EXPLORING THE FITNESS LANDSCAPES OF LATTICE PROTEINS

ALEXANDER RENNER [a] and ERICH BORNBERG-BAUER [b c d]
(b) Abt. Theoretische Bioinformatik, Deutsches Krebsforschungszentrum
Im Neuenheimer Feld 280, D - 69 120 Heidelberg

We present methods to investigate the sequence to structure relation for proteins. We use random structures of HP-type lattice models as a coarse grained model to study generic properties of biopolymers. To circumvent the computational limitations imposed by most lattice protein folding algorithms we apply a simple and fast deterministic approximation algorithm with a tunable accuracy. We investigate ensemble properties such as the conditional probability to find structures with a certain similarity at a given distance of the underlying sequence for various alphabets. Our results suggest that the structure landscapes for lattice proteins are generally very rugged, while larger alphabets fine tune the folding process and smoothen the map. This implies a simplification for evolutionary strategies. The applied methods appear to be helpful in the study of the complex interplay between folding strategies, energy functions and alphabets. Possible implications to the investigation of evolutionary strategies or the optimization of biopolymers are discussed.

1 Introduction

Under physiological conditions *in vitro* biopolymers generally fold to a unique structure. It is often assumed that only the sequence determines this "native" state and that it corresponds to the MFE (equilibrium minimum free energy) state (the *thermodynamic hypotheses*). The search space is astronomically large, yet proteins fold in seconds. Many mechanisms were proposed to understand protein folding, but there is no consensus yet [1]. Folding *in vivo* is even more involved since several agents prevent misfolds, aggregation etc. During the last decades several highly simplified models, among them lattice proteins have been derived to investigate the basic principles that govern the folding process of biopolymers and enable natural proteins to evolve under the constraints of functional adaptation and natural foldability [e]. In the **HP** model [2,3] (where **H** stands for a hydrophobic residue and **P** for a polar one) it is assumed that the non specific *hydrophobic force* is the dominant contribution to stability. It therefore to a large extent determines the 3D structure of the backbone [4,3,2]. In this framework side-chain packing selects structures within this relatively small set of compact states and hence allows for detailed functional fine tuning. For an excellent review of current methods the reader is referred to Dill *et al.* [3].

[a] Institut für Theoretische Chemie, Universität Wien, Währingerstraße17, A-1090 Wien
[c] Inst. für Mathematik, Universität Wien, Strudlhofg. 4, A-1090 Wien (Austria ⊂ Europe)
[d] correspondence: bornberg@dkfz-heidelberg.de , http://www.dkfz-heidelberg.de/tbi
[e] i.e. the ability to also attain the functional state within a reasonable time.

For our investigations we will use a sequential folding procedure and apply it to **HP**-type lattice proteins. Sequential folding may be relevant within several frameworks e.g. for the folding of sub-domains *in vitro* and the early steps of forming a nucleus or locally ordered structures. It may also account for the case of unguided folding of a nascending chain [f] *in vivo* being extruded from the ribosome to the lumen. This was proposed by Levinthal [5]. Some evidences for the relevance of sequential folding were recently summarized [6].

There is a promising strategy to construct biopolymers without disposing of details about folding: applied molecular evolution is intended to complement or even replace rational design. There, starting from an initial pool of random sequences, the principles of evolutionary optimization, error prone replication and selection of fitter offsprings, are applied in a test tube system. This illustrates the importance of studying not only the foldability of single sequences but the sequence structure relation of ensembles of random ensembles as well. To understand and describe at a molecular level how the principles of Darwinian evolution act in shaping biopolymers is also crucial for the understanding of prebiotic evolution. These principles can be exploited for biotechnology. Evolving entities must in principle accomplish two tasks: to *conserve* acquired features in their genotype and to *adapt* to new requirements on the phenotypic level as well. Since there is a tradeoff between these tasks, it is crucial to understand roles, interdependencies and interrelations between genotype and phenotype.
Early concepts (developed in the thirties by S. Wright and R. Fisher) coined the term of *fitness landscapes*. Evolution is viewed as an adaptive walk over the set of genotypes preferring "fitter" offsprings by selecting for some functional criterion, a phenotype property. Later considerations focused on influence and importance of phenotypically neutral mutations. Only few mutations can be advantageous but a continuous gradient of fitness must be maintained so that mutated offsprings survive [7,8]. Applied to biopolymers, this implies that residues essential for function will be rather conserved and non-essential ones will be replaced by evolutionary diverse sequences [g].
Since it is difficult to define fitness *a priory* and it is generally assumed that structure largely determines function [h], we are primarily interested in the *sequence to structure map*. "Simple exact models" [3] such as lattice proteins or RNA secondary structures are an ideal playground to explore these issues on large ensembles of biopolymers. The impact of parameters on structure formation can be studied in full detail and computational demands are reduced to a manageable level. Since in principle the structure prediction problem is of comparable complexity for real proteins and (fully represented) RNA, we were motivated by the recent success in characterizing the sequence to structure map-

[f] i.e. a chain under construction
[g] For the **HP** representation one would expect **HH** contacts in the core to be conserved.
[h] in the sense that it is a *conditio sine qua non* [9]

ping for RNA secondary structures[10,11,12,13,14]. This problem is, however, more involved for real proteins: 1) in contrast to RNA, proteins do not comprise genotype and phenotype in one molecule 2) there is a level of neutrality that arises from the redundancy of the genetic code at the genotype level *and* from structural robustness of folding at the phenotype level 3) structure representations simpler than the lattice approximation are not available. This in turn implies the need for computationally demanding almost-exhaustive or approximation algorithms.
In this work we are not so much interested in folding single instances. This has been solved at a reasonable level for the **HP** model[3,15]. For a different lattice model, a 3x3x3 cube with a broad spectrum of interactions the thermodynamic property of a pronounced energy gap between the ground state and other states was proposed to be a necessary and sufficient condition for fast folding in a Monte Carlo run[16,17,18]. We will also not discuss evolutionary issues any further (such as the nature of a possible primordial alphabet and the number of possible structures[19,20,21] that can be realized). It is our goal to present some techniques to give an idea how questions that we consider as relevant to understand limits and possibilities of biopolymer evolution can be addressed at different levels of simplification. Some recent results that we hope will clarify some aspects of the complex interplay of folding mechanisms, alphabets, and potentials will be presented.

2 Methods

2.1 "Generic" Lattice Proteins

The HP model: Here we refer to one of the most popular models of lattice proteins, the subclass of **HP**-models, introduced by Dill *et al.*[2,3]. All residues have the same size. The peptide chain is constructed by placing residues sequentially on the beads of a regular lattice. The resulting chain has identical bond lengths and discrete bond angles. We use relative moves for storing and comparing structures: the structure is represented as a self avoiding walk on a regular lattice and the movement of the chain is represented as a sequence of moves where each is encoded relative to the prior. The method is well known (see e.g. [2]); our version has been adapted to apply to any regular lattice (a detailed description will be given[22]). The algorithm has several advantages over representing structures by absolute moves or integer coordinates: 1) lattice independent programming of folding algorithms and structure comparison is possible 2) point mutations are pivot moves[23] 3) concatenation of strings corresponds to elongation of the walks 4) storage requirements are kept small and 5) structures can be compared utilizing classical string comparison methods[24].

Potentials: The generalized energy function for a sequence with n residues $S = (s_1, s_2, ..., s_n)$ with $s_i \in \mathcal{A} = \{a_1, a_2, \ldots, a_b\}$, the alphabet of b residues and an overall configuration $X = (x_1, x_2, ..., x_n)$ on a lattice $\mathcal{L}$ can be written as the sum of all pairwise inter-residue interactions:

$$E(S, X) = \sum_{i}^{n} \sum_{j>i+1}^{n} \mathcal{E}(s_i, s_j) d_{ij}^{\alpha} f(s_i, s_j, |i-j|) \qquad (1)$$

where $d_{ij} = ||x_i - x_j||$ is the Euclidian distance, $\mathcal{E}_{ij} = \mathcal{E}(s_i, s_j)$ a pair-potential retrieved from the energy matrix. In our implementation, contributions are considered up to a certain cutoff distance: $d_{ij}^{\alpha} = 0$ if $d_{ij} >$ cutoff. For consistency we used $cutoff = 1$ whenever direct comparison to Dill's model was considered and $f = 1$ throughout this work.
We implemented three different potentials: In the "classical" **HP**-model (random) heteropolymers are composed from $\mathcal{A} = \{$ **H**, **P** $\}$ with only one stabilizing interaction if and only if hydrophobic residues (**H**) are neighbors on the lattice but not along the chain. Polar residues (**P**) do not explicitly contribute to the energy. The salient features of real protein structures are implicitly considered: the hydrophobic effect comprises solvent-driven collapse to a native state, the self-avoiding walk constraint accounts for the excluded volume effect. The **HP'** set includes a strong overall interaction as well. The **HPNX**-model is a generic extension of the **HP** model and mimics "electrostatic" interactions between negative residues (**N**) and those with a positive charge (**P**) as well as repulsions within these classes. A third class of apolar residues is "neutral" (**X**) [i].

$\mathcal{E}_{ij}$	H	P
H	-1	0
P	0	0

$\mathcal{E}_{ij}$	H	P
H	-3	-1
P	-1	-0

$\mathcal{E}_{ij}$	H	P	N	X
H	-4	1	1	0
P	1	0	-1	0
N	1	-1	0	0
X	0	0	0	0

Table 1: Energy potentials for alphabets **HP**, **HP'** and **HPNX**.

Lattice Protein Folding is NP-hard [25]. A large variety of approximation algorithms was therefore developed [15,26,27]. Most of these are not fast enough to investigate large ensembles of structures and stochastic optimization techniques (see e.g. [28]) are not useful either to study ensemble properties of specifically folded single chains[j]. Hart and Istrail [29] recently presented an algorithm for the **HP** model that guarantee folding within at least 3/8 of the optimum energy. It

[i] The frequency of **H**s is the same as in the **HP** model, such that a random distribution of the **HX** subset corresponds exactly to the **HP** model.

[j] another reasons is given in the next section 2.2

is deterministic, works in $\mathcal{O}(n)$, but does not consider different potentials and cross-space interactions. It will be compared in future work.
Here we use a straightforward deterministic algorithm, termed the *greedy Chain Growth Algorithm* (gCGA) [30]. In its simplest version the algorithm is fast but there is, of course, a trade-off between accuracy and speed. Starting with an initial move, it proceeds along the chain. The next m residues in the sequence are added without consideration of the following residues. Energy contributions, retrieved from $\mathcal{E}_{IJ}$, are evaluated for all neighbors. The next move is determined by sorting these configurations with respect to energies, selecting the best and appending the first move of this chosen configuration to the "frozen" core. The gCGA was shown [30,24] to yield good results for short chains on a square lattice.

2.2 Landscapes

Sequence Space $\mathcal{S}^{(n)}$ is defined as the set of all b^n sequences $S_i(n)$ of a given length n that can be converted by well defined string-edit operation; we regard only point mutations. For two strings of equal length n, the number of positions by which they differ is known as *Hamming distance* h and defines a metric in *Sequence Space* $\mathcal{S}^{(n)}$. The probability $P[h]$ that two randomly chosen sequences have distance h is given by:

$$P[h] := P[d_S(S_1(n), S_2(n)) = h] = (b-1)^h \binom{n}{h} b^{-n} \tag{2}$$

The *Shape Space* $\mathcal{X}^{(n')}$ is defined as the set of all possible structures X_i for sequences of length n. Following Guttman *et al.* [31,32] the number of *self avoiding walks* (SAWs) on a lattice is $\#(SAWs) = a\hat{c}_{eff}^{n-2} n^{\beta}$ where β is a scaling exponent and c_{eff} the effective connectivity of the lattice [k].

Our description of *landscapes* follows that of Fontana *et al.* [10,11] on RNA secondary structures. A general notion starts with the definition of *Combinatory Maps* (**CM**) which are maps from one metric space $(\mathcal{G}, d_G)$ into another metric space $(\mathcal{F}, d_F)$. If a scalar quantity is assigned, the mapping $\phi : (\mathcal{G}, d_G) \to \mathbb{R}^1$ was also termed a *combinatorial landscape* (**CL**).
For reasons mentioned above we are interested in the sequence to structure map and functional properties associated with the structure. CMs are then viewed as generalizations of mappings from genotype (*sequence space* $(\mathcal{S}, d_S = h)$) to phenotype (*shape* or *structure space* $(\mathcal{X}, d_x = t)$), CLs are generalizations of mappings from genotype into fitness values. Biopolymer folding can now be understood as a mapping Φ from one space into another: $\Phi : (S, h) \Longrightarrow (X, t)$.

[k] For a square lattice $\hat{c}_{eff} = 2.63$ and $\beta = 0.33$ for the cubic lattice 4.68 and 1.16 respectively.

Scalar phenotype characteristics f_i for biopolymers are, for example, the radius of gyration $F_i := G_i(S_i, X_i)$ or the minimum free energy $F_i := E_i(S_i, X_i)$. A metric is simply $d_F(i,j) = |F_i - F_j|$. We define *neighbors* of a genotype G_i as the set of genotypes G_j with the smallest possible distance in genotype space $\mathcal{G}$: $N(G_i, d_G) = \{G_j | d_G(i,j) = 1\}$. *Neutral Neighbors* $NN(G_i)$ in a CL or a CM are the set $N(G_i)$ of neighbors that fall into the same phenotype with respect to the chosen d_F and Φ: $NN(G_i, d_G, d_F) = \{N(G_j) | d_F(i,j) = 0\}$. An instance (G_i, F_i) is called a *local optimum* if all neighbors $N(G_i)$ have fitness values lower than $F(G_i)$.

Landscapes have a characteristic topology. If there is a large number of local optima near any point, the landscape is called *rugged* and global optimization strategies may fail. Most descriptions are based on the definition of an autocorrelation function where $\langle . \rangle$ denote expectation values:

$$\rho(h) := \rho(d_G = h) = 1 - \frac{\langle d_F^2 | h \rangle}{\langle d_F^2 \rangle} \tag{3}$$

This expression can be viewed as a measure of the average similarity d_F of phenotype properties (energies, radius of gyration, structures etc.) for a fixed genotype distance h of the underlying genotype (sequence) [l]. It is obvious that for $h = 0$ (i.e. two identical sequences) a deterministic procedure (but not necessarily a stochastic one) will yield the same structure. Consequently, the autocorrelation function yields 1 at $h = 0$ and decays to a value of $\rho(h) = 0$ when all similarity is destroyed. A suitable *characteristic length* is the *correlation length* l_{d_F}. It is defined as the solution of $\rho_F(h) = 1/e$ [m]. As analytical solutions are not available for most landscapes we use large statistical ensembles of computationally folded biopolymers to compute $\rho(h)$. A two-dimensional probability density surface $P[d_F = t | d_G = h]$ was proposed for easier visualization [10,11]. It expresses the joint probability of two genotypes $G_i(n), G_j(n)$ having phenotype distance $d_F(i,j)$ at a given genotype distance $d_G(i,j) = h$.

Structure representation: We use the string of relative moves $\mathcal{R}_i := R(X_i) = (r_1, r_2, \ldots, r_i), r_i \in R$ (where R is the alphabet of relative moves), the distance matrix $DM^{(n \times n)}$ (which is symmetric and contains the Euclidian distances between two residues) [n] and the contact matrix $CM(X_i)$, which contains a 1 where the entries $d_{ij} = 1$ and 0 else. Scalar measures of compactness are the *radius of gyration*, and the number of contacts C_C, defined for all $(\frac{b(b+1)}{2})$ pairs of interactions as: $C_C^{(a_i, a_j)}(X_I) = |\{C_{ij} | (a = i, b = j)\}|$.

Defining *distance measures* is essential for comparing structures and to characterize landscapes: the number of identical contacts is defined as $D_C^{a_i, a_j}(X_1, X_2) =$

[l] When e.g. structure distances are correlated to sequence differences, measured by h, we obtain a characteristics of the sequence to structure mapping.

[m] where e denotes Euler's constant

[n] All information except the nature of the bonds and the chirality are retained.

$|\{i,j \in N, i < j | c_{ij}(X_1) = c_{ij}(X_2) = 1\}|$ and can be normalized e.g. as $\hat{D}_C^{a_i,a_j}(X_1,X_2) = \frac{2D_C(X_1,X_2)}{C_C(X_1)+C_C(X_2)}$ such that only contact regions in two structures are taken into account. Comparing two structures R_i, R_j is simple: the Hamming distance counts the number of pairs of identical directions at identical positions $D_{R_h} = (n - h(R_1, R_2))$. R_i-s can also be aligned using standard dynamic programming procedures defining gap-penalties and edit costs for the exchange of directions [24].

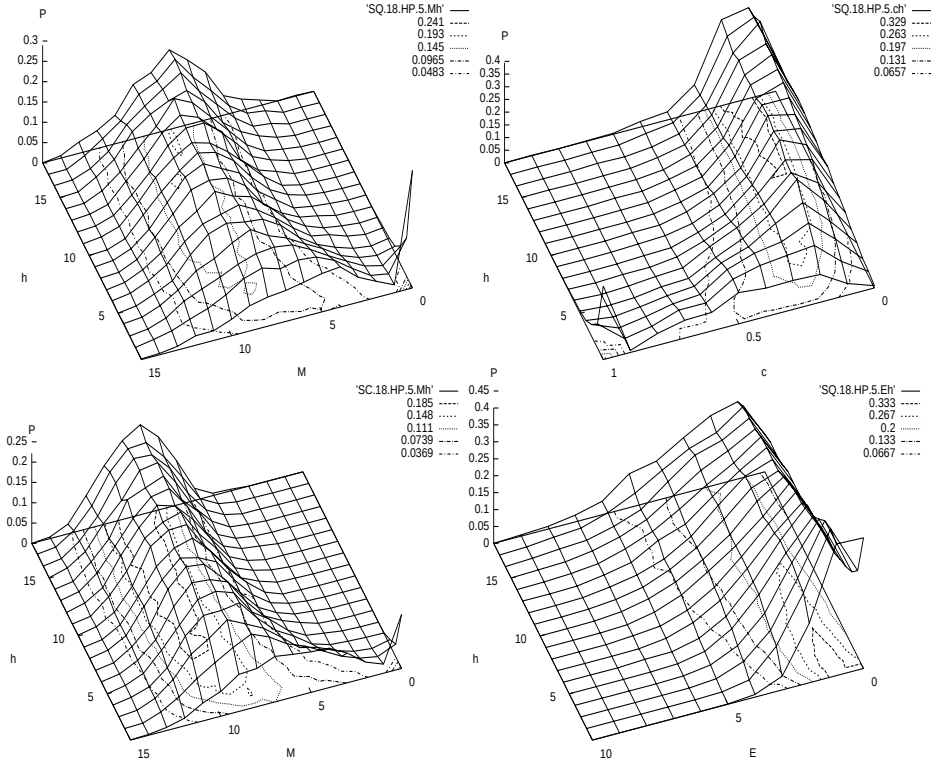

Figure 1: Probability density surfaces for lattice proteins ($n = 18$, **HP** alphabet, m=5, square lattice). (a): structure distance (relative moves) vs. h, (b): structure distance (contacts) vs. h, (c): same as (a) but cubic lattice, (d): energy distance vs. h.

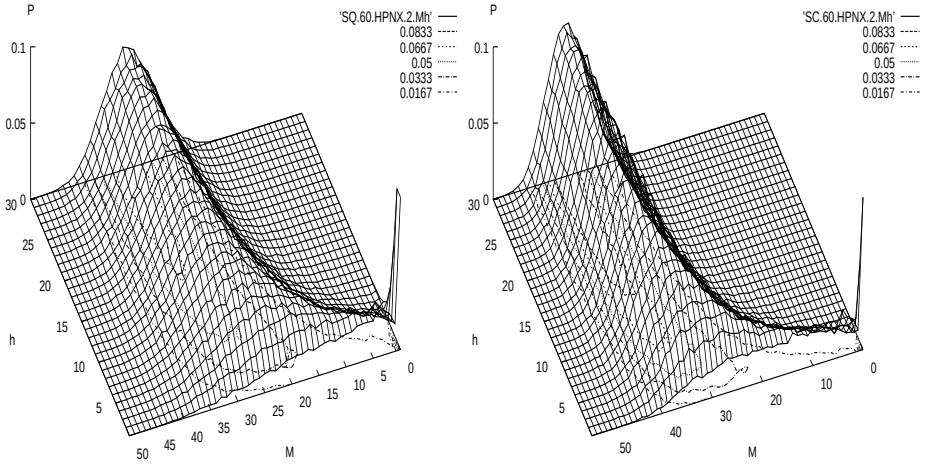

Figure 2: Probability density surfaces for lattice proteins ($n = 60$, **HPNX** alphabet, m=2): structure distance (relative moves M) vs. h on a square lattice (a) and a cubic lattice (b). (Data are cutoff at $h = 30$.)

3 Computations

Assessing the performance of the gCGA we have shown that increasing m yields better results[24,30]. At fairly small look-ahead values an average success rate of 10% and a performance within 80% of the optimal energy for can be obtained. In general, increasing m lowers energy and increases compactness and the number of contacts. The contacts, except **HH** remain rather unchanged, indicating that compactness results from a tighter core. A small number of **PP**-contacts implies that they are surface exposed without being explicitly penalized. The major reason for improved efficiency is that, the more the chain "looks ahead", the deeper a trap along the folding pathway can be overcome[24].

We compute large ensembles of random structures for short chains ($n = 18$) on a square and a simple cubic lattice. We generated 500 reference strings and 5 mutations for all hamming distances. Convergence of this uniform sampling procedure is fast. We checked the influence of the alphabet, the look ahead parameter m and the lattice. We calculated the conditional probability p that two structures (energies) have a distance m or c (or e respectively) given that their underlying sequences have Hamming distance h. Some instructive examples are reported in figs. 1 to 3, more comprehensive results will be reported elsewhere. The overall shape looks, similar to RNA landscapes[10,11], like half a horseshoe. Peaks at $h = 1$ and $M = 0$ refer to the number of neutral point mutations i.e. strictly identical structures. The probability to find a closely related structure

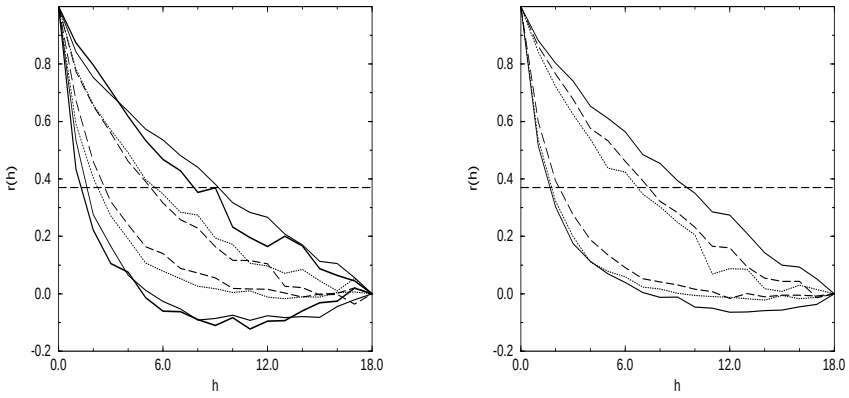

Figure 3: Lattice protein landscapes for the square lattice (a) and the cubic lattice (b). Autocorrelation functions $r(h)$ for energies (upper, nearly straight lines) and structure (relative moves, lower graphs) distances are shown for search depth $m = 5$ (thin lines) and 11 (thick lines), for the **HP** (full line), the **HP**' (dotted line) and the **HPNX** (dashed lines) alphabet.

is rapidly shifted to a random distribution with increasing h. Above a certain "critical" value h_{cr} the probability density becomes independent from h and the structure ensemble is essentially randomized. This defines something like a quantitative measure for a neighborhood size and corresponds roughly to the characteristic lengths (see below). This suggests that the majority of local optima as obtained by the gCGA can be found in a close neighborhood of any random structure. The structure density surfaces for D_{R_h} (Fig. 1a) and $\tilde{D}_C$ (Fig. 1b) have a similar shape although the structure measures are based on completely different definitions. Hence the overall shape of the density surfaces do not depend on the usage of a certain structure notion. The density surface for the cubic lattice (Fig. 1d) shows faster randomization which is intuitively clear since shape space is much larger. Still there is a significant number of neutral mutations. To illustrate the versatility of our approach we also report the density surfaces for length $n = 60$ and the **HPNX** alphabet on the square and the cubic lattice (Fig. 2). There the number of neutral mutations is significantly larger which supports an assumption by Lipman and Wilbur[33] about the increasing probability of neutral mutations with larger chains. Also the shift of the average structure distance to higher values becomes more pronounced for the cubic lattice.

The energy density surfaces (Fig. 1c) look different: energies are stronger correlated and again there is a significant number of neutral mutations. At small

h, however, the distribution is rather bell shaped and rapidly broadens. Since most structures are relatively compact, in the case of the **HP** alphabet most structures have 8,9 or at most 10 contacts and the most frequent energy distance is 1. This is of course not the case for different alphabets and longer chains (data not shown).
We then computed autocorrelation functions following Equn. 3. It can be clearly seen that correlations are influenced only slightly by the search depth. Results of structure statistics depend strongly on the particular alphabet and correlations are approximately HPNX > HP' > HP. Energies are less sensitive to mutations than structures which is a result of the high degeneracy of lattice models, that is the correspondence of more than one structure to the MFE state [3]. Larger alphabets, however, have energy landscapes that are more rugged. This reflects the larger number of possible states in energy space and a smaller degeneracy. It is certainly interesting to note that most of these results are similar to findings from RNA secondary structures [11,13].

4 Discussion

We presented and implemented an approach to characterize fitness landscapes of lattice proteins. Clearly enough our results on folding single instances are not unexpected from the "lattice protein point of view". We think, however, that our results are significant in the sense they constitute an important new method for understanding certain features of relevance for the evolution of biopolymers:

- Combining the **HP** model with the concept of relative moves and applying a fast approximation algorithm, makes it feasible to investigate ensembles large enough for a statistical characterization of fitness landscapes. It was also shown that the performance tradeoffs of the algorithm allow it to handle larger chains and thereby address biologically meaningful problems.

- Structure Landscapes of **HP**-type lattice proteins are very rugged. This suggests that there are many local optima and evolutionary strategies may easily get stuck. Yet energies are higher correlated i.e. less sensitive towards point mutations than structures. This is definitely a consequence of using random sequences that usually fold to multiple ground states [3]. Since uniquely folding sequences are rare it is reasonable to assume that evolution from one "unique folder" to another is even more difficult and requires more mutations. Larger alphabets reduce this degeneracy and energy and structure correlations correspond to a similar fitness criterion.

- The ruggedness depends strongly on the size of sequence space and shape space. Larger alphabets smoothen both, the folding landscape and the fit-

ness landscapes. This is another analogy to the fitness landscapes of RNA secondary structures and meets some earlier claims [24] that real proteins need a larger sequence space not only for chemical diversity but also for smooth evolution.

- The probability density surfaces also imply that a significant amount of neutrality complies with the possibility of a fast exploration of shape space [o]. The number of neutral mutations for larger Hamming distances, however, is small. This seems to be a clear contradiction to observations in real proteins that are very robust with respect to point mutations. On one hand this can be attributed to the crudeness of the **HP** model since each mutation actually corresponds to more mutations in a larger alphabet. On the other hand it may result from using an approximation algorithm that typically finds local optima.
  The issue of neutral evolution also deserves a closer look since earlier work in the **HP** - models by Lipman and Wilbur [33] postulated the existence of extended neutral sets in sequence space and that their frequency increases with sequence length.

- In spite of significant changes on single structure properties, ensemble properties are hardly influenced by the search depth. This is particularly interesting in the light of most recent results on RNA secondary structures [13] where ensemble properties are robust with respect to the chosen algorithm and will be the subject of more detailed studies.

Since we used several simplifications, we view our results as a very crude model of realistic processes of the "real world" analogy. Major caveats to our studies are certainly the restriction to very simple models and the usage of an approximation algorithm without performance guarantee. Future work will focus on a comparison to other algorithms, alphabets and fitness criteria to refine the methods presented in this work.

Acknowledgments

We thank Prof. P. Schuster for excellent facilities, W. Hart, K. Dill (at PMMB IV), W. Fontana, M. Vingron for useful discussions, P. Stadler and I. Hofacker for computational help and Sean ODonohue for proofreading. We are indebted to 2 (of 3) referees who helped with constructive criticism to refine the work by commenting its contents and not the language.

[o] This was termed *shape space covering* [14,11].

References

1. H.S. Chan and K.A. Dill. *Physics today*, **2:**24, 1993.
2. K.F. Lau and K.A. Dill. *Macromolecules*, **22:**3986, 1989.
3. K.A. Dill, *et al. Protein Science*, **4**:561, 1995.
4. K.A. Dill, *Biochemistry*, **29**:7133, 1990.
5. C. Levinthal. *J. Chim. Phys.*, **65**:44, 1968.
6. D. Shortle. *Protein Science*, **7**:991, 1996.
7. M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, 1983.
8. J. Maynard-Smith. *Nature*, **225**:563, 1970.
9. P. Schuster. *J. Biotech.*, **41**:239, 1995.
10. W. Fontana, *et al. Phys. Rev. E*, **47**:2083, 1993.
11. W. Fontana, *et al. Biopolymers*, **33**:1389, 1993.
12. M. Huynen, *et al. Proc. Natl. Acad. Sci. USA*, **93**:397, 1996.
13. M. Tacker, *et al. Eur. Biophys. J.*, subm, 1996.
14. P. Schuster, *et al. Proc. R. Soc. Lond. B*, **255**:279, 1994.
15. K. Yue, *et al. Proc.Natl.Acad.Sci.USA*, **92**:325, 1995.
16. E.I. Shakhnovich and A. Gutin. *J. Chem. Phys.*, **93**:5967, 1990.
17. E. I. Shakhnovich and A.M. Gutin. *Proc. Natl. Acad. Sci., USA*, **90**:7195, 1993.
18. M. Karplus and A. Sali. *Curr. Opn. Struct. Biol.*, **5**:58, 1995.
19. C. Chothia. *1992*, **357**:543, Nature.
20. G. M. Crippen and V. N. Mairov. In *Proc. Pacific Symposion on Biocomputing 96.* L. Hunter, T. Klein *eds.*:160, World Scientific, 1996.
21. S. Govindarajan and R. A. Goldstein. *Proc. Natl. Acad. Sci. USA*, **93**:3341, 1996.
22. A. Renner, *et al. preprint*, 1996.
23. N. Madras and A. D. Sokal. *J. Stat. Phys.*, **50**:109, 1987.
24. E. Bornberg-Bauer. PhD thesis, University of Vienna, 1995.
25. R. Unger and J. Moult. *Bull. Math. Biol.*, **55**:1183, 1993.
26. R. Unger and J. Moult. *J. Mol. Biol.*, **231**:75, 1993.
27. P. Stolorz. *Proc. 27th Hawaii Intl. Conf. on System Sciences*, 1994.
28. J. E. Solomon and D. Liney. *Biopolymers*, **36**:579, 1995.
29. W. E. Hart and S. C. Istrail. *J. Comp. Biol.*, **3**:53, 1996.
30. E. Bornberg-Bauer. In *Proceedings of German Conference on Bioinformatics.* Leipzig, 1996.
31. A.J. Guttmann and J. Wang. *J. Phys. A: Math*, **24**:3107, 1991.
32. A.J. Guttmann. *J. Phys A: Math*, **22**:2807, 1989.
33. D. J. Lipman and W. J. Wilbur. *Proc. R. Soc. Lond. B*, **245**:7, 1991.