, the whole shape space can be covered. In the **HP** model, however, the structures are well separated and direct transformations to another structure are rare. This is also intuitively clear, since, in contrast to the boolean logic of base pairing in RNA, in the **HP** model one stabilizing interaction cannot be substituted with another, simply because there is no other. It will be interesting to see if this feature is sensitive to the alphabet size.

I would like to thank H. S. Chan and K. A. Dill, University of California, San Francisco, who generously provided ground state data. Thanks to the hospitality of PMMB (Program in Mathematics and Molecular Biology), where this cooperation was initialized. Critical discussions with M. Vingron, R. Goldstein, W. Fontana, H. S. Chan, and R. Backofen were useful. Thanks to P. Mutzel for help with the graph drawing and useful comments on the manuscript by H. S. Chan, M. Rehmsmeier, E. Rivals, and the referees.

## REFERENCES

- Abkevich, V., A. Gutin, and E. Shakhnovich. 1995. How the first biopolymers could have evolved. *Proc. Natl. Acad. Sci. USA*. 93:839–844.
- Benner, S. A., M. A. Cohen, and G. H. Gonnet. 1993. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.* 229:1065–1082.
- Bornberg-Bauer, E. 1996. Random structures and evolution of biopolymers: a computational case study on RNA secondary structures. *Pharm. Acta Helv.* 71:79–85.
- Bornberg-Bauer, E. 1997. Chain growth algorithms for HP type lattice proteins. *RECOMB Proceedings*, 47–55, ACM Press, New York.
- Bryngelson, J., and P. G. Wolynes. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA*. 84:7524–7528.
- Camacho, C. J., and D. Thirumalai. 1993. Minimum energy compact structures of random sequences of heteropolymers. *Phys. Rev. Lett.* 71:2505–2508.
- Casari, G., and M. J. Sippl. 1992. Structure-derived hydrophobic potential. *J. Mol. Biol.* 224:725–732.
- Chan, H. S., and K. A. Dill. 1991a. Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Chem.* 20:447–490.
- Chan, H. S., and K. A. Dill. 1991b. Sequence space soup of proteins and copolymers. *J. Chem. Phys.* 95:3775–3787.
- Chan, H. S., and K. A. Dill. 1994. Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys.* 100:9238–9257.
- Chan, H. S., and K. A. Dill. 1996. Comparing folding codes for proteins and polymers. *Proteins*. 24:335–344.
- Chothia, C. 1992. One thousand families for the molecular biologist. *Nature*. 357:543–544.
- Cordes, M. H., A. R. Davidson, and R. T. Sauer. 1996. Sequence space, folding and protein design. *Curr. Opin. Struct. Biol.* 6:3–10.
- Czirok, A., R. N. Mantegna, S. Havlin, and H. Stanley. 1995. Correlations in binary sequences and a generalized Zipf analysis. *Physiol. Rev. E*. 52:446–452.
- Davidson, A. R., K. J. Lumb, and R. Sauer. 1995. Cooperatively folded proteins in random sequence libraries. *Nat. Struct. Biol.* 2:856–864.
- 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.
- Eigen, M. 1987. *Stufen des Lebens*. Piper, München.
- Fontana, W., P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. 1993. RNA folding and combinatorial landscapes. *Physiol. Rev. E*. 47:2083–2099.
- Goldstein, R. A., Z. A. Luthey-Schulten, and P. G. Wolynes. 1992. The statistical mechanical basis of sequence alignment algorithms for protein structure recognition. *Proc. Natl. Acad. Sci. USA*. 89:4918–4922.
- Gonnet, G. H., and R. Baeza-Yates. 1991. *Handbook of Algorithms and Data Structures*. Addison-Wesley, Don Mills.
- Goodsell, D., and A. Olson. 1993. Soluble proteins: size, shape and function. *TIBS*. 18:65–68.
- Govindarajan, S., and R. A. Goldstein. 1996. Why are some protein structures so common? *Proc. Natl. Acad. Sci. USA*. 93:3341–3345.
- Govindarajan, S., and R. A. Goldstein. 1997a. Evolution of model proteins on a foldability landscape. *Proteins*. In press.
- Govindarajan, S., and R. A. Goldstein. 1997b. The foldability landscape of model proteins. *Biopolymers*. In press.
- Huang, E. S., S. Subbiah, and M. Levitt. 1995. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* 252:709–720.
- Hunt, N. G., L. M. Gregoret, and F. E. Cohen. 1994. The origins of protein secondary structure effects of packing density and hydrogen bonding studied by a fast conformational search. *J. Mol. Biol.* 241:214–225.
- Huynen, M., P. Stadler, and W. Fontana. 1996. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA*. 93:397–401.
- Kamtekar, S., J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science*. 262:1680–1685.
- Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature*. 217:624–626.
- King, J. L., and T. H. Jukes. 1969. Non-Darwinian evolution. *Science*. 164:788–798.
- Koshi, J. M., and R. A. Goldstein. 1997. Mutation matrices and physical-chemical properties: correlations and implications. *Proteins*. 27:336–344.
- Lau, K. F., and K. A. Dill. 1989. A lattice statistical mechanics model of 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.
- Lipman, D. J., and W. J. Wilbur. 1991. Modelling neutral and selective evolution of protein folding. *Proc. R. Soc. Lond. B*. 245:7–11.
- Lupas, A. 1996. Coiled coils: new structures and new functions. *TIBS*. 21:375–382.
- Maynard-Smith, J. 1970. Natural selection and the concept of a protein space. *Nature*. 225:563–564.
- Onuchic, J. N., P. Wolynes, Z. Luthey-Schulten, and N. Socci. 1995. Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl. Acad. Sci. USA*. 92:3626–3630.
- Orengo, C. 1994. Classification of protein folds. *Curr. Opin. Struct. Biol.* 4:429–440.
- Orengo, C. A., D. T. Jones, and J. M. Thornton. 1994. Protein superfamilies and domain superfolds. *Nature*. 372:631–634.
- Prusiner, S. B. 1995. Prion diseases. *Scientific American*. 3:44–54.
- Reidhaar-Olson, J., and R. Sauer. 1988. Combinatorial cassette mutagenesis as a probe of the information content of protein. *Science*. 241:53–57.
- Renner, A., and E. Bornberg-Bauer. 1997. Exploring the fitness landscapes of lattice proteins. *Proceedings of the 1997 Pacific Symposium on Biocomputing*. R. Altman, K. Dunker, L. Hunter, and T. Klein, editors. World Scientific, London. 361–373.
- Sali, 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.
- Schulman, B. A., and P. S. Kim. 1996. Proline scanning mutagenesis of a molten globule reveals noncooperative formation of a protein's overall topology. *Nat. Struct. Biol.* 3:682–687.
- Schuster, P., W. Fontana, P. F. Stadler, and I. L. Hofacker. 1994. From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. Lond. B*. 255:279–284.
- Shakhnovich, E., and A. Gutin. 1990. Enumeration of all compact conformations of copolymers with random sequence of links. *J. Chem. Phys.* 93:5967–5971.
- Skolnick, J., and A. Kolinski. 1990. Simulations of the folding of a globular protein. *Science*. 250:1121–1125.
- Strait, B., and G. T. Dewey. 1996. The Shannon information entropy of protein sequences. *Biophys. J.* 71:148–155.
- Tacker, M., P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. 1996. Algorithm independent properties of RNA secondary structure predictions. *Eur. Biophys. J.* 25:115–130.
- Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth Intl. Congress on Genetics*. Vol. 1. D. F. Jones, editor. Brooklyn Botanic Gardens, New York. 356–366.
- Yee, D. P., H. S. Chan, T. F. Havel, and K. A. Dill. 1994. Does compactness induce secondary structure in proteins? A study of poly-alanine chains computed by distance geometry. *J. Mol. Biol.* 241:557–573.
- Yue, K., M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill. 1995. A test of lattice protein folding algorithms. *Proc. Natl. Acad. Sci. USA*. 92:325–329.