PROTEIN FOLDING

A Benchmark Combinatorial Optimisation Problem

Technical Report 6-02

Jason Brownlee
jbrownlee@ict.swin.edu.au

PhD Candidate
Master of Information Technology, Swinburne University of Technology, 2004
Bachelor of Applied Science, Computing, Swinburne University of Technology, 2002

Centre for Intelligent Systems and Complex Processes
Faculty of Information and Communication Technologies
Swinburne University of Technology
Melbourne, Victoria, Australia

July 2005

Introduction

The intention of this work is to briefly describe the protein folding as a field of study and to define a model employed in the field to protein tertiary structure prediction called the hydrophobic polar (HP) model. It is proposed that problem instances of this model can be used as both a benchmark and application problem domain for a broad range of combinational optimisation algorithms. This work will provide sufficient information for the protein-folding novice to gain a preliminary understanding of the nature of the problem, so that it could be addressed by a broad range of canonical probabilistic approximation based global search techniques.

Proteins and Protein Folding

Proteins are complex organic compounds that are among the most studied molecules in chemistry and its peripheries from the aspect of activity. A protein molecule is composed of a string of amino acids (monomers) that are joined together in a series by peptide bonds. A long string of amino acids is typically referred to as a protein, though shorter sequences are commonly referred to as peptides or polypeptides. Proteins are essential to the functioning of living cells and viruses and typically take on roles such as enzymes, compositional parts of enzymes, as well as broader mechanical and structural roles (on the cellular as opposed to the molecular level).

A protein can be described in terms of its four main structural descriptions, as follows:

1. **Primary Structure:** The amino acid sequence. These sequences act as instructions for proteins, and were an outcome of the human genome project for mapping DNA.
2. **Secondary Structure:** This describes the protein in terms of regular patterns or substructures in the polypeptide backbone (thus independent of the amino-acid side chains or offshoots from the main backbone of the molecule). These substructures are classified as one of three motifs; alpha-helix, beta-sheet and the random coil.
3. **Tertiary Structure:** The three-dimensional shape of the protein molecule. This refers to the spatial relationship between the molecules secondary substructures. A protein is not functional until it is “folded” into its tertiary structure.
4. **Quaternary Structure:** Refers to structures and molecular relationships when protein molecules form in groups to construct larger molecules.

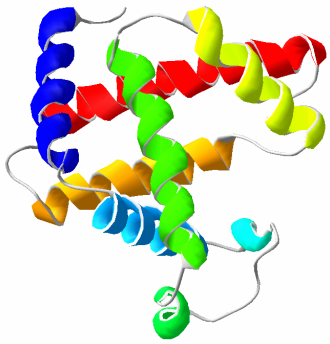

Figure 1 - A 3D representation of a protein’s tertiary structure, the secondary structure is represented as the colour-coded spirals (apha-helies). Picture from [31]

Protein folding [16,22] is the study of the process by which the protein assumes it native conformation (its tertiary structure), and is generally considered one of the most significant tasks in computational biology or bioinformatics. Through learning, more about how proteins fold and function biologists are able to understand more about the development of disease where proteins misfold such as Alzheimer's disease, Mad Cow disease (BSE) and Cystic Vibriosis.

The primary form of the protein defines the three-dimensional tertiary structure the protein will fold into itself. This self-folding requires specific environmental conditions to occur such as the correct pH and in some cases chaperone or helper molecules. Proteins can be unfolded (denatured) using thermal or chemical kinetics, and some proteins can refold into their tertiary structures when the agent is removed. The spontaneous self-folding process can occur in microseconds (millionths of a second), though large molecules can take minutes or hours to fold.

Two large areas of study in protein folding are protein design and protein structure prediction. Protein design is the design of new protein molecules from scratch, which involves devising the set of amino acid sequences that will fold reliably into a native conformation from the larger set of all possible amino acid sequences. Protein structure prediction is an area of study interested in accurately determining the tertiary structure of a protein from its primary structure. The tertiary structure can be determined through conventional structure elucidation experimentally, using X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy to determine the position of atoms within the molecular environment. This is a slow and expensive process (not to mention possible ambiguity), and as a result, biologists have turned to computer algorithms to model the proteins and assist primary to tertiary structure predictive problem.

Structure prediction is hard for a number of reasons; the main arguments are as follows:

1. The number of tertiary structures a protein can express is very large and many structures are functionally not relevant or important
2. The physical bases for protein structural stability is not fully understood
3. The primary structure may not completely define the tertiary structure
4. The direct simulation of proteins is generally intractable

There are many ways in which to address this problem, though the approach of interest to this discussion is called de novo (or ab initio) protein modelling [11]. This approach takes an amino acid sequences as input and searches for or builds tertiary structures for the sequence from scratch. Search approaches are typically stochastic in approach (Monte Carlo) where approximations are desirable given the size of the search space and complex of the problem. Typically, only small protein structures are investigated due the computational resources required.

The Hydrophobic Polar (HP) Model

A thermodynamical approach to the problem of tertiary structure prediction[1] assumes that a folded protein seeks to minimise Gibbs free energy[2]. One of the simplest models of this assumption further assumes that the major contribution to the free energy of a proteins natural conformation is due to interactions between hydrophobic amino acids [20]. That is that amino acids with a hydrophobic residue (not able to form hydrogen bonds - non-polarised) in a protein sequence tend to push together, and that those amino acids with a hydrophilic residue (capable of forming hydrogen bonds - polarised) tend to push to the outside of the molecule [19].

This simple and well studied abstraction is called the hydrophobic polar (HP) model of protein folding [17,18,20,21]. Commonly studied proteins consists of approximately 20 different amino acids. The HP model discards all amino-acid information in the sequence except for two properties, whether the amino acid is hydrophobic (H) or whether the amino acid is polar (P), (also known as hydrophilic). The amino acids are then organised on a two-dimensional or three-dimensional lattice such that the number of topologically adjacent H-H connections is maximised. The number of these hydrophobic connections is calculated and multiplied by a free energy constant, (typically -1) thus transforming the problem to that of minimising the free energy of lattice structures. The amino acid sequence is embedded in a lattice as a self-avoiding walk where a single vertex can be occupied by one amino acid and the walk cannot intersect itself.

Problem Definition: A sequence of length k is given where each $k_i \in \{H, P\}$. The sequence must be embedded as a self-avoiding walk in a discrete lattice with dimensionality d (typically two or three) such that the number of topologically adjacent H's are maximised. The lattice is a regular structure (typically square or cubic) which limits single set movements (nodes) of the walk to graph arcs (edges). For example, the valid movement set for a two dimensional square lattice would be $move \in \{L, R, U, D\}$ (left, right, up down relative to the lattice).

Figure 1 provides an example of a sequence (*PPHPPHHPPHHPPPHHHPPP*) embedded in a two-dimensional square lattice. A white square represents P (polar), and a black square represents H (hydrophobic) amino acids. The start of the sequence is highlighted with a dark border around the first P square. The solution could be specified in lattice relative movements as follows: *LUULDLULDDLDRRURDDL*. The total number of topologically adjacent hydrophobic residues is nine, giving the solution an energy value of -9 (calculated as: $9 \cdot -1$).

[1] Specifically the first law of thermodynamics – the conservation of energy

[2] Free energy is the energy available in the system to do useful work. The Gibbs free energy function is a thermodynamic function commonly used in chemistry to calculate a free energy quantity (G).

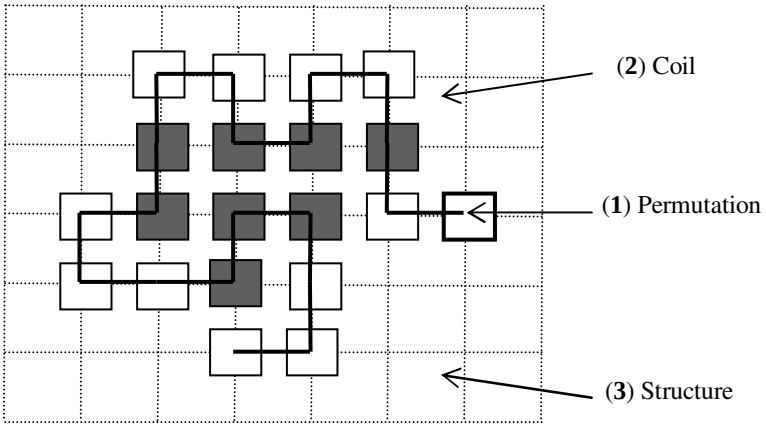

Figure 2 - Example sequence embedded as a self-avoiding walk in a 2D square lattice.

Although this abstraction bares little resemblance to the structure and essence of protein molecules, for the sake of continuality it is useful to point out the three structural descriptions of protein in the context of this model. The primary structure (1) is taken as the sequence (permutation) of hydrophobic (H) and polar (P) amino acid residues. The secondary structure (2) of the model could be interpreted as common structural motifs, such as zigzags and coils. Finally, the tertiary structure (3) is the structure in its entirety embedded in the lattice.

The HP model cannot be used to describe real proteins or demonstrate how real proteins fold. The model is a simple abstraction that reduces all amino acid properties to two binary properties, and all structure properties of the protein to a discrete lattice. A problem with this reduction is that not all amino acids can be described as either hydrophobic or hydrophilic. The model is easy to describe, analyse, and visualise which means that it can studied by both biochemists and non-biochemists. The model can be used by biochemists to gain some insight into the principles of what primary structures can and cannot fold into stable tertiary structures.
The HP model is one of the simplest and most popular biophysical models of protein folding. It does have some parallels with biochemistry theories of energy minimisation and compactness of protein structure by encouraging H-H clustering toward the centre of the structure. The discrete representation permits small proteins to be enumerated in the computer and studied with statistical analysis. Exhaustive enumeration and search of larger sequences is computationally expensive, and has been demonstrated to be intractable (NP-complete) on both a square two-dimensional [25] and a cubic three-dimensional [8] lattices. For a sequence of length k, there are 2^k permutations, and there are approximately 2.6^{k-1} folded structures. It is interesting to note that nature is capable of solving this exponential hard problem in microseconds an area of study which may prove useful for the study of NP-completeness [9].

Although the tertiary structures are conventionally modelled on 2D and 3D lattices (due to its simplicity), there has been research into extending the model to triangular (hexagonal) lattices [26,27,36]. There are quantitative differences between these representations, though qualitatively the structures show little difference in that the model still engenders compact tertiary structures with H-H amino acid residue cores.

Final Comment

The model has been addressed with simple algorithms that seek complexity and solution quality guarantees such as depth-first search [21], and heuristic approximation techniques [1,4,7,14,35,36]. There have been numerous approaches to imbue additional attributes into the model to improve algorithmic performance such as alternative energy functions, representations, amino acid attributes, and structural features. Monte Carlo (Simulated Annealing) search techniques were adopted to provide quick approximate solutions [10,12,13], of note is the PERM approach [15,29,30,34]. Evolutionary algorithms (genetic algorithms) have also been applied at length [2,23,24,28,29,32,33]. More recently, collective intelligence approaches have been used such, specifically the Ant Colony Optimisation (ACO) algorithm [3,5,6].

There are common pseudo-standard benchmark problem instances for both the two-dimensional and three-dimensional model. These are available in the literature, though it is useful to note that there is a repository of problems available on the web with references [37].

The HP model provides an interesting combinatorial problem domain for probabilistic approximation search techniques. Such techniques have already been applied to the domain with success. It is clear that approximation techniques are only a useful application to this problem when the optimal structure cannot be located using conventional deterministic and heuristic approaches. Specifically, suitable application problem instances are those with an increased sequence length for which conventional techniques do not effectively scale to address. Approximation search algorithms do not provide any guarantees as to solution quality or runtime complexity and this is perhaps the general approaches most sighted weakness in the field of computational biology.

This work has defined and described the hydrophobic polar protein-folding model and has demonstrated its application within the field, and its usefulness as a benchmark for combinatorial optimisation algorithms. The problem domain is well defined in the literature, extensively studied and there are standard short length sequences with known optimal solutions available – suitable as benchmark problem instances. It is the hope that this work has promoted interested in the protein-folding area of research and that the use of simple protein-folding models can be more widely adopted as a *demonstration* (benchmark) domain for probabilistic search algorithms, and perhaps further as an *application* problem domain worthy of attention.

References

[1] A. Newman and M. Ruhl, "Combinatorial Problems on Strings with Applications to Protein Folding," *Proceedings LATIN 2004: Theoretical Informatics, 6th Latin American Symposium,* Buenos Aires, Argentina, pp. 369-378, 2004.

[2] A. Patton, W. P. III, and E. Goldman, "A standard ga approach to native protein conformation prediction," *Proceedings of the 6th International Conference of Genetic Algorithms,* pp. 574-581, 1995.

[3] A Shmygelska and H. H. Hoos, "An Improved Ant Colony Optimisation

Algorithm for the 2D HP Protein Folding Problem," *Proceedings of the 16th Canadian Conference on Artificial Intelligence,* Canada, pp. 400-417, 2003.

[4] Alantha Newman, "A new algorithm for protein folding in the HP model," *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms,* San Francisco, California, pp. 876-884, 2002.

[5] Alena Shmygelska and Holger H Hoos, An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem *BMC Bioinformatics*, vol. 6, 2005.

[6] Alena Shmygelska, Rosalía Aguirre Hernández, and Holger H. Hoos, "An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem," *Proceedings of the Third International Workshop on Ant Algorithms,* pp. 40-53, 2002.

[7] Beutler TC and Dill KA., A fast conformational search strategy for finding low energy structures of model proteins *Protein Science*, vol. 5, pp. 2037-2043, 1996.

[8] Bonnie Berger and Tom Leighton, "Protein folding in the hydrophobic-hydrophilic (HP) is NP-complete," *Proceedings of the second annual international conference on Computational molecular biology,* New York, New York, United States, pp. 30-39, 1998.

[9] Brian Hayes, Prototeins *American Scientist*, vol. 86, no. 3, pp. 216-221, May, 1998-Jun 30, 1998.

[10] David G. Covell and Robert L. Jernigan, Conformations of folded proteins in restricted spaces *Biochemistry*, vol. 29, pp. 3287-3294, 1990.

[11] Dylan Chivian, Timothy Robertson, Richard Bonneau, and David Baker, Ab Initio Methods *Methods Biochemistry Annual*, vol. 44, pp. 547-557, 2003.

[12] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus , Protein folding bottlenecks: A lattice Monte Carlo simulation *Physical review letters*, vol. 67, pp. 1665–16681991.

[13] Faming Liang and Wing Hung Wong, Evolutionary Monte Carlo for protein folding simulations *The Journal of Chemical Physics*, vol. 115, pp. 3374-3380, Aug 15, 2001.

[14] Giancarlo Mauri, Giulio Pavesi, and Antonio Piccolboni , "Approximation algorithms for protein folding prediction," *Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms,* Baltimore, Maryland, United States, pp. 945-946, 1999.

[15] Hsu, H.-P., Mehra, V., Nadler, W., and Grassberger, P., Growth Algorithms for Lattice Heteropolymers at Low Temperatures *Journal of Chemical Physics*, vol. 118, pp. 444-451, 2003.

[16] Jeffrey L. Cleland. *Protein Folding - In Vivo and In Vitro*, USA: American Chemical Society, 1993.

[17] 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 Science*, vol. 4, pp. 561-602, 1995.

[18] K. F. Lau and K. A. Dill, "Theory for protein mutability and biogenesis," *Proceedings of the National Academy of Sciences,* pp. 638-642, 1990.

[19] Ken A Dill, Theory for the folding and stability of globular proteins *Biochemistry*, vol. 24, pp. 1501-1509, 1985.

[20] Ken A. Dill , Dominant forces in protein folding *Biochemistry*, vol. 29, pp. 7133-7155, 1990.

[21] Kit Fun Lau and Ken A Dill, A lattice statistical mechanics model of the conformational and sequence spaces of proteins *Macromolecules*, vol. 22, pp. 3986-3997, 1989.

[22] Lila M. Gierasch and Jonathan King. *Protein Folding - Deciphering the Second Half of the Genetic Code*, Washington, USA: American Association for the Advancement of Science, 1990.

[23] Natalio Krasnogor, David Pelta, Pablo E. Martinez Lopez, and Esteban de la Canal, "Genetic Algorithms for the Protein Folding Problem: a Critical View," *Proceedings of Engineering of Intelligent Systems, EIS'98,* pp. 353-360, 1998.

[24] Natalio Krasnogor, William E. Hart, Jim Smith, and David A. Pelta, "Protein Structure Prediction With Evolutionary Algorithms," *Proceedings of the Genetic and Evolutionary Computation Conference,* pp. 1596-1601, 1999.

[25] Pierluigi Crescenzi , Deborah Goldman, Christos Papadimitriou, Antonio Piccolboni, and Mihalis Yannakakis, "On the complexity of protein folding (extended abstract)," *Proceedings of the thirtieth annual ACM symposium on Theory of computing,* Dallas, Texas, United States, pp. 597-603, 1998.

[26] R. Agarwala , S. Batzoglou, V. Dancik, S. E. Decatur, S. Hannenhalli, M. Farach, S. Muthukrishnan, and S. Skiena, "Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model," *Proceedings of the first annual international conference on Computational molecular biology,* Santa Fe, New Mexico, United States, pp. 1-2, 1997.

[27] R. Agarwala , S. Batzoglou, V. Dancik, S. E. Decatur, S. Hannenhalli, M. Farach, and S. Skiena, "Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model," *Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms,* New Orleans, Louisiana, United States, pp. 390-399, 1997.

[28] R. Unger and J. Moult, "A genetic algorithm for three dimensional protein folding simulations," *Proceedings of the 5th International Conference on Genetic Algorithms,* pp. 581-588, 1993.

[29] R. Unger and J. Moult, Genetic Algorithms for protein folding simulations *Journal of Molecular Biology*, vol. 231, pp. 75-81, 1993.

[30] Ramakrishnan, R. and Ramachandran, B., A dynamic monte carlo algorithm for exploration of dense conformational spaces in heteropolymers *Journal of Chemical Physics*, vol. 106, pp. 2418-2425, 1997.

[31] S.E.V. Phillips, Structure and refinement of oxymyoglobin at 1.6 Ĺ resolution *Journal of Molecular Biology*, vol. 142, pp. 5311980.

[32] Thang N. Bui and Gnanasekaran Sundarraj , "An efficient genetic algorithm for predicting protein tertiary structures in the 2D HP model," *Proceedings of the 2005 conference on Genetic and evolutionary computation,* Washington DC, USA, 2005.

[33] Tianzi Jiang, Qinghua Cui, Guihua Shi, and Songde Ma, Protein folding simulations of the hydrophobic–hydrophilic model by combining tabu search with genetic algorithms *Journal of Chemical Physics*, vol. 119, pp. 4592-4596, 2003.

[34] Ugo Bastolla, Helge Frauenkron, Erwin Gerstner, Peter Grassberger, and Walter Nadler, Testing a new Monte Carlo algorithm for protein folding *Proteins: Structure, Function, and Genetics*, vol. 32, pp. 52-66, 1998.

[35] William E. Hart and Sorin Istrail, "Fast protein folding in the hydrophobic-hydrophilic model within three-eights of optimal," *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing,* Las Vegas, Nevada, United States, pp. 157-168, 1995.

[36] William E. Hart and Sorin Istrail, "Lattice and off-lattice side chain models of protein folding (extended abstract): linear time structure prediction better than 86% of optimal (Extended Abstract)," *Proceedings of the first annual international conference on Computational molecular biology ,* Santa Fe, New Mexico, United States, pp. 137-146, 1997.

[37] William Hart and Sorin Istrail. HP Benchmarks. [Online] http://www.cs.sandia.gov/tech_reports/compbio/tortilla-hp-benchmarks.html . 97. 2005.