Combining Machine Learning and Optimization Techniques to Determine 3-D Structures of Polypeptides *

Márcio Dorn, Luciana S. Buriol and Luis C. Lamb
Institute of Informatics, Federal University of Rio Grande do Sul
Porto Alegre, RS, Brazil.
marcio.dorn@acm.org; buriol@inf.ufrgs.br; luislamb@acm.org

Abstract

One of the main research problems in Structural Bioinformatics is the analysis and prediction of three-dimensional structures (3-D) of polypeptides or proteins. The 1990's Genome projects resulted in a large increase in the number of protein sequences. However, the number of identified 3-D protein structures has not followed the same trend. The determination of protein structure is experimentally expensive and time consuming. This makes scientists largely dependent on computational methods that can predict correct 3-D protein structures only from extended and full amino acid sequences. Several computational methodologies and algorithms have been proposed as a solution to the Protein Structure Prediction (PSP) problem. We briefly describe the AI techniques we have been used to tackle this problem.

1 Introduction

Currently, one of the main research problems in Structural Bioinformatics is related to the analysis prediction of three-dimensional structures (3-D) of polypeptides or proteins [8]. The recent Genome projects resulted in a large increase in the number of protein sequences. However, the number of identified 3-D protein structures has not followed the same trend. Currently, the number of protein sequences is far higher than the number of known 3-D structures. If we compare the number of non-redundant sequences of protein sequences ($\sim$10.6 million non-redundant on January 2011) stored in the GenBank [2] with the number of 3-D protein structures with distinct folds (1,198, SCOP 1.5 on January 2011) stored in the Protein Data Bank (PDB) [1] we observe that only $\sim$0.01% of protein sequences are represented in the PDB. Clearly, there is a large gap between the number of protein sequences we can generate and the number of new protein folds we can determine by experimental methods such as X-ray diffraction and NMR. The determination of protein structure is experimentally expensive and time consuming [4]. Therefore, the use of computational techniques that can predict the correct 3-D protein structure from only extended and full amino acid sequences is unavoidable.

Proteins are long sequences of 20 different amino acid residues that in physiological conditions adopt a unique 3-D structure. This structure determines the function of the protein in the cell (structural functions, catalysts in chemical reactions, transport and storage, regulation, and recognition control) [9]. A *peptide* is a molecule composed of two or more amino acid residues chained by a chemical bond called the *peptide bond*. This peptide bond is formed when the carboxyl group of one residue reacts with the amino group of the other residue, thereby releasing a water molecule. Two or more linked amino acid residues are referred to as a peptide, and larger peptides are generally referred to as *polypeptides* or *proteins*. A peptide has three main chain torsion angles, namely ϕ, ψ and ω. In the peptide the bonds between $N - C_\alpha$ (ϕ), and between $C_\alpha - C$ (ψ) are free to rotate. This freedom is mostly responsible for the conformation adopted by the main chain. The rotational freedom around the ϕ (N–C_α) and ψ (C_α-C) angles is limited by steric hindrance between the side chain of the amino acid residue and the peptide backbone [3]. As a consequence, the possible conformation of a given polypeptide is quite limited and depends on the amino acid chemical properties.

2 Computing Techniques for the PSP Problem

Several computational techniques have been proposed as a solution to the PSP problem. They are divided into four classes [8]: (a) First principle methods without database information; (b) First principle methods with database information; (c) Fold recognition methods; and (d) Comparative modeling methods. However, these methodologies have limitations. Group (d) above can only predict structures of protein sequences, which are similar or nearly identical to protein sequences of known structure. Group (c) is limited to the fold library derived from PDB. Group (a) can obtain novel structures with new folds; unfortunately, the complexity and high dimensionality of the search space even for a small protein molecule renders the problem intractable. Predicting the correct 3-D structure of a protein molecule is an intricate and often arduous task. The PSP problem is classified in computational complexity theory as a NP-complete problem [4]. This complexity is due to the folding process of a protein being highly selective. A long amino acid chain ends up in one

*Research supported by the Brazilian Research Council CNPq.

out of a huge number of 3-D conformations. In contrast, the conformational preferences of single amino acid residues is weak. Thus, the high selectivity of protein folding is only possible through the interaction of many residues. Therefore, non-local interactions play an import role in protein three-dimensional structure, as local sequence-structure relationships are not absolute. The prediction of 3-D protein structure can be seen as an optimization problem, where, the goal is to determine the position of each atom in the 3-D space, the bond lengths, the bond angles and the dihedral angles formed between the atoms of the polypeptide.

Currently, we are applying Machine Learning (ML) techniques in order to build computational models to reduce the conformational search space presented in *ab initio* methods that can predict new protein folds. In order to induce polypeptide models we process huge amounts of data obtained from experimental proteins from the PDB. Structural information from protein templates are used in order to build the structure of unknown proteins. Examining structural protein motifs in detail is highly difficult since the mapping from a local sequence of amino acid residues to a local 3-D protein structure is very complex. In this context, we use statistical fragment-based methods to acquire structural information from small protein template samples and use this information in order to train an artificial neural network model and predict approximative 3-D polypeptides structures [7]. The secondary structure of the templates is combined with the information of the torsion angles from the templates obtained from the PDB [6]. It provides a more efficient form to manipulate and obtain structural information from protein templates. Neural networks are used in order to predicting the conformational state of an amino acid residue. This enables new folds to be predicted even when we utilize principles of knowledge-based methods. We are not limited to Protein Data Bank. These structures are built through the use of a sequence-to-structure mapping function. The search space is expected to be greatly reduced and the *ab initio* methods can demand a reduced computational time to achieve a more accurate polypeptide structure. As observed in the experiments the developed method can produce accurate predictions, and the secondary and tertiary structures are close to their experimental structures. These approximate structures can reduce the total time demanded by *ab initio* methods to fold a sequence of unknown structures [5]. The main contributions so far are in: (1) proposing a novel approach for the generation of an approximate 3-D conformation for polypeptides; (2) using secondary structure information combined with ϕ, ψ torsion angles about the central residue in contiguous fragments of a target sequence; (3) using this information, through a neural network, to predict new polypeptides structures.

3 Current Results and Directions

Recently we have developed an heuristic search strategy to refine the approximate structures obtained with ML techniques. We developed a hybrid genetic algorithm (HGA) to optimize the approximated structures. In our proposal, a genetic algorithm is combined with a structured population, and it is hybridized with a path-relinking procedure that helps the algorithm to scape from local minima [10]. The solution structure is a set of n genes, where each gene corresponds to a set of angles of the protein. Each set of angles are comprised of two dihedral angles (ϕ, ψ) from the protein backbone and a number of χ angles that varies according to the type of amino acid residue. The crossover operator is a random key scheme that prioritizes (given 70% of chances) genes originated from individuals selected from an elite set of solutions. The population is structured in “castes”. Initial solutions are generated at random, with angles selected in a special range generated by the neural network strategy that we developed before. The developed method allows efficient mechanisms for protein structure prediction. This is achieved by the use of efficient genetic operators and a path-relinking procedure which helps the HGA to improve the solutions over the generations.

As corroborated by our experiments, the developed method can produce accurate predictions where the 3D protein structures are similar to their experimental structures. At this stage we identify the following contributions: First, the use of AI techniques to develop a new, effective algorithm for the 3D PSP problem, showing that such techniques are useful in an important knowledge domain. Second, the use of genetic algorithms with path-relinking shows that these combined techniques lead to efficient applications. We expect this research opens several interesting research avenues, with a range of applications in bioinformatics.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bath, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, 2000.

[2] D. A. Beson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res.*, 37(Database-Issue):26–31, 2009.

[3] C. Branden and J. Tooze. *Introduction to protein structure*. Garlang Publ. Inc., New York, 2 edition, 1998.

[4] P. Crescenzi, D. Goldman, C.H. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. *J. Comput. Biol.*, 5(3):423–466, 1998.

[5] M. Dorn, A. Breda, and O. Norberto de Souza. A hybrid method for the protein structure prediction problem. *Lecture Notes in Bioinformatics*, 5167:47–56, 2008.

[6] M. Dorn and O. Norberto de Souza. Mining the protein data bank with cref to predict approximate 3-d structures of polypeptides. *International Journal of Data Mining and Bioinformatics*, 4(3):281–299, 2010.

[7] M. Dorn and O. Norberto de Souza. A3n: an artificial neural network n-gram-based method to approximate 3-d polypeptides structure prediction. *Expert Systems with Applications*, 37(12):7497–7508, 2010B.

[8] C.A. Floudas, H.K. Fung, S.R. McAllister, M. Moennigmann, and R. Rajgaria. Advances in protein structure prediction and de novo protein design: A review. *Chem. Eng. Sci.*, 61(3):966–988, 2006.

[9] A. M. Lesk. *Introduction to Bioinformatics*. Oxford University Press Inc., New York, USA, 1 edition, 2002.

[10] M. G. C. Resende, C.C. Ribeiro, F. Glover, and R. Martí. Scatter search and path-relinking: Fundamentals, advances, and applications. In M. Gendreau and J.-Y. Potvin, editors, *Handbook of Metaheuristics*, pages 87–107. Springer, 2010.