Protein Folding Simulation by Two-Stage Optimization*

A. Dayem Ullah[1], L. Kapsokalivas[1], M. Mann[2], and K. Steinhöfel[1]

[1] King's College London, Department of Computer Science, London WC2R 2LS, UK
[2] Bioinformatics Group, University of Freiburg, Georges-Köhler-Allee 106, 79016 Freiburg, Germany

Abstract. This paper proposes a two-stage optimization approach for protein folding simulation in the FCC lattice, inspired from the phenomenon of hydrophobic collapse. Given a protein sequence, the first stage of the approach produces compact protein structures with the maximal number of contacts among hydrophobic monomers, using the CPSP tools for optimal structure prediction in the HP model. The second stage uses those compact structures as starting points to further optimize the protein structure for the input sequence by employing simulated annealing local search and a 20 amino acid pairwise interactions energy function. Experiment results with PDB sequences show that compact structures produced by the CPSP tools are up to two orders of magnitude better, in terms of the pairwise energy function, than randomly generated ones. Also, initializing simulated annealing with these compact structures, yields better structures in fewer iterations than initializing with random structures. Hence, the proposed two-stage optimization outperforms a local search procedure based on simulated annealing alone.

1 Introduction

The question of how proteins fold and whether we can efficiently predict their structure remain the most challenging open problems in modern science. Proteins regulate almost all cellular functions in an organism. They are composed of amino acids connected in a linear chain. These chains fold in three-dimensional space. The 3D structure of proteins, also referred to as tertiary structure, plays a key role in their functionality. According to Anfinsen's thermodynamic hypothesis, proteins fold into states of minimum free energy and their tertiary structure can be predicted from the linear sequence of amino acids [2]. In nature, proteins fold very rapidly, despite the enormous number of possible configurations. This observation is known as the Levinthal paradox and implies that protein folding can not be a random search for the global minimum [20].

One of the driving forces of folding, mainly in globular proteins, is the hydrophobic interaction, which tends to pack hydrophobic amino acids in the center of the protein. This effect can be captured by the HP model, a coarse

* Research partially supported by EPSRC Grant No. EP/D062012/1.

Z. Cai et al. (Eds.): ISICA 2009, CCIS 51, pp. 138–145, 2009.

grained model, where the twenty different amino acids are classified into two classes, namely hydrophobic and polar [13]. Protein structure prediction is an NP-complete problem in the HP model [9]. Consequently, one can resort to constraint programming and stochastic local search to tackle this problem. Both techniques are commonly used to approach NP-complete problems.

Previous approaches using local search methods for protein structure prediction include tabu search [19,6], simulated annealing [3], and a population-based local search method [16]. Constraint programming techniques have been successfully applied to the protein structure prediction [4,11] as well as to resolve protein structures from NMR data [18].

The constraint programming approach, employed in [4], predicts the optimal structure of a protein in the HP model in very short computational time. Nevertheless, it is computationally intractable for more elaborate energy functions such as a 20 amino acid pairwise interactions energy function. Local search approaches, on the other hand, work well in practice for elaborate energy functions, despite the large number of iterations required. In this paper we aim to combine the advantages of both approaches.

We introduce a protein folding simulation procedure that employs two stages of optimization in order to find structures of minimum energy. The input protein sequence first collapses to a compact structure and then a slower annealing procedure follows to find the minimum energy structure. Specifically we employ the Constraint-based Protein Structure Prediction (CPSP) tools introduced in [22] to obtain an HP model conformation with maximal number of contacts among hydrophobic monomers in the FCC lattice. Then the CPSP output is given as input to a simulated annealing-based local search procedure which employs the pairwise energy function introduced in [7]. The choice of the FCC lattice is motivated by the fact that it was shown to yield very good approximations of real protein structures [23]. Also it does not suffer from the bipartiteness of the cubic lattice, which allows interactions only between amino acids of opposite parity in the chain. The two-stage optimization introduced in this paper, produces better conformations with less computational cost than local search alone that starts with randomly generated initial structures.

This paper is organized as follows. In Section 2 we first give the outline of the two-stage optimization as well as useful definitions for the detailed illustration of the method. In Section 3 we present experimental results for benchmarks along with a discussion. Finally, Section 4 contains the concluding remarks.

2 The Two-Stage Optimization

The approach works in three phases, namely, the sequence conversion, the constraint programming and the local search. The sequence conversion phase takes as input a protein sequence in the 20 letter amino acid alphabet and returns a converted sequence in the HP model. The resulting sequence consists of two kinds of amino acids, namely, hydrophobic and polar. Let us denote the input sequence as S_{orig} and the resulting sequence as S_{HP}. The constraint programming

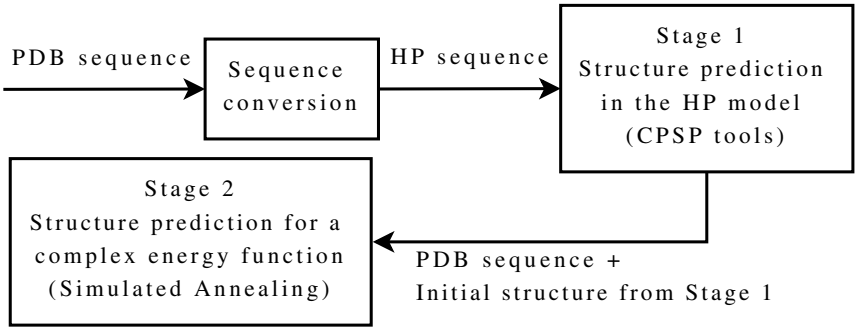


Fig. 1. An outline of the two-stage optimization

phase is the first stage of optimization and utilizes the CPSP tool *HPstruct* for optimal protein structure prediction in the HP model, introduced in [22]. The input to this tool is a sequence S_{HP}. For each sequence S_{HP}, the tool provides a set of structures in the FCC lattice with maximal number of H-monomer contacts. Let us denote this set of structures as $\mathcal{L}_{HP}$ to distinguish it from a set of randomly generated structures $\mathcal{L}_{rand}$. The final phase is the local search, which is also the second stage of optimization. It employs the simulated annealing algorithm with the pull-moves for triangular lattices and optimizes a more complex energy function. During this phase, each sequence S_{HP} is converted back to its original composition S_{orig}, first. Then, the local search is executed for a number of iterations with an initial structure from $\mathcal{L}_{HP}$. In the subsections below we analyze each phase of the approach.

2.1 Sequence Conversion

The conversion of a 20 letter amino acid sequence into an HP sequence utilizes the approach in [10]. This approach establishes a classification of the 20 amino acids into hydrophobic and polar as the result of a hierarchical clustering applied to the Miyazawa-Jernigan [21] pairwise contact values. In Table 1 we give the classification of amino acids used in the sequence conversion phase.

Table 1. Amino acid classification for sequence conversion

Hydrophobic	Polar	
C - Cysteine	H - Histidine	N - Asparagine
F - Phenylalanine	A - Alanine	D - Aspartic Acid
I - Isoleucine	T - Threonine	E - Glutamic acid
L - Leucine	G - Glycine	K - Lysine
M - Methionine	P - Proline	
V - Valine	S - Serine	
W - Trytophan	Q - Glutamine	
Y - Tyrosine	R - Arginine	

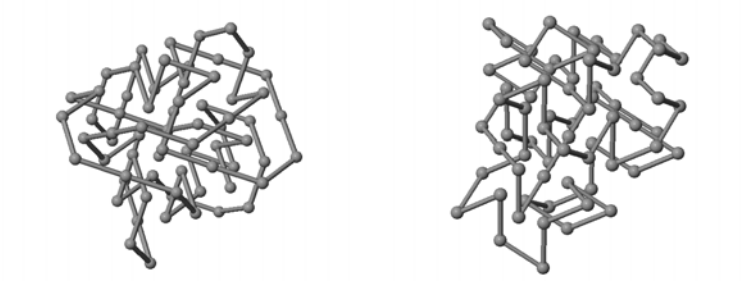
2.2 Constraint-Based Optimal Structure Prediction in HP-Models

The Constraint-based Protein Structure Prediction (CPSP) approach enables the calculation of optimal structures in 3D HP-models [22]. Using its implementation *HPstruct* [22], we enumerate for a given sequence S_{HP} a representing set of optimal structures $\mathcal{L}_{HP}$, all showing a compact hydrophobic core and shape.

The CPSP approach utilizes a database of precalculated (sub)optimal H-monomer placements (H-cores) [5]. These are sequence independent, defined by the number of H-monomers and the lattice. For a sequence S_{HP}, a self-avoiding walk describing constraint satisfaction problem is formulated that in addition constrains the H-monomers of the sequence to a given H-core. Any solution yields an optimal solution of the optimal structure prediction problem [4]. A screen through all appropriate H-cores enables the prediction of all optimal structures. For details on the CPSP approach see [4].

2.3 Local Search

Simulated annealing was introduced as an optimization tool independently in [17,8] (see also [1]). The algorithm traverses the space of conformations employing the pull-move neighborhood relation in triangular lattices [6]. The objective function to be optimized is the empirical contact potential described in [7], which is a 20 amino acid pairwise interactions energy function. A logarithmic cooling schedule is employed which was shown to converge to optimal solutions [14].



(a) (b)

Fig. 2. 1CTF: (a) Structure produced by CPSP tool. (b) Predicted structure by two-stage optimization.

3 Experiments

Let us now describe the benchmark selection and the protocol we followed in our experiments. The protocol serves the purpose of a fair performance comparison between the two-stage optimization presented above and an optimization procedure, based on local search alone. Although the new approach involves the CPSP tool in addition to local search, in practice the CPSP tool's runtime is very short and can be neglected [22]. Thus, the performance of each method depends on

Table 2. Benchmark sequences from Protein Data Base (PDB) and the derived HP-sequences

PDB id:	4RXN
length:	54
S_{orig}:	MKKYTCTVCGYIYDPEDGDPDDGVNPGTDFKDIPDDWVCPL CGVGKDEFEEVEE
S_{HP}:	HPPHPHPHHPHHHPPPPPPPPPPHPPPPPHPPHPPPHHHP HHPHPPPPHPPHPP
PDB id:	1ENH
length:	54
S_{orig}:	RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEA QIKIWFQNKRAKI
S_{HP}:	PPPPPHPPPPHPPHPPPHPPPPHHPPPPPPPHPPPHPHPPP PHPHHHPPPPPPH
PDB id:	4PTI
length:	58
S_{orig}:	RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNF KSAEDCMRTCGGA
S_{HP}:	PPPHHHPPPHPPPHPPPHHPHHHPPPPPHHPPHHHPPHPPPPPPH PPPPPHHPPHPPP
PDB id:	2IGD
length:	61
S_{orig}:	MTPAVTTYKLVINGKTLKGETTTKAVDAETAEKAFKQYANDNGVDGVW TYDDATKTFTVTE
S_{HP}:	HPPPHPPHPHHHPPPPHPPPPPPPPHPPPPPPPPPHPPHPPPPPHPPHH PHPPPPPPHPHPP
PDB id:	1YPA
length:	64
S_{orig}:	MKTEWPELVGKAVAAAKKVILQDKPEAQIIVLPVGTIVTMEYRIDRVRLFVD KLDNIAQVPRVG
S_{HP}:	HPPPHPPHHPPPHPPPPPHHHPPPPPPPPHHHHPHPPHHPHPHPHPPHPHHHP PHPPHPPHPPHP
PDB id:	1R69
length:	69
S_{orig}:	SISSRVKSKRIQLGLNQAELAQKVGTTQQSIEQLENGKTKRPRFLPELASALG VSVDWLLNGTSDSNVR
S_{HP}:	PHPPPHPPPPHPHPHPPPPHPPPHPPPPPPPHPPHPPPPPPPPPHHPPHPPPH PHPHPHHHPPPPPPPHP
PDB id:	1CTF
length:	74
S_{orig}:	AAEEKTEFDVILKAAGANKVAVIKAVRGATGLGLKEAKDLVESAPAALKEGVSK DDAEALKKALEEAGAEVEVK
S_{HP}:	PPPPPPPHPHHHPPPPPPPHPHHPPHPPPPPHPHPPPPPHHPPPPPPHPPPHPP PPPPPHPPPHPPPPPPPHPHP

the energy reached, given a limited number of iterations for the local search. In other words, we examine the performance of simulated annealing with $\mathcal{L}_{HP}$ as the set of initial structures versus its performance with a randomly generated set of initial structures $\mathcal{L}_{rand}$.

Table 2 shows the benchmark sequences that we used for our experiments and their corresponding derived HP sequences. Benchmarks 4RXN, 4PTI, 1R69 and 1CTF are taken from [7]. In [7], the authors show that the empirical contact potential we employ in our approach, is able to discriminate the native structures of these 4 benchmarks. Benchmarks 1ENH, 2IGD and 1YPA are taken from [11].

For each protein sequence we performed 10 independent local search runs starting with random initial structures ($\mathcal{L}_{rand}$). Then we performed 10 independent runs for the two-stage approach where the initial structures for the local search phase are taken from the CPSP tool *HPstruct*, namely, the set of

Table 3. Comparison between the two-stage optimization and the local search alone

PDB id.	Length	Method	Avg S.E.	Avg F.E.	B.E.	Avg It.
4RXN	54	SA-only	-2.405	-161.625	-165.401	1,019,588
		2-stage	-140.377	-164.483	-167.781	816,844
1ENH	54	SA-only	-2.395	-149.456	-152.747	926,785
		2-stage	-127.347	-151.36	-153.098	904,368
4PTI	58	SA-only	-3.4799	-208.969	-215.698	1,056,287
		2-stage	-179.196	-210.357	-212.500	652,600
2IGD	61	SA-only	-2.5611	-178.941	-180.893	1,160,557
		2-stage	-163.201	-182.564	-183.205	706,773
1YPA	64	SA-only	-3.1447	-252.556	-256.017	1,004,750
		2-stage	-236.895	-256.504	-257.81	1,142,827
1R69	69	SA-only	-3.055	-202.338	-215.166	1,073,051
		2-stage	-188.966	-216.708	-219.402	1,001,264
1CTF	74	SA-only	-1.804	-221.713	-228.921	1,176,490
		2-stage	-176.088	-231.225	-233.86	1,043,517

Avg S.E. - Average Start Energy
Avg F.E. - Average Final Energy
B.E. - Best Energy Observed
Avg It. - Average Iterations

structures $\mathcal{L}_{HP}$. The number of initial structures in $\mathcal{L}_{HP}$ per benchmark was limited to 10 by setting appropriety the argument -maxSol for HPstruct. Each structure was used to initialize an independent run of simulated annealing for the two-stage approach. The initial temperature for simulated annealing in both approaches was set equal to $D * \ln(2)$, where D is an estimation for the maximum depth of local minima of the underlying energy landscape. In a similar fashion to [3], D was set equal to $n^{2/3}/c$, where n is the sequence length and c was chosen to be 1.5. Moreover, the maximum number of iterations of each run in local search phase was limited to 1,500,000 for both approaches.

In Table 3, for each protein sequence, the first row corresponds to the results observed from local search alone, whereas the second row corresponds to the results observed from the two-stage optimization. Figure 2 shows the best initial structure provided by the CPSP tools and the best structure obtained by the two-stage optimization for benchmark 1CTF.

As we can see in Table 3, the average energy of $\mathcal{L}_{HP}$ structures for the empirical contact potential is up to two orders of magnitude lower than the average energy of $\mathcal{L}_{rand}$ structures. We observe that, given the same maximum iteration limit to both approaches, the two-stage optimization always leads to conformations of lower energy on average compared to simulated annealing alone. Also, the two-stage optimization reached lower best energy conformations within the time limit for all benchmarks except 4PTI. Moreover, it requires on average less number of iterations to produce conformations within the average final energy level, except for benchmark 1YPA. In general, the two-stage optimization

approach outperforms simulated annealing alone, since it reaches better final conformations in fewer iterations for the majority of benchmarks.

4 Conclusions

In this paper we introduced a two-stage optimization approach for protein folding simulation which combines the advantages of Constraint-based Protein Structure Prediction (CPSP) and local search. CPSP is very efficient for the HP model but computationally infeasible for a 20 amino acid pairwise interactions energy function. At the same time, local search methods are applicable to the problem, despite the considerable amount of computational effort required. Experimental results with PDB sequences show that the CPSP tool *HPstruct* produces compact structures, whose energy for the pairwise energy function is up to two orders of magnitude better than the energy of a randomly generated structure. Further experimentation with a simulated annealing-based local search procedure starting from these compact structures, shows that better structures are obtained in fewer iterations compared to simulated annealing with a random initialization. Hence, the proposed two-stage optimization outperforms a local search procedure based on simulated annealing alone.

References

1. Aarts, E.H.L.: Local search in combinatorial optimization. Wiley, New York (1998)
2. Anfinsen, C.B.: Principles that govern the folding of protein chains. Science 181, 223–230 (1973)
3. Albrecht, A.A., Skaliotis, A., Steinhöfel, K.: Stochastic protein folding simulation in the three-dimensional HP-model. Computational Biology and Chemistry 32(4), 248–255 (2008)
4. Backofen, R., Will, S.: A Constraint-Based Approach to Fast and Exact Structure Prediction in Three-Dimensional Protein Models. Constraints 11(1), 5–30 (2006)
5. Backofen, R., Will, S.: Optimally Compact Finite Sphere Packings - Hydrophobic Cores in the FCC. In: Amir, A., Landau, G.M. (eds.) CPM 2001. LNCS, vol. 2089, pp. 257–272. Springer, Heidelberg (2001)
6. Böckenhauer, H.-J., Dayem Ullah, A.Z.M., Kapsokalivas, L., Steinhöfel, K.: A Local Move Set for Protein Folding in Triangular Lattice Models. In: Crandall, K.A., Lagergren, J. (eds.) WABI 2008. LNCS (LNBI), vol. 5251, pp. 369–381. Springer, Heidelberg (2008)
7. Berrera, M., Molinari, H., Fogolari, F.: Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. BMC Bioinformatics 4, 8 (2003)
8. Cerny, V.: A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. Journal of Optimization Theory and Applications 45, 41–51 (1985)
9. Crescenzi, P., Goldman, D., Papadimitriou, C., et al.: On the complexity of protein folding. Journal of Computational Biology 5, 423–465 (1998)
10. Cheon, M., Chang, I.: Clustering of the Protein Design Alphabets by Using Hierarchical Self-Organizing Map. Journal of the Korean Physical Society 44, 1577–1580 (2004)

11. Dal Palú, A., Dovier, A., Fogolari, F.: Constraint Logic Programming approach to protein structure prediction. BMC Bioinformatics 5(1) (2004)
12. DeLano, W.L.: The PyMOL Molecular Graphics System. DeLano Scientific, Palo Alto, CA, USA (2002), http://www.pymol.org
13. Dill, K.A., Bromberg, S., Yue, K., et al.: Principles of protein folding - A perspective from simple exact models. Protein Sci. 4, 561–602 (1995)
14. Hajek, B.: Cooling schedules for optimal annealing. Mathem. Oper. Res. 13, 311–329 (1988)
15. Herráez, A.: Biomolecules in the Computer: Jmol to the rescue. Biochem. Educ. 34(4), 255–261 (2006)
16. Kapsokalivas, L., Gan, X., Albrecht, A.A., Steinhöfel, K.: Two Local Search Methods for Protein Folding Simulation in the HP and the MJ Lattice Models. In: Proc. BIRD 2008. CCIS, vol. 13, pp. 167–179. Springer, Heidelberg (2008)
17. Kirkpatrick, S., Gelatt Jr., C., Vecchi, M.P.: Optimization by simulated annealing. Science 220, 671–680 (1983)
18. Krippahl, L., Barahona, P.: PSICO: Solving Protein Structures with Constraint Programming and Optimization. Constraints 7(4-3), 317–331 (2002)
19. Lesh, N., Mitzenmacher, M., Whitesides, S.: A complete and effective move set for simplified protein folding. In: Proc. 7$^{\text{th}}$ Annual International Conference on Computational Biology, pp. 188–195. ACM Press, New York (2003)
20. Levinthal, C.: Are there pathways for protein folding? J. de Chimie Physique et de Physico-Chimie Biologique 65, 44–45 (1968)
21. Miyazawa, S., Jernigan, R.L.: Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18, 534–552 (1985)
22. Mann, M., Will, S., Backofen, R.: CPSP-tools - Exact and Complete Algorithms for High-throughput 3D Lattice Protein Studies. BMC Bioinformatics 9 (2008)
23. Park, B.H., Levitt, M.: The complexity and accuracy of discrete state models of protein structure. Journal of Molecular Biology 249(2), 493–507 (1995)