Protein Folding by Hydration Aided Search of Minimum Energy

JARMO T. ALANDER
Department of Information Technology and Production Economics
University of Vaasa
P.O. Box 700, FIN-65101 Vaasa
FINLAND
Jarmo.Alander@uwasa.fi http://www.uwasa.fi/

Abstract: The main result of the computational experiments made in this work is that fitness landscape smoothing has a significant effect on search efficiency. Our test case is related to the famous Levinthal's paradox i.e. how a protein molecule finds its global energy minimum. The energy surface is smoothed by a simple hydration model, which also results in longer range interactions. Hydration also brings more specific correlation between free energy and conformational compactness when compared to the popular lattice models thus aiding conformational search in a way that is not only computationally but also physically sound. Results were compared to those gotten using both the basic local HP model and global Coulombic type models. As a conclusion we can state that based on empirical observations our hydration model implies correlation between energy and compactness, which is beneficial for folding.

Key-Words: fitness landscape, optimisation, protein folding, search

1 Introduction

A *protein* molecule is a chain of *amino acids* linked together by *peptide bonds* i.e. it is an organic *heteropolymer*. The role of *globular proteins* in cells is mainly to function as highly specific and efficient chemicals e.g. *enzymes* and *antibodies* of the *immune system* of *vertebrates*. About half of the proteins of *Escherichia coli* consist of thousands of enzymes [1]. The rate enhancements produced by enzymes can be extremely high, in excess of 10^{10} [2].

The three-dimensional (native) conformation of protein is called *tertiary structure*. It is widely assumed that the globular shape of globular proteins is due to hydrophobic interactions of the nonpolar atoms tending to create a hydrophobic core covered by polar residues in contact with the solvent (water).

Typically a globular protein consists of 1,000–20,000 atoms and has a diameter of about 35-100Å. In small proteins about half of the atoms are located at the surface while in larger proteins the number falls below 20 percent. Especially in many small proteins *disulfide bridges* largely contribute to the stability by binding distant parts of the protein together.

Levinthal's paradox: "If a protein is to find its functional conformation by wandering randomly through conformational space, in excess of 10^{50} years would be required for folding." [3, 4, 5, 6]

The number of protein conformations is approximately 8^n, where n is the number of residues i.e. each residue has on the average 8 possible conformations. It has been estimated that the number of proteins that could possibly have existed during evolution on Earth is somewhere between 10^{40} and 10^{50} [7, 8].

2 Lattice models

Formally a folding lattice model of a protein can be seen as a finite set V of configurations, an *energy function* $f : V \to \mathbb{R}$, and the concept of neighbourhood between the configurations, which means that V is actually a vertex set of a graph Γ, the configuration space, and edge set E defined by the neighbourhood relation. [9]

One of the most reduced protein models is the HP model, in which the chain consists of only two types of residues, hydrophobic (H) and polar (P) on a regular 2D or 3D grid, which can further be rectangular or hexagonal. Eaton E. Lattman, Klaus M. Fiebig and Ken A. Dill define H amino acids to be *Ala*, *Ile*, *Leu*, *Met*, *Phe*, *Trp*, *Tyr*, *Val*, and *Cys* while the rest they define to be polar (P) [10]. After this definition every protein sequence can be mapped into a corresponding sequence in HP model space.

The *Hamiltonian* H of the simple HP model can be written in the form

$$H(\vec{p},\vec{s}) \;=\; -\vec{p}\cdot\vec{s} \;=\; \frac{1}{2}(|\vec{s}-\vec{p}|^2-\vec{p}^{\,2}-\vec{s}^{\,2}),$$

where $\vec{p}$ is a vector telling whether the ith residue is hydrophobic (1) or not (0) and $\vec{s}$ is a vector telling whether the ith site is in the core (1) or on the surface (0). Hence only local interactions are explicitly included in the Hamiltonian, while the longer range interactions only present themselves in the form of steric constraints i.e. as the self avoiding walk (SAW) condition.

The simple HP model is extremely simple. There is no doubt that it lacks many vital features of real proteins. Luckily it is quite easy to make it somewhat more general. In this work we have generalised it by a simple hydration model.

2.1 Fitness landscapes

There are several eminent similarities among the fitness landscapes of complex systems, like protein evolution, spin glasses and satisfiability problems. There are also some differences in the dynamics, which have a profound effect on the long-term evolution with respect to local minima. The problem is that the evolution, or optimisation, easily stops at local extreme. The more rugged landscape the more local extremes and the shorter the average walk to the nearest extremum. While *noise*, thermal or artificial, is able to help to escape local extremes, it may take too long in practise to use this annealing, real or simulated, scheme when solving complex problems. The situation with protein evolution is totally different. The fitness landscape is rugged, but there is usually always some freedom to move around the local extremal point. This is because the topology of the fitness landscape: because of *neutral mutation*s and the high dimensionality of the sequence space, vast number of sequence combinations, and the short distances between different points, which are actually caused by the high dimensionality.

3 Search method

The key with respect to search efficiency is to combine stochastic and deterministic approaches and in addition to utilise the results already evaluated i.e. to benefit computer memory facility by concentrating search to the neighbourhood of already encountered good trials. This schema is depicted in figure 1: search is done under stochastic control, which inputs best trials from memory to a routine performing local deterministic search. At the beginning of processing the pool of trials is filled with default starting conformation.

The random search procedure is modified so that it keeps the best solutions in memory and tries to find better trials by updating the known best trials. This kind of population based approach is used e.g. in *simulated annealing* and *genetic algorithms*.

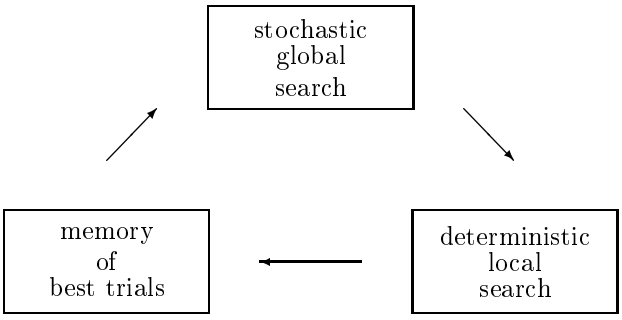

Figure 1: A hybrid conformation search approach.

3.1 Conformation optimisation

Conformation trials are saved to a population where they are after a random number of steps retrieved by a subroutine called `loadConformation`. The population is initialised by the starting conformation, which is usually the most elongated one. The starting conformation is set elsewhere

in the program—usually explicitly supplied by the user. Trials, which are better than the average of the trials, are saved to the population. Corner and crankshaft moves together with totally random point mutation of the conformation are used to find better solutions.

4 Hydration in a lattice model

Hydration can be seen as a process aiding protein folding in several ways. Firstly it mediates the interaction between hydrophobic residues over distances considerable in atomic scale. Secondly hydration inevitably filters, both temporally and especially spatially, interaction potentials. Both phenomena effectively smooth the otherwise very rugged free energy landscape, and thus make the folding process faster. In a way the hydration layer can also be seen to act much as a lubricant between rough surfaces, which reduces friction.

One type of *global optimisation* method is based on smoothing the object function $f(r)$ by replacing it by a transformed function $f(r,t)$, where t is a control parameter that determines the extent of smoothing. The parameter t is initially set to a large value and then slowly reduced until $t = 0$, at which point the object function is "in focus"[1] and the global minimum is likely to be revealed. This kind of *smoothing algorithm* is actually a deterministic analog of the stochastic *simulated annealing*. Smoothing is also what the next represented method based on a simple hydration model of protein does.

In a conventional lattice model only interactions between a residue and its six nearest neighbours are taken into account when evaluating the free energy. Here we will simulate hydration by interactions depending on a larger neighbourhood having radius of r_H lattice units.

Let us consider the way of evaluating hydration interaction between residues i and j. We have the following principal cases for the distance $D(i,j)$ between residues i and j:

- $i = j \pm 1$ i.e. the bonded case, which will be omitted in our energy calculations,
- $D(i,j) \leq r_H$ i.e. the distance $D(i,j)$ between residues i and j is less or equal to the radius of the environment and thus we have to evaluate the energy taking also hydration into account, and
- $D(i,j) > r_H$ i.e. the interaction between residues is omitted due to the long distance $D(i,j)$.

The solution adopted here is to simulate hydration iteratively much like running a 3D *lattice automaton* [11]. In practise the solvent type is replaced by a set of solvent unit cubes having a different hydration degree or level. The number of iteration rounds needed is proportional to r_H [12].

The total free energy due to hydration E_h is evaluated by the formula

$$E_h = -\sum_{i \in NN} h_i,$$

where h_i is the hydration level of voxel i and NN is the set of voxels surrounding (nearest neighbours) the protein model.

HPPHPHPHPHPH	12mer
HPPHHPPHHPPHHPPH	16mer
HPPPPHHHHPPHPHPHHHPHPPHHPPH	27mer
PHPPPPPPHPPPPPHPPPPPHPPPPPH	27merPivot
PHPPHPPHPPHPPHPPHPPHPPHPPHP	27merRep
HPPHHPPHHPPHHPPHHPPHHPPHHPPHHPPH	32mer

Table 1: Test sequences

5 Results

The above hydration model was tested using several sequences and hydration parameter combinations. Some of the test sequences can be seen in table 1. The *27merPivot* sequence is the least complex of the three sequences that are 27 residues long. It takes about 40,000 local steps to find its native conformation (see table 2). The dependence on r_H does not seem to be high. This is understandable because the simple structure of this sequence supposedly results in a relatively smooth free energy landscape. The relatively smooth landscape is apparently also the reason for the approximately as fast folding

[1] smoothing is equivalent to defocusing in optical systems

for the Coulombic type interaction. Folding using the basic HP model is about 30 times slower. Obviously the HP model is spending most of its time to find a needle in a haystack, which is known to be an extremely time consuming search task.

Hydration interaction:

w_{HH}	$Q1$	median	$Q3$	N
0.5	22,976	41,861	71,518	160
1.0	23,622	42,507	73,734	160
r_H	$Q1$	median	$Q3$	N
2	33,478	53,797	104,598	80
3	20,832	36,671	58,058	80
4	24,052	40,682	75,003	80
5	19,716	38,169	70,330	80
Psize	$Q1$	median	$Q3$	N
1	19,158	33,595	60,830	80
4	23,956	44,107	73,336	80
16	28,528	43,170	80,916	80
64	25,293	46,546	71,623	80

Basic HP model ($r_H = 0$):

Psize	$Q1$	median	$Q3$	N
1	743,948	1,336,524	time-out	20
4	937,276	1,574,986	time-out	20
16	614,626	1,310,792	time-out	20
64	393,936	957,568	time-out	20
All	665,128	1,310,792	time-out	80

Coulombic type interaction ($E \propto 1/r$):

Psize	$Q1$	median	$Q3$	N
1	19,592	38,187	50,324	10
4	17,812	56,281	76,922	9
16	17,448	41,673	55,192	9
64	17,105	39,300	72,212	9
All	18,399	41,673	62,604	37

Table 2: *27merPivot*: The quartiles ($Q1$, median, $Q3$) of the number of local steps n_e needed to reveal the native state vs. r_H, population size ($Psize$), and weight w_{HH}. For the basic HP model $\max(n_e) = 2 \times 10^6$ steps, for the others $\max(n_e) = 10^6$ steps. N = number of samples.

5.1 Scaling

The above experiments could be extended in many ways. We could collect more statistics on the reference models. But doing this using the current program version and environment would be quite time consuming and that is why it was not included in this work, but left for further studies, which are needed to fully reveal the characteristics of the proposed hydration model and ways to develop it. In any case our simple hydration model was among the fastest model for each test sequence (c.f. tab. 3). Both the basic HP model and the Coulombic model were found to be considerably slower for some of the test sequences.

sequence	Hydration	Basic HP	Coulombic
12-mer	1	3	2
16-mer	1	3	2
27-mer	2	3	1
27merPivot	1-2	3	1-2
27merRep	1	3	2
32-mer	1	2	3

Table 3: Ranking of models vs. test sequences

5.2 Folding as a landscape dependent algorithm

Figure 3 shows schematically how the correlation between free energy E and conformation compactness K varies during protein folding. When the correlation is strongly negative, as it must be at the beginning and also at the end of the process if a stable native state exists, the protein conformation is relatively fast becoming more compact while losing free energy. Ideally a strong negative correlation means efficient hill climbing to the nearest energy minimum. In case of a complex molecule like protein this unfortunately means a high probability of the search getting stuck to a local minimum i.e. a premature convergence. That is why a macromolecule like protein cannot just fall to its free energy minimum conformation. In this case the shortest way to the goal is not the fastest. It is obvious that for most sequences there exists a critical compactness value K_c, which is the highest value still giving for most conformations an energetically and sterically easy access to the native state. From the computational point of view keeping the ensemble below this value is necessary but not sufficient for efficient folding. From the physics point of view passing this value means a high probability of experiencing a glass transition like phase transition.

If the correlation is highly positive, it means that the protein avoids compact states while

it loses free energy i.e. there is a strong tendency to avoid trapping to local energy minima. By definition this zone gives an excellent implementation of backtracking the low energy states. Drift away from solution is sometimes called *deception* and it is usually considered a property making the search for optimum difficult.

The third zone in our schematic figure, corr $\approx$ 0, means that there is no correlation between the free energy and the compactness of the conformation. In practise the protein is performing a pure random search driven by thermal noise.

In order for a protein to fold properly within a finite time scale the folding path must in practise draw away from the hill climbing zone to avoid premature convergence. Once the folding ensemble has passed the critical compactness zone it is more and more difficult to have major reconfiguration changes due to the energy and steric constraints. However, the molecule should be as compact as possible when approaching the deception zone, where it is actually driven more towards less compact conformations than the native state. The price to be paid for these conditions is the slowing down of the search, but to avoid *Levinthal's paradox* the protein must actually slow down the greedy hill climbing search in order to be able to scan the most promising low energy conformations.

6 Conclusions

The effect of hydration on conformation search was tested by searching the native conformations of several test sequences. In general hydration seems to be beneficial. Without it our algorithm needs considerably more local steps. Similarly it seems to be beneficial to count *H-H contacts*, but their weight used in the free energy formula does not seem to be so critical. Hence it seems that both local and longer range interactions should be modelled in order to have a realistic fast folding process. Moreover a small population size seems to be better than a large one. This is easy to explain: the bigger the population size the more time it takes to process the best trials. Population based search is beneficial because search without a population of trials is clearly less efficient than that using a population.

References

[1] James D. Watson, Nancy H. Hopkins, Jeffrey W. Roberts, Joan Argetsinger Steitz, and Alan M. Weiner. *Molecular Biology of the Gene*, volume 1. The Benjamin/Cummings Publishing Company, Inc., Menlo Park, CA, 4 edition, 1987.

[2] Kenneth E. Neet. Enzyme catalytic power minireview series. *The Journal of Biological Chemistry*, 273(40):25527–25528, 2. October 1998.

[3] Cyrus Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique*, 65:44–45, 1968.

[4] Cyrus Levinthal. [the proper reference to Levinthal's paradox]. In P. Debrunner, J. C. M. Tsibris, and E. Münck, editors, *Proceedings of a Meeting held at Allerton House, Monticello, IL*, pages 22–24, Monticello, IL, 1969. University of Illinois Press, Urbana. (ref. in [6]).

[5] John Moult and Ron Unger. An analysis of protein folding pathways. *Biochemistry*, 30(16):3816–3824, 23. April 1991.

[6] Robert Zwanzig, Alexander M. Gutin, and Biman Bagchi. Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America*, 89(1):20–22, 1. January 1992.

[7] Manfred Eigen. *Steps Toward Life*. Oxford University Press, Oxford (UK), 1992.

[8] Bonnie J. Strait and T. Gregory Dewey. The Shannon information entropy of protein sequences. *Biophysical Journal*, 71(1):148–155, July 1996.

[9] Robert Happel and Peter F. Stadler. Canonical approximation of fitness landscapes. *Complexity*, 2():53–58, 1996. (`http://www.tbi.univie.ac.at/~studla/publications.html`).

[10] Eaton E. Lattman, Klaus M. Fiebig, and Ken A. Dill. Modeling compact denatured states of proteins. *Biochemistry*, 33(20):6158–9166, 24. May 1994.

[11] John von Neumann. *Theory of Self-Reproducing Automata*. University of Illinois Press, Urbana, 1966.

[12] Jarmo T. Alander. Extending HP lattice model with non-local hydration. In Satoru Miyano, Ron Shamir, and Toshihisa Takagi, editors, *Currents in Computational Molecular Biology*, volume 30 of *Frontiers Science*, pages 118–119, Tokyo (Japan), 9.-11. March 2000. Universal Academy Press, Inc.

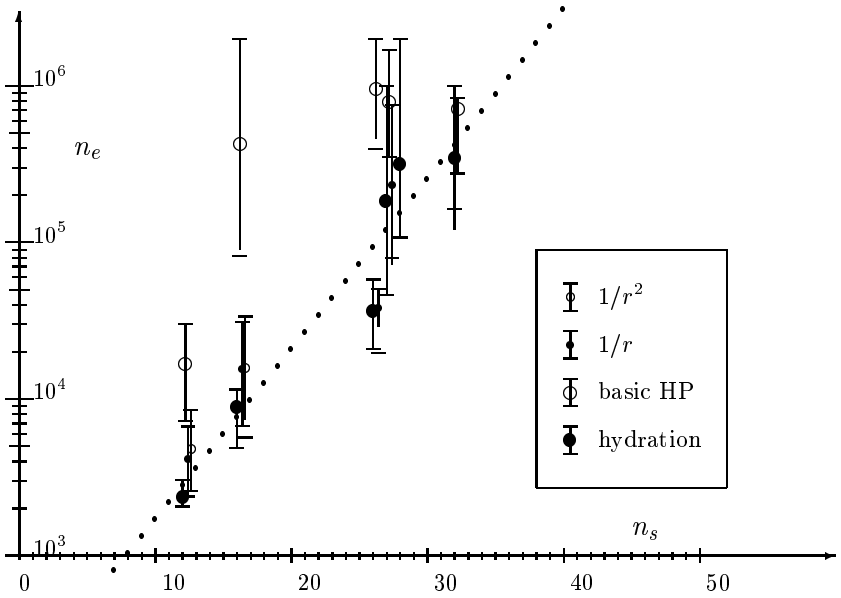

Figure 2: The median number of local steps n_e vs. sequence length n_s The dotted line shows the least squares estimate $N_0 \times 10^{n_s/n_{10}}$, where $N_0 \approx 140$ and $n_{10} \approx 9.2$. As a comparison the dotted curve is $C \times n_s^{3\frac{2}{3}}$, which is a *power-law* based estimate. For clarity different model shown shifted.

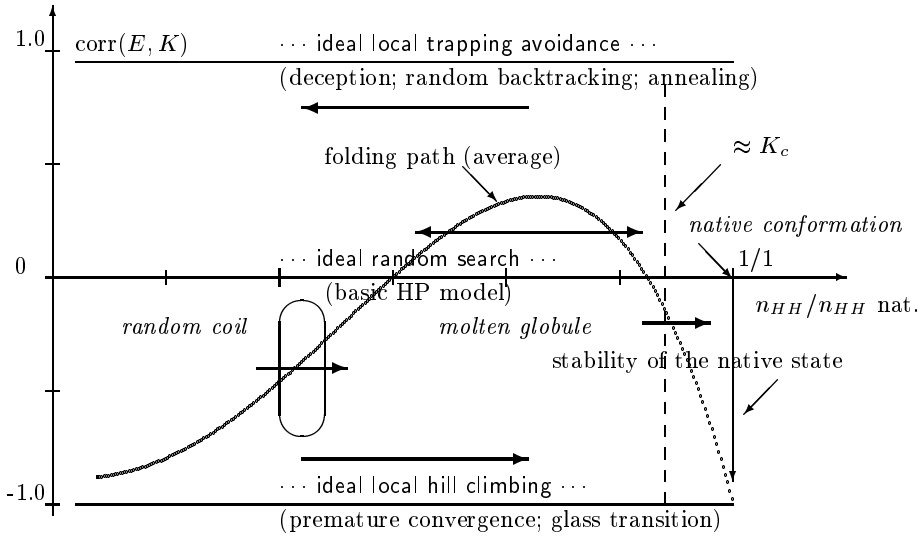

Figure 3: Interpretation of the protein folding path as search algorithms vs. the correlation of energy and conformation compactness K when varying the relative number of H-H contacts n_{HH}/n_{HH} nat. Bold arrows show the direction for decreasing free energy. Notations: K_c = critical compactness and $\cap$ = search ensemble.