CHAIN GROWTH ALGORITHMS FOR HP-TYPE LATTICE PROTEINS

ERICH BORNBERG - BAUER

Abteilung Theoretische Bioinformatik, Deutsches Krebsforschungszentrum, Heidelberg
bornberg@dkfz-heidelberg.de , Im Neuenheimer Feld 280, D - 69 120 Germany
and
Institut für Mathematik, Universität Wien, A - 1090 Wien, Austria

Abstract

We describe a novel fast straightforward folding algorithm for **HP** (hydrophobic - polar) type lattice proteins. It is designed after the concept of unguided, cotranslational folding of a nascent peptide. It is deterministic and runs in $\mathcal{O}(n)$ in the chain length. Accuracy of prediction is governed by the search depth of the algorithm that is "looked ahead" at each chain growth step. Long range interactions are significantly increased and energy barriers become less prohibitive with increasing search depth. The efficiency of sequential folding is tested and results compared to related methods. All characteristics of the **HP**-model such as formation of a hydrophobic core and overall compact structures are observed. Since the procedure is very fast and flexible we obtain a useful tool to approximate the sequence to structure mapping of biopolymers in general and to study the complex interplay of folding strategies, potentials and alphabets with large ensembles of random structures.

Keywords: Lattice models, protein folding, cotranslational foldability, chain growth algorithm, look ahead

1 Background

1.1 Protein Folding

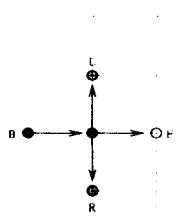
In biological settings proteins generally fold to a unique 3-D structure. It is commonly assumed that only the sequence of amino acids determines this "native" structure and that it corresponds to the equilibrium minimum free energy (MFE) state (the *thermodynamic hypothesis*). The search space is astronomically large, yet proteins fold in seconds. This suggests that proteins do not search all possible configurations - the so called *Levinthal-paradox*. Therefore folding most likely follows some strong and possibly deterministic rules encoded in the sequence. If so, these rules should comply to a mechanistic framework of reproducible processes. A remarkable number of possible models was proposed during the last decades to solve this notorious *Protein Folding Problem* [1], (Some popular hypotheses are: 2-stage molten globe collapse, diffusion collision, hydrophobic zipper, nuclear condensation, nucleation propagation etc.) still a clear and consistent concept is unknown [8]. This is cumbersome, since the ability to determine the structure *ab initio* (i.e. without knowledge beyond sequence and solvent properties), might give rise to a solution for a large number of pharmaceutical and biotechnological problems. Traditional computer simulations suffer from the lack of proper potential functions and the tremendous need for resources. Several strongly simplified models have therefore been derived during the last decade to investigate at least the most basic principles that govern the protein folding process. They are also useful to study basic principles of functional adaptation and natural foldability, i.e. the ability to also attain the functional state within a reasonable time and following a series of defined events. Lattice Proteins are one of these and will be described in short in the next section.

Our motivation to study sequential folding in the **HP** - framework [14] arises from a number of findings. It represents a subprocess of several models; these are the folding of sub-domains *in vitro* and the early steps of forming a nucleus or locally ordered structures. One of the most intriguing concepts for reasoning about folding *in vivo* is known as *cotranslational folding*. One can easily imagine that folding of a nascent peptide chain starts as soon as the N' terminus is extruded to the lumen and becomes solvent exposed while the C' end is still shielded in the ribosome One of the earliest assumptions for this was given by Levinthal [24]. The algorithms to be presented here are modeled after this view of folding in vivo. They also provide estimates for the need of optimization - expressed by the "look ahead " parameter - such that cotranslational folding successfully yields stable structures. They also serve as a tool to investigate the influence of short and long range interactions. This is crucial to explain not only specifity and stability of structures but also the foldability since the need to *optimize* for long range interactions - whatever their nature may be - during a folding procedure implies a less intuitive and probably more difficult search problem.

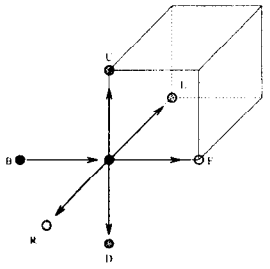
At the same time the method is fast enough to be applicable for statistical investigations on large ensembles of structures

RECOMB 97, Santa Fe New Mexico USA

[1] There are actually several aspects to the problem: 1) To design (a) sequence(s) that fold(s) into a target structure with a desirable function (*inverse folding*) 2) to convey a consistent method to predict the structure from sequence information only (*structure prediction*) 3) to describe the biological pathway of folding in full detail (*folding pathway*) 4) to explain the uniqueness of a structure encoded in the sequence (*encoding problem*) etc.



	B	F	L	R
f	b	f	l	r
l	r	l	b	f
r	l	r	f	b
b	f	b	r	l



	B	F	L	R	U	D
b	f	b	r	l	d	u
f	b	f	l	r	u	d
l	r	l	u	d	f	b
r	l	r	d	u	b	f
u	d	u	f	b	l	r
d	u	d	b	f	r	l

Figure 1: a) Relative moves on the square lattice and c) on the simple cubic lattice. b) Encoding relative moves for the square lattice: relative moves are given in capitals, lower case letters refer to absolute moves (see text for description). Given the prior move, read from the leftmost column, the current absolute move is searched in the line and the corresponding relative move is read from the uppermost line. e) Encoding scheme for the simple cubic lattice.

so that we obtain a tool to investigate the complex interplay of range of interactions, folding mechanisms and potentials. This and biochemical implications will be discussed more in detail elsewhere.

1.2 Lattice Proteins

Lattice proteins are abstractions of biopolymers: residues are represented at a unified size by placing each, but at most one at a time on one bead of a regular lattice. Bond lengths are unified since they are represented by vertices, bond angles are discrete. Lattice proteins have become a valuable tool to address basic questions of the sequence to structure relation in biopolymers. This is due to the easy computational implementation, the well-definedness of structure representation and the reduced search space. Several different models are known (see e.g. [33, 36, 1, 20, 14, 43]). For an excellent review and critical evaluation of current methods the reader is referred to Dill *et al.* [14]. Following Dill we use the well known subclass of **HP**-models [23, 14] strictly as a model system to investigate specific questions concerning the sequence to structure relationship and generic properties of biopolymers in general but not as a realistic representation of proteins.

1.3 Related Work

Since the *Lattice Protein Folding Problem* was reported to be NP-hard [41, 17, 28] a large variety of approximation algorithms was proposed. Unger and Moult [42] developed a resource intensive algorithm based on principles of genetic algorithms and Monte Carlo techniques in the square lattice with excellent results for fairly long chains ($n = 60$). Stolorz [38] presented a method to approximate the whole density of states (DOS). It utilizes an approach somewhat similar to the algorithms presented here with a "window" that is shifted through the sequence and recursively counts up low energy states. Further algorithms are the hydrophobic zipper model, proposed by Fiebig *et al.* which starts with a random **HH** contact and zips up the rest [15, 14]. Another, branch and bound - like algorithm designed by Yue *et al.* [45] constructs an ensemble of cores that fits within some, theoretically approximated, bounds and follows some constraints. It is extremely successful but computationally demanding [45]. None of these algorithms seem to be fast enough for statistical investigations of large ensembles of random structures.
Hart and Istrail [19] recently presented an algorithm for the **HP** - model that guarantees structures with energy better than 3/8 of optimum. Accuracy depends strongly on the lattice and energy function. It produces not very compact chains by applying a hierarchical folding model that, as they claim, is compatible with the diffusion collision mechanism. It is also deterministic, works in $\mathcal{O}(n)$ but does not consider different potentials and cross-space interactions.
Several attempts have utilized Monte Carlo-techniques for a chain-growth procedure. Moves are accepted or rejected following the Rosenbluth and Rosenbluth criterion [32]. These approaches basically aim to generate an ensemble of chains with a large fraction of very low energy states based on the Boltzmann like distribution of states. They are useful to investigate thermodynamic properties such as transition temperatures [37, 29].

2 Methods

2.1 Lattices and Relative Moves

Using relative moves is well established (see e.g. [23]). We use a version that can be applied to any regular lattice [2]. A detailed description will be given [31].
A regular Lattice $\mathcal{L}$ can be characterized as the set of basis-vectors of equal length. Characteristic lattice constants are the dimension d, the number of nearest neighbors c^* (i.e. the lattice coordination number), possible moves c and the number of allowed moves $\hat{c}$. Basis-vectors for a lattice (embedded in euclidian space) are called (absolute) moves on the lattice $\mathcal{M}_{\mathcal{L}} = \{m_1, ..., m_c\}$. For the square lattice SQ we have $\mathcal{M}_{SQ} = \{m_1 = [0,1] := f, m_2 = [0,1] := l, m_3 = [-1,0] := r, m_4 = [0,-1] := b\}$. Let $x_{i+1} = x_i + m$ denote the lattice point obtained by attaching the move m to site x (e.g. the position of the i-th residue in a chain), then there is a move m^* such that $x_i = x_{i+1} + m^*$. For all lattice points x, y and all moves m', m'' there is a symmetry operation ψ on the lattice such that $y = \psi(x)$ and $y + m'' = \psi(x + m')$. A walk on a lattice for a sequence of length n is completely described by its initial point, $x_0 = 0$, and the ordered list of the $n-1$ moves. Let us denote $\mathcal{M}'_{\mathcal{L}}(x_i) \subseteq \mathcal{M}$, the set of all possible moves at a point x_i (e.g. on the hexagonal lattice, there are only 3 moves allowed, depending on the "class" of points under consideration: $\mathcal{M}'_{HEX}(x_i) = \{b, f, l\}$ or $\mathcal{M}''_{HEX}(x_{i+1}) = \{r, u, d\}$ but $\mathcal{M}'_{SQ} = \{b, f, l, r\}$ for any x.) One can also define a set of *relative moves* $\mathcal{R} = \{r_1, ..., r_{\hat{c}}\}$ with a corresponding symmetry operation

[2] We call a lattice regular if for any two lattice points x and y there is a symmetry operation σ of the lattice such that $x = \sigma(y)$ and if for any two pairs of neighbors (u, v) and (x, y) there is a symmetry operation τ of the lattice such that $u = \tau(x)$ and $v = \tau(y)$.

$\mathcal{E}_{ij}$		H	P
	H	-1	0
	P	0	0
%		48	52

$\mathcal{E}_{ij}$		H	P
	H	-3	-1
	P	-1	-0
%		48	52

$\mathcal{E}_{ij}$		H	P	N	X
	H	-4	0	0	0
	P	0	0	-1	0
	N	0	-1	0	0
	X	0	0	0	0
%		48	12	14	26

$\mathcal{E}_{ij}$		h	H	Y	X
	h	-2	-4	-1	2
	H	-4	-3	-1	0
	Y	-1	-1	0	2
	X	2	0	2	0
%		16	36	10	28

Table 1: Energy potentials **HP**, **HP'**, **HPNX**, **hHYX** for alphabets as implemented in the computer programm. Integer values are used for computational convenience. Numbers in the lowest line denote the frequencies of corresponding amino acids in natural proteins.

Φ such that, when the coordinate frame is rotated after each move r_k, the move $-r_k$ is assigned the backwards direction B in the rotated coordinate system. In the following we use capital letters for relative moves and lower case for absolute moves. One yields e.g. $\mathcal{R}'_{HEX}(x) = \{B, A, C\}\forall x$ and $\mathcal{R}'_{SQ}(x) = \{B, F, L, R\}\forall x$. For self avoiding walks we only need $z = m - 1$ relative directions, since the backwards direction B never occurs, hence $\mathcal{R} = \mathcal{R}' \setminus B$. Since the coordinates of the point x_k obtained at the k-th step of the walk are given by $x_k = x_{k-1} + m_k$, the absolute direction m_k of the k-th step is uniquely defined by the relative direction r_k of this step, *and* the absolute direction of the previous step, m_{k-1}. Consequently, there is a mapping $\tau : \mathcal{M} \times \mathcal{R} \to \mathcal{M}$, $(m, R) \mapsto \hat{R}^{-1}m$ for each lattice determining the absolute direction of a step given its relative direction and the absolute direction of the previous step. Once we have defined τ we can represent a walk on the lattice as a string of length $n - 1$ over $\mathcal{R}$.
We use relative moves for convenience since the method has several advantages over representing structures by absolute moves or coordinates:

1. Point mutations are pivot moves which proves the ergodicity of any two self avoiding walks [25].
2. Algorithms for folding and structure comparison can be programmed independently from the chosen lattice [7].
3. Concatenation of strings corresponds to elongation of the first walk.
4. Storage requirements are kept small.
5. Structure comparison can be achieved with classical string comparison methods: Hamming distance and sequence alignment define a metric distance measure in shape space [4].

2.2 Potentials

The generalized energy function for a sequence with n residues $S = (s_1, s_2, ..., s_n)$ with $s_i \in \mathcal{A}$, the alphabet of 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

$$E(S, X) = \sum_{i}^{n} \sum_{j \geq i+2}^{n} \mathcal{E}(s_i, s_j) d_{ij}^{\alpha} f(s_i, s_j, |i - j|) \qquad (1)$$

where $d_{ij} = ||x_i - x_j||$ is the Euclidian distance, $\mathcal{E}_{ij} = \mathcal{E}(s_i, s_j)$ is a pair-potential retrieved from the energy matrix. In our implementation contributions can be considered up to a certain cutoff distance: $d_{ij}^{\alpha} = 0$ if $d_{ij} >$ cutoff, α in general is -1. The function f respects the dependency of distance within the sequence and takes on 1 in all calculations presented in the following.
We implemented four different potentials: the "classical" **HP** - model, a more sophisticated one **HP** and the two four - letter alphabets (**HPNX**) and (**hHYX**). empirical potential on a square lattice. (For consistency we used **HP** and *cutoff = 1* whenever direct comparison to Dill's model was considered and *cutoff = 1.8* else. The latter includes diagonal interactions on the plane and through space e.g. in a simple cube.) In the **HP**-model [23, 14] the spectrum of various inter-atomic forces is reduced to one inter residue interaction, the *hydrophobic force*: it is assumed that this unspecific force is the dominant contribution to stability and therefore to a large extent determines the 3D structure of the backbone. (Random) Heteropolymers are composed from a two-letter alphabet $\mathcal{A}$ = { **H**, **P** } where there is 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 overall energy. The model is a crude abstraction but captures several salient features of real protein structures: 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).
HP sequences in general show a large structural degeneracy i.e. the number of configurations that correspond to one lowest energy state [14, 45]. For $n = 18$ on a square lattice only ca. 2.4% of all chains fall into a unique ground state. This is remarkable since real proteins, synthesized from binary patterns, also frequently appear with little structural specifity [11]. A detail study on the influence for 2 letter alphabets was recently given by Chan and Dill [9].
The **HP'** alphabet includes a stronger overall attracting force.
The **HPNX** - set is a generic extension of the **HP** model, inspired by the electrostatic interactions between a negatively charged residue (**N**) and one with a positive charge (**P**) as well. It is useful to investigate the influence of a larger alphabet and the consequence of adding a less dominant force. The **HX** subset corresponds exactly to the **HP** model. The **YhHX**-set is a modified form of Crippen's empirical potential [12] which consists of four classes of residues (and originally 4 classes of separation along the chain). Energy matrices and frequencies of corresponding residues in natural proteins are given in table 1. Comparison in between the different potentials will be reported in future work.

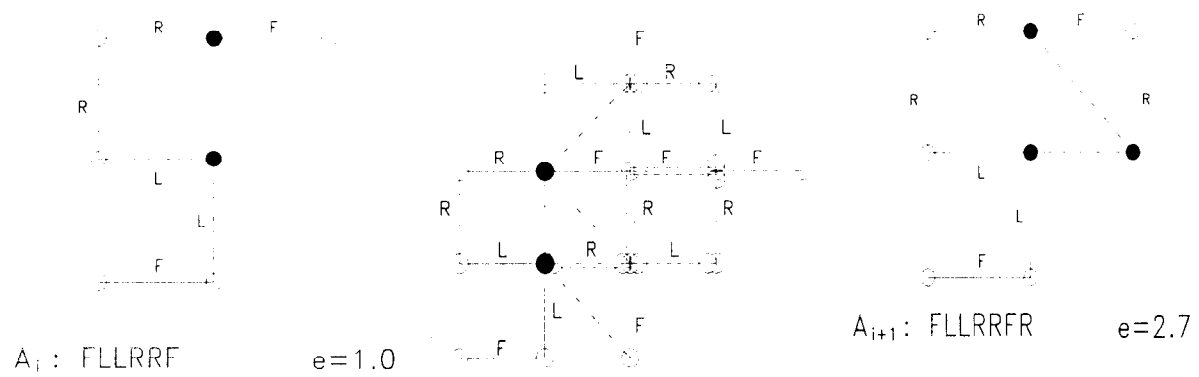


Figure 2: The gCGA: A sequence $S = (PPHPPHPHH)$ $(n = 9)$ is folded (search depth $m = 2$, cutoff $u = 1.9$): a) 7 residues are already "frozen" (i.e. the description starts with $k = 5$ and $X = (FLLRRF)$) black dots are **Hs**, circles symbolize **Ps**, full lines bonds of the "frozen core" and long dashed lines energy contributions within u. b) All possible moves (short dashed lines) are tried: One of the next $c^m = 9$ possible solutions (RR) is forbidden and hence yields an infinite energy value. One configuration (RF) gives the "best" energy (e=3.4) and therefore c) the next residue with the move R is appended. This configuration $X = (FLLRRFR)$ is used for the next iteration step if the chain was longer, in this case "RF" completes the procedure to $X = (FLLRRFRF)$. (Note that in this case we yield the maximum compact state, which is however degenerate with respect to the associated energy e.g. to $X = (FLLRRLRR)$.)

3 Chain Growth Algorithms (CGA)

Following our view of sequential folding in vivo we implemented some versions of a *chain growth algorithm (CGA)*. In the simplest version the algorithm is very fast. There is of course a trade-off between accuracy and speed which frequently comes with NP-completeness. The common principle is described in the following, an example is illustrated in Fig. 2, pseudo code notions are appended below.
In the following assume a given configuration at the k-th iteration step as a list of relative moves $X_{1,k} = (r_1, r_2, \ldots, r_k)$, $r_i \in \mathcal{R}_\mathcal{L}$. After starting with an initial move, all $z = \hat{c}^m$ possible configurations that can be formed from the search depth m (the "look - ahead parameter") are generated and temporarily appended. The corresponding energies of these overall configurations are evaluated following Eqn. 1. The lowest energy configuration is chosen and from this appendix the first p moves only (in general $p = 1$) are appended to the "frozen core" and never detached again. If there are more than one configuration with equal lowest energy ("degenerate states") they are lexicographically sorted with respect to a default hierarchy of the move list. This configuration is then used for the next iteration. This procedure is repeated till the complete chain is scanned. Occasional traps (e.g. $E^\star > 0$) are escaped by backtracking and searching for alternate solutions (which is very unlikely on 3D-lattices) or instances are regarded as misfolds and eliminated from the sample.

In the "greedy" version the chain growth step is realized deterministically and termed gCGA hereafter. In the stochastic version sCGA the next move is appended from a configuration that is chosen not lexicographically but with a probability that follows a Gibbs distribution. k is the Boltzmann constant and T is the "temperature", i.e. a tunable parameter to calibrate the randomization of the algorithm. For $m = 1$ and $T \to 0$ the sCGA degenerates to the gCGA except for the default hierarchy. If $p > 1$ moves are appended at once we yield for the stochastic version an algorithm ($msCGA$) similar to the one proposed by O'Toole [29]. They used a *3sCGA* on a simple cubic lattice for ensemble studies. (If the ratio n/p yields no integer the last iteration in the chain growth scheme is modified for a shorter p.)

Method ChainGrowth $(S(n), m, p)$

start with forward move $r_1 := F$;
while $(k \leq n - m - 2)$ **do**
  $(r^\star_{k+1}, \ldots, r^\star_{k+m}) :=$
  $SelectConfig(S(n), (r_1, \ldots, r_k), m)$;
  for all $(i = 1, \ldots, p)$ **do**
    $(r_{k+i} := r^\star_{k+i})$;
  end for
  $k := k + p$
end while
output $X = (r_1, \ldots, r_n), E(X)$;

Method greedy

Procedure SelectConfig$(S(n), (r_1, \ldots, r_k), m)$

for all $\hat{c}^m$ configurations
$X^l_{1,k+m} = (r_1, \ldots, r_k, r^l_{k+1}, \ldots, r^l_{k+m}), r^l_i \in \mathcal{R}_\mathcal{L}$ **do**
  $E^l_{1,k+m} := EvaluateEnergy(S(n), X^l_{1,k+m})$;
end for
output lexicographically first $(r^l_{k+1}, \ldots, r^l_{k+m})$ with minimal $(E^l_{1,k+m})$;

Using *tries* [22], it is possible to efficiently keep track of all neighboring positions of residues that have already been scanned. Evaluating energy contributions from neighbors for all newly appended residues is executed n times. Scanning through all appended configurations of course demands an effort that is exponential with m. The overall time requirements scale with $\mathcal{O}(n \times \hat{c}^m)$.

4 Computations

4.1 Finding the ground state in 2D

We tested the algorithm for 3189 sequences (data kindly supplied by P.Stolorz) with a non-degenerate ground-state on

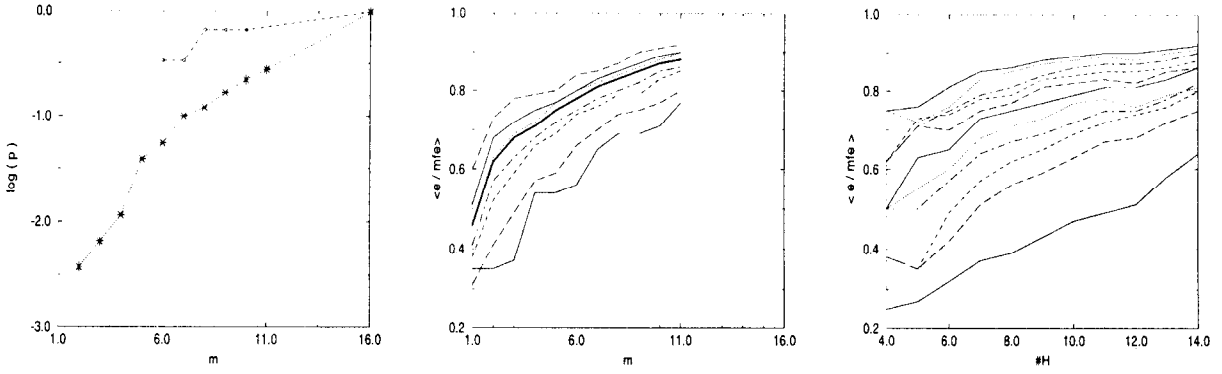


Figure 3: Performance of the gCGA. a) Stars: the success rate (log(p) of correctly predicted structures) for all 3189 sequences of length $n = 18$ with a non-degenerate ground-state on the square lattice vs. the search depth. Triangles denote selected values for structures without a first excited state ("energy gap") which fold much easier. b) The relative performance (gCGA results divided by ground state energies) at various search depths m. Results are shown for different ground-state energies from $mfe = 9$ (uppermost, dashed line) to $mfe = 4$ (full, lowest line). The thick (full) line stands for the average mfe value. c) The relative performance vs. the number of hydrophobic residues for search depth $m = 11$ (uppermost line) to $m = 4$ (lowest line).

Method stochastic

Procedure SelectConfig(r, m)

.....

let $(r^{\star}_{k+1}, \ldots, r^{\star}_{m}) = (r^{l}_{k+1}, \ldots, r^{l}_{k+m})$;

with probability $[\frac{exp(-E^l/kT)}{\sum_l exp(-E^l/kT)}]$

output $(r^{\star}_{k+1}, \ldots, r^{\star}_{m})$

---

a square lattice with $n = 18$ (see Fig. 3). The success-rate in achieving the ground state folding from *one* end scales roughly linearly with the exponent of the search depth (see Fig. 3a), yet a remarkable jump can be observed at $m \approx 5$ which is probably due to the possibility of forming small folding units of that size. We think that the success rate is quite good if one considers the crudeness of the **HP** model and the low connectivity of the square lattice. Since the core is "frozen", it must be assumed that the performance decreases strongly with longer n but increases with higher dimension of the lattice. We are, however, only interested in ranges of n that may correspond to independent folding units since it has often been argued, that - especially two dimensional models - actually resemble larger units of residues [14]. Hence it is interesting to see that the gCGA algorithm works especially good when sequences contain more hydrophobic residues (Fig. 3c). This is an interesting contrast to the algorithm from Hart and Istrail that was proposed to performs better the smaller the **H** content is (W. Hart, personal communication). Also the average energy is nearly always above 50% of the optimum which indicates the success to find a structure that is pretty low in the density of states spectrum [38]. Especially at small m a significant improvement can be achieved when the MFE is low (Fig. 3b). The 3 sequences with an "energy gap", i.e. no structure that corresponds to the energy level next to the ground state, fold nearly an order of magnitude easier. According to the small sample size this is certainly not statistically significant. Yet is is interesting that this feature, which has been claimed to represent a necessary and sufficient criterion for fast folding in Monte Carlo simulations of a lattice model [33] can also be found in a simple model as presented here.

4.2 Random Structures in 3D

We also tested the gCGA for long chains on a 3D lattice. No large data sets of ground states are available for comparison. Used length of $n = 125$ is considerable compared to other work [45, 41, 38, 29, 37] and we do not expect to find the groundstate with a CGA. We considered a set of 1000 random structures on the simple cubic lattice folded at increasing search depth $m = 1, 2, 4, 6$. Ensemble data [4] are shown in Fig. 5. Increasing m in general lowers energy, increases compactness and the number of contacts. The major reason for improved efficiency is that, the more the polymer "looks ahead", the deeper an energetic trap during the folding procedure can be overcome [4]. The number of contacts except **HH** stays rather constant which indicates that improvement results from formation of a tighter core. **PP** - contacts are on the surface and, as are the expectations of the **HP** - model, "solvent" exposed without being explicitly penalized. For $n = 125$ on a simple cubic lattice for ~62 **H**s e.g. one can estimate a maximum of 176 overall contacts (**HH, HP, PP**). In the sample most structures exhibit ca.. 140 contacts, some instances approach 155 for $m = 6$. **HH**-contacts are ≈ 68 with special instances up to ≈ 75. (These instances of course result from random sequences with more than 50% **H**.) Assuming a perfect core yields an upper limit of $\sim$ 77 **HH** interactions when ~50% are **H**s. This is of course also a natural limit for the MFE and shows we are in the range of 85% - 90% of the optimum energy and compactness with only looking 6 residues ahead for a chain of length 125. One should keep in mind that in general the MFE - structure is not necessarily maximally compact [14] and a perfect core in general not possible [14, 19] since some space can not be optimally packed and only residues with odd separation along the chain can form contacts. Consequently the estimate is actually much better. In Fig. 4 we illustrate the influence of the m for a selected example: long range interactions are significantly increased to a much higher extent than m itself. This shows that, even with local optimization long-range interactions may be important and achieved, even if not explicitly aimed for. Obviously it is relatively simple to form structures with a compact core

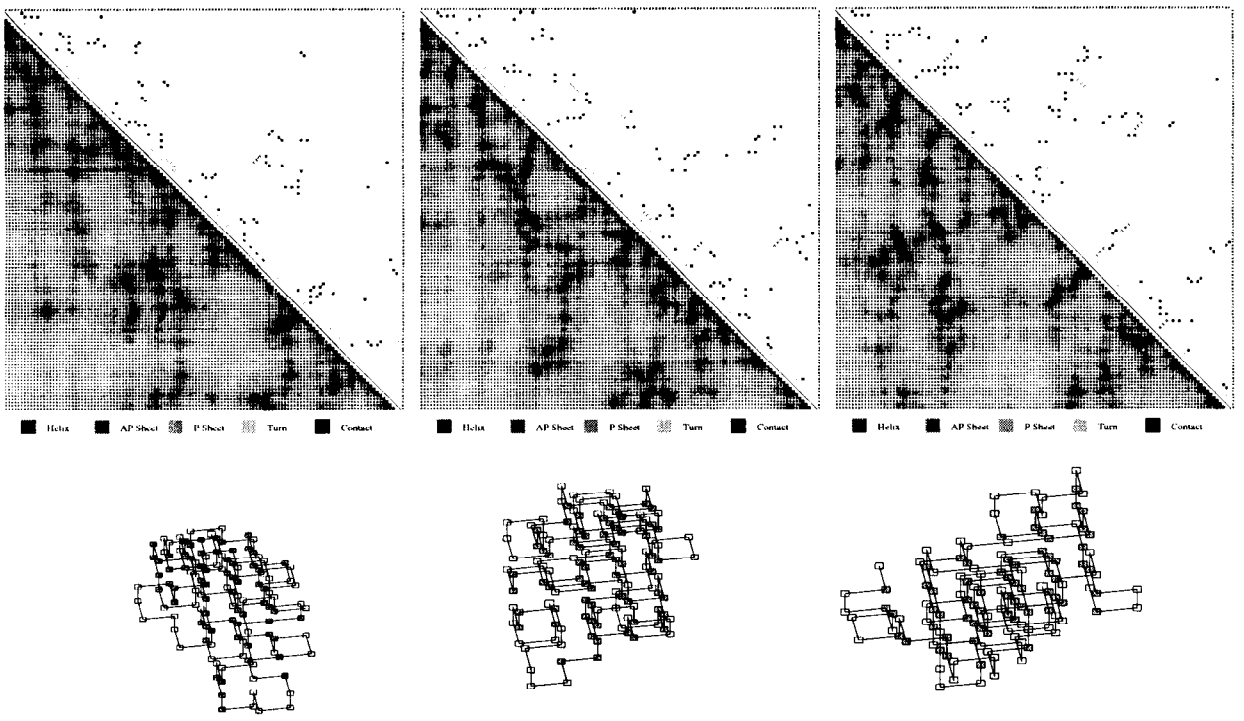


Figure 4: A random sequence, folded on the simple cubic lattice with the gCGA (**HP** - alphabet, u=1.9, n=125) with "look-ahead" m=1 (left, 115 contacts), 2 (middle, 136 contacts) and 4 (right, 139 contacts). Dark cubes symbolize **H**, white ones **P**s. The upper triangle is the contact map: secondary structural elements are indicated in grey shades as annotated. The lower triangle is the distance matrix, large squares denote short distances. Long range interactions are increased with larger m and short range, locally ordered structural elements begin to arise.

and many long range interactions that stabilize an overall configuration.

A major caveat for the gCGA is the default hierarchy. From large ensembles of random structures, however, it can be seen that the hierarchy is suppressed after a relatively small core (i.e. ca. 25 residues for a simple cubic lattice) is frozen for small m (ca. 4-6). The distribution of moves is then unbiased and it can be assumed that the influence of the move hierarchy on the structure formation is only marginal compared to the folding constraints encoded in sequence and potential function (data were reported in [4]).

5 Discussion

5.1 Conclusions

5.1.1 Performance

We presented simple algorithms with tunable accuracy. At a relatively small expense reasonable solutions can be found as shown for large ensembles of random structures in 3D and comparison to 2D structures with known ground states.

Compared to other methods there are some advantages and some caveats: on the credit side we have the speed, the easy implementation, the applicability to any potential, lattice and the possibilities to regard cross-space interactions as well as to tune the desired accuracy. Furthermore the strength of our methods lies in the applicability to study the complex interplay between folding strategies, potentials and alphabets for large ensembles of 3D model of biopolymers. For that purpose, the technique is - at least to our knowledge - the only one available at the moment. Its major drawback is certainly the decreasing efficiency for longer chains and the lack of a lower bound. We expect the algorithm is more efficient for lattices with a higher dimension since the frozen core then represents a less stringent obstacle for further improvements.

5.1.2 Structure specifity

Results can of course only be viewed as a very crude approximation of realistic processes of protein folding, details on structure similarities or comparison to other algorithms are currently under investigation. However, since the **HP** - model itself is very coarse grained, the crudeness of the folding process may seem adequate. If one aims to really determine the global minimum of structures by the means of more sophisticated techniques, the **HP**-model would probably not be the right choice since then a number of interactions which is not accounted for come into play. **HP** sequences in general exhibit a large structural degeneracy of

the ground state energy [14, 41, 45]. This is probably due to the fact that the **HH** force is a very unspecific one [45]. Consequently, in comparison to random-energy-like models [33] or contact potential derived methods [1] a **HP** sequence with a non-degenerate ground-state might correspond to a structure that, viewed at with a more sophisticated potential exhibits a very pronounced minimum and hence a relatively wide energy gap (see also section 4.1). From a dynamic point of view coarse grained "native states" then actually become ensembles of finer grained structures with a large number of fluctuations [18, 45].

5.1.3 Consistency with folding in vivo

From a biological point of view gCGA type folding can be interpreted as the case of an unguided, straightforward folding event of a nascent chain at a ribosome, where the backbone formation is determined by the hydrophobic pattern. Nowadays there is plenty of evidence that cotranslational folding may indeed play an important role for folding in vivo. These ideas (recently reviewed in [35]) were also supported by the detection of greater structural compactness and stronger homologies of the N' region (see [2, 40, 26] and Refs. therein). Though not surprising from a computational point of view, the performance can also be viewed as an approximation to test foldability since these sequences yield stable, fast folding structures. They may serve as starting points for further evolutionary optimization under the fitness constraint of cotranslational foldability. These structures may then be reshaped during evolutionary optimization following other fitness criteria. significant amount of structures folds to the global minimum in a straightforward way.

5.1.4 Possible implications for protein evolution

The relatively high success rate may be meaningful in an evolutionary context: the **HP** pattern would give rise to a reasonable selection criterion - maybe independent from the folding strategy. Further fine-tuning would be achieved by selecting the right packing by choosing proper side-chains. This shows that the *foldability* (which is, in our context, the ability to fold into a functional native state in a fast and straightforward way) represents no obstacle but may rather be a reasonable selection criterion for random sequences in the **HP** framework. As these fold in a straightforward manner to the ground state they may be selected to satisfy the constraint of the thermodynamic hypothesis. This is particularly meaningful if one considers that non-degenerate ground states may correspond to very pronounced minima if viewed at the same structure with a more fine grained potential.

5.1.5 Comparison to other models and in vitro folding theories

We think that this view of folding might also be compatible to some concepts of folding *in vitro*, namely that of a nucleation process of small sub-domains (e.g. following the simple arguments from Wetlaufer [44]) that rapidly propagates along the chain. It is, however, unlikely to be relevant for folding of whole proteins. Yet it is interesting to see that even such a simple, straightforward procedure can be quite successful.

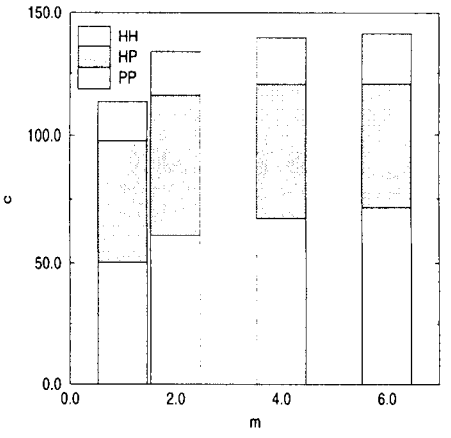


Figure 5: The number of **HH**, **HP** and **PP** contacts as a function of the search depth m. A sample of 1000 random sequences with the **HP** alphabet and length n=125 was folded on the simple cubic lattice.

An interpretation for folding *in vitro* is difficult since literature is vast and inconsistent. There are, however, several experimental findings and some theoretical considerations that are compatible: 1) Local interactions dominate folding [44, 27] (this does, of course not mean that long-range interaction may not be responsible for the *specifity* of the overall fold). Similar conclusions were recently reported by Unger and Moult from a lattice model [43]. 2) Local nucleation appears to be a good approximation for the early steps of folding for realistic folding domains: following the simple arguments from Wetlaufer [44] about the size of nuclei, a rough estimation of 8 - 18 residues was given. 3) Clearly enough the speed is well within the requirements of *Levinthal Paradox*. A combination with other, globally optimizing procedures such as the algorithm from Hart and Istrail [19] might be reasonable to improve the efficiency. 4) The dominance of the **HP** pattern for the overall architecture [14, 13, 21] is widely accepted now. 5) Local optimality is maybe not sufficient for global optimality but makes it easier. This can also be observed in other models [38, 42]. Combining these aspects puts us into the position to assume a reasonable relevance at least for small, independently folding nuclei.
Realistic folding in the sense that there is an overall optimization criterion (e.g. with explicit implementation of long-range interactions and respecting side chains) however is beyond both, the limits and the intention of the introduced method.

5.2 Further Applications

The deterministic algorithm appears to be - within the narrow limits of the **HP**-lattice-protein model - efficient enough to apply to statistical investigations of the sequence to structure map. Recently the techniques of *landscapes* and *complex combinatory maps* have been successfully applied to characterize the sequence to structure map for *RNA secondary structures* [16, 6, 34, 39]. There the average depen-

dency of e.g. structure- , energy, etc. similarity (*phenotype properties*) are related to the sequence similarity (*genotype property*), e.g. the Hamming-distance between the underlying sequences. Since the strength of the gCGA lies in its speed and hence the applicability for statistical investigations we will follow this line and investigate the influence of mutations on the fold-ability and the success rate with respect to structural similarities for several other alphabets. Some recent results from statistical investigations on large ensembles of random structures with different lattices and alphabets were recently reported [3, 30]. There are several remarkable findings that are important to understand biopolymer evolution: 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. Larger alphabets reduce this degeneracy and energies and structures become equally sensitive towards mutations [30]. In spite of significant changes on single structure properties, ensemble properties are hardly influenced by the search depth. Furthermore we find few very frequent and many rare structures. This is significant since it implies that few structures dominate the ensemble and can be more easily found from many different starting sequence during any evolutionary optimization process. The structure distribution is analogous to the abundance of certain fold classes among "real world" proteins [10].
These ensemble studies show a striking similarity to RNA secondary structures [34, 5, 16]. This suggests that these features – neutrality, proximity of structures in sequence space and the distribution of structure frequencies following Zipf's law – are generic properties of the sequence to structure map of biopolymers in general and of simple exact models in particular.

Acknowledgments:

Work was started during my PhD at the Institute for Theoretical Chemistry, University of Vienna, with Prof. P. Schuster who provided excellent facilities. Computer package is joint work with A. Renner. Stimulating discussions with W. Fontana, M. Vingron, W. Hart and B. Schuler, help by P. Stadler and I. Hofacker and useful comments by B. Schwikowski and R. Spang are gratefully acknowledged.

References

[1] V. Abkeviech, A. Gutin, and E. Shakhnovich. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemestry*, 33:10026 – 10036, 1994.

[2] N. Alexandrov. Structural argument for N-terminal initiation of protein folding. *Protein Sci.*, 2:1989 – 1991, 1993.

[3] E. Bornberg-Bauer. Simple folding model for hp lattice proteins. In *German Conference on Bioinformatics*, volume forthcoming of *Lecture Notes in Computer Science*. Springer.

[4] E. Bornberg-Bauer. *Random Structures and the Evolution of Biopolymers*. PhD thesis, University of Vienna, 1995.

[5] E. Bornberg-Bauer. Random structures and evolution of biopolymers: A computational case study on RNA secondary structures. *Pharm. Acta Helv.*, 71:79–85, 1996.

[6] E. Bornberg-Bauer. Structure formation of biopolymers is complex, their evolution may be simple. In L. Hunter and T. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing*, pages 97 – 108. World Scientific, London, 1996.

[7] E. Bornberg-Bauer and A. Renner. tofolapo-0.07. available upon request from bornberg@dkfz-heidelberg.de.

[8] H. S. Chan and K. A. Dill. The protein folding problem. *Physics today*, 2:24 – 32, 1993.

[9] H. S. Chan and K. A. Dill. Comparing folding codes for proteins and polymers. *Proteins Struct. Funct. Gen.*, 24:335 – 344, 1996.

[10] C. Chothia. One thousand families for the molecular biologist. *Nature*, 357:543 – 544, 1992.

[11] M. H. Cordes, A. R. Davidson, and R. T. Sauer. Sequence space, folding and protein design. *Curr. Opn. Struct. Biol.*, 6:3 – 10, 1996.

[12] G. M. Crippen. Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry*, 30:4232 – 4237, 1991.

[13] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133 – 7155, 1990.

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

[15] K. M. Fiebig and K. A. Dill. Protein core assembly process. *J. Chem. Phys.*, 98:3475 – 3487, 1993.

[16] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Phys. Rev. E*, 47:2083 – 2099, 1993.

[17] A. Fraenkel. Complexity of protein folding. *Bull. Math. Biol.*, 55:1199 – 1210, 1993.

[18] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscape and motions of proteins. *Science*, 254:1598 – 1603, 1991.

[19] W. E. Hart and S. C. Istrail. Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. *J. Comp. Biol.*, 3:53 – 96, 1996.

[20] D. A. Hinds and M. Levitt. Exploring conformation space with a simple lattice model for protein structure. *J. Mol. Biol.*, 243:668 – 682, 1994.

[21] E. S. Huang, S. Subbiah, and M. Levitt. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.*, 252:709 – 720, 1995.

[22] D. E. Knuth. *The Art of Computer Programming, Vol 3*. Addison Wesley, Reading, MA (USA), 1973.

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

[24] C. Levinthal. Are there pathways for protein folding? *J. Chim. Phys.*, 65:44 – 45, 1968.

[25] N. Madras and A. D. Sokal. The pivot algorithm: A highly efficient monte carlo method for the self avoiding-walk. *J. Stat. Phys.*, 50:109–186, 1987.

[26] E. V. Makeyev, V. A. Kolb, and A. S. Spirin. Enzymatic activity of the ribosome-bound nascent polypeptide. *FEBS Letters*, 378:166 – 170, 1996.

[27] J. Moult and R. Unger. An analysis of protein folding pathways. *Biochemistry*, 30:3816 – 3824, 1991.

[28] J. T. Ngo and J. Marks. Computational complexity of a problem in molecular structure prediction. *Protein Eng.*, 5:313 – 321, 1992.

[29] E. M. O'Toole and A. Z. Panagiotopoulos. Monte carlo simulation of folding transitions of simple model proteins using a chain growth algorithm. *J. Chem. Phys.*, 97:8644 – 8652, 1992.

[30] A. Renner and E. Bornberg-Bauer. Exploring the fitness landscapes of lattice proteins. In L. Hunter and T. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing*. World Scientific, London, in press.

[31] A. Renner, E. Bornberg-Bauer, I. L. Hofacker, and P. F. Stadler. Self-avoiding walk models for non-random heteropolymers. *preprint*.

[32] M. N. Rosenbluth and A. W. Rosenbluth. Monte carlo calculation of the average extension of molecular chains. *J. Chem. Phys.*, 23:356 359, 1954.

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

[34] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc.Roy.Soc.(London)B*, 255:279–284, 1994.

[35] D. Shortle. Protein folding for realists. *Protein Sci.*, 7:991 – 1000, 1996.

[36] J. Skolnik and A. Kolinski. Dynamic monte carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J. Mol. Biol.*, 221:499 – 531, 1991.

[37] J. E. Solomon and D. Liney. Exploration of compact protein conformations using the guided replication monte carlo method. *Biopolymers*, 36:579 – 597, 1995.

[38] P. Stolorz. Recursive approaches to the statistical physiscs of lattice proteins. *Proc. 27th Hawaii International Conference on System Sciences*, 1994.

[39] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Algorithm independent properties of RNA secondary structure predictions. *Eur. Biophys. J.*, in press.

[40] T. Thanaraj and P. Argos. Ribosome-mediated translation pause and protein domain organization. *Protein Sci.*, 5:1594 - 1612, 1996.

[41] R. Unger and J. Moult. Finding lowest free energy conformation of a protein is an NP-hard problem: Proof and implications. *Bull. Math. Biol.*, 55:1183 – 1198, 1993.

[42] R. Unger and J. Moult. Genetic algorithms for protein folding simulations. *J. Mol. Biol.*, 231:75 – 81, 1993.

[43] R. Unger and J. Moult. Local interactions dominate folding in a simple protein model. *J. Mol. Biol.*, 259:988 994, 1996.

[44] D. B. Wetlaufer. Nucleation, rapid folding and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci., USA*, 70:697 – 701, 1973.

[45] K. Yue, M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill. A test of lattice protein folding algorithms. *Proc. Natl. Acad. Sci., USA*, 92:325 - 329, 1995.