

Dynamic Programming Algorithms for Protein Structure Comparison

Manish Kumar¹, Ajay Prakash²

¹Phd Scholar, Shri Venkateshwara University, Uttar Pradesh, India

²Assistant Professor, S.M College Chandausi, Uttar Pradesh, India

Abstract: *Bioinformatics is the computer-assisted data management field that assists in gathering, analysing, and representing biological information. A brief history of bioinformatics is that it arose after computerized protein and DNA sequencing started from the 1970s onward and after computers started to be used as central sources, which took place in the mid- to late 1980s. Bioinformatics widely uses many algorithms in artificial intelligence soft computing and simulation. These algorithms depend on theoretical basics such as discrete mathematics, control and system theory, and statistics. The main objective of bioinformatics is to development new algorithms and statistics, which will be used to evaluate the relationships among members of large data arrangements.*

Keywords: dynamic programming, structural similarity, algorithm, FATCAT, RMSD.

1. Introduction

Proteins are classified based on their nature and properties [1]. Proteins perform a variety of functions, including enzymatic catalysis, transporting ions and molecules from one organ to another, nutrients, contractile system of muscles, antibodies, cartilage, tendons, and regulating the physiological and activities. The functional properties of proteins depend on their three-dimensional structures. The native structure of a protein can be experimentally determined using X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy, etc. Over the past 43 years, the structures of more than 100147 proteins (as of May 2014) have been determined. On the opposite hand, the amino acid sequences are determined for over eight million proteins (as of May 2014). The specific sequence of amino acids in a very peptide chain folds to come up with compact domains with a selected 3D structure of the protein. The peptide chain itself contains all the information necessary to specify its three-dimensional structure [2]. Deciphering the 3D structure of a protein from its amino acid sequence is a long-standing goal in molecular and computational biology.

Proteins fold into 3-D structures, and protein structures are more preserved than protein sequences. Thus, for given a protein structure, it is necessary to go looking for geometrically comparable proteins through protein structure analysis. This is often done in circumstances wherever the likeness at the sequence level is simply to borderline to be detected by any sequence-based similarity search program [3]. Therefore, the object of protein structure comparison is to induce the most important structural similarity between two structures [4]. Protein families are known to retain the shape of the fold even when sequences have diverged below the limit of detection of significant similarities at the sequence level [5].

To find the protein structure comparison, we can use the global measurement of structural comparison method or design a scoring function that reveals the global structural similarity then apply a dynamic programming algorithm to

get the alignment with the maximum score. In this case, we focus on the latter approach, in which dynamic programming algorithms are applied to get the solution.

2. Methods

2.1 Structure Based Sequence Alignment

This method is used to speed up the structural similarity search. Its main goal is to find a better 1D demonstration of 3D structures so that spatial information can be reserved as much as possible. A dynamic algorithm for sequence alignment is used directly to solve this problem [6]. Structure based sequence alignments are potentially more accurate than simple sequence alignments.

To compare protein structure we have to represent the protein structure in 1D representation i.e we have to represent the protein structure as a sequence of characters and this will explain the structural environment of a residue in a protein called 3D environment [7]. Then the protein structure comparison will be primarily reworked as a general sequence alignment drawback, and the standard dynamic programming algorithms for sequence alignment will be used directly for resolution these issues, simply with totally different sets of characters and with a special evaluation perform. This type of structure alignment may well be as quick as sequence alignment. Though they'll not be as correct as those strategies that take into account

Real 3D structural data, however a minimum of will function a fast filter to hurry up the structural similarity search, that is usually rather more time overwhelming. The main development on this direction includes searching out a far better 1D illustration of 3D structures in order that spatial data will be preserved the maximum amount as potential.

2.2 Double Dynamic Programming

Double dynamic programming algorithm is a programming method used for structure comparison. It is applied at two different levels. The first being at a low level to get the top

score that defines the similarity of spatial environment of residues e and f. It is measured by a more complex function that assumes residues e in protein A is correspondent to residue f in protein B [7]. In double dynamic we have to align a given sequence with real coordinates of a structure, taking into account the detailed pairwise interactions, a process which we called threading. In this method we have to match pairwise interactions relates to the requirement of structural comparison methods. Here we have to define the potential terms involving all other residues $f \neq e$. This is a similar definition to that of the structural environment of a residue [8].

2.3 Partial Order Structural Alignment

The algorithm that can perform and visualize multiple alignments of protein structures is the Partial Order Structural Alignment This is the premier algorithm that can perform and visualize multiple alignments of protein structures, concurrently considering their conformational flexibility. It takes into account both the partial order alignment representation and the flexible structure alignment FATCAT POSA is advantageous in circumstances whereby structural flexibilities exist and provides new understandings by visualizing the mosaic nature of multiple structural alignments [9].

Given two protein structures, denote a match of two fragments, one from each protein as an aligned fragment pair. The starting position of an aligned fragment pair q in the two proteins as $j^1(q)$ and $k^2(q)$, and its ending positions in the two proteins as $j^1(q)$ and $k^2(q)$, respectively. Each aligned fragment pair describes one way of superimposing one protein on the other. FATCAT and FlexProt are the two programs which use the formulation of structure alignment as finding a chain of aligned fragment pair and adopt dynamic programming algorithm to find the optimal. Both of these programs allow the structural flexibility in structure comparison. FlexProt first searches for the largest set of congruent aligned fragment pair. FlexProt then looks for a subset of the aligned fragment pairs that describes a possible alignment of two structures with flexibility by clustering consecutive aligned fragment pairs that have a similar 3D transformation. In contrast, FATCAT searches for the best chain of aligned fragment pairs considering the gaps and structural changes between consecutive aligned fragment pairs, each with its own score penalty, therefore, the minimization algorithm compares on the fly solutions involving structural changes and simple extensions and in this way it performs the alignment and structural flexibility detection simultaneously. It would also be important to examine the classes and determine which groups of proteins remain in the same family [10].

3. Results

Randomization was introduced as defined in the methods, on every cycle at the point of residue selection in such that the score values were not altered. This provides better results as opposed to a "best shot" approach based on the residue pairs that score highest. There is no sole measure of the structural superposition of two proteins. Three methods are considered: the best root-mean-square deviation (RMSD)

calculated over all aligned α -carbon atoms (aRMSD), the RMSD over the highest scoring residue pairs (bRMSD), and the weighted RMSD (RMSD). Knowing the maximum score for the Y comparison allows the effects of the number of restatements and the number of residue pair selections on each recapitulation to be investigated to see if they are restrictive factors in accomplishing the maximum score [11].

4. Discussion

The double dynamic programming method provides direction through which parameters can be optimized to achieve its global maximum. The results showed that strong gap penalties cause deposits to be aligned compromising the increased root mean square deviation (RMSD). At first sight, this is convincing, but the likelihood that a weaker penalty might have allowed a lower RMSD to be achieved with the same number of matches could not have passed unnoticed. Although there is a doubt in the suitability of a single RMSD measure, it provides a value that is comparable with other methods. Nevertheless, different methods often result in different numbers of aligned residues making direct comparison hard [7]. The use of stochastic element in double programming method, together with varying gap penalties, certified this problem to be overcome through bringing about the alignment populations with an extensive range of variation in the number of aligned deposits.

5. Conclusion

Dynamic programming is one of the most commonly used algorithms in bioinformatics, and it has been practical to various research topics. Mostly, it is used in the analysis of nonlinear representations of biomolecules as opposed to the comparison of linear sequences. It should be emphasized that although dynamic programming gives an optimal solution, the solution may not be biologically the meaningful. The biological explanation depends not only on the algorithm, but also on how appropriately the construction of the computational problem tells the truth of the biological systems.

References

- [1] Kumar, Manish, Kapil Govil, and Chanchal Chawla. "Comparison Between The Various Protein Classification Schemes." *Journal of Engineering Computers & Applied Sciences* 2.8 (2013): 59-61.
- [2] Anfinsen C.B, *Principles that govern the folding of protein chains*, *Science* (1973); **181** (96):223:230.
- [3] Kumar, Manish, and Govil, Kapil. "Protein Structure Comparison and Classifications into Domains," *International Journal of Science and Research (IJSR)*, Volume 2, Issue 10, 20-22, (2013).
- [4] Reed, D. (2011). *A balanced introduction to computer science*. Boston: Prentice Hall.
- [5] Kumar, Manish, and Govil, Kapil. "The FSSP database: Fold Classification based on Structure--Structure alignment of Proteins" *International Journal of Science and Research (IJSR)*, Volume 2, Issue 10, 23-25, (2013).

- [8] Sperschneider, V. et.al, Sperschneider, J. et.al, & Scheubert, L. et.al (2008). *Bioinformatics: Problem solving paradigms*. Berlin: Springer.
- [9] Ion I. Maandoiu, Zelinkovsky A.. *Bioinformatics Algorithms Techniques and Applications*.
- [10] Liu, L. A. et.al, Wei, D. et.al, Li, Y. et.al, & IGI Global. (2011). *Computational biology and environmental sciences*. Hershey, Pa: IGI Global (701 E. Chocolate Avenue, Hershey, Pennsylvania, 17033, USA.
- [11] Maji, P., & Pal, S. K. (2012). *Rough-fuzzy pattern recognition: Applications in bioinformatics and medical imaging*. Hoboken, N.J: John Wiley & Sons.
- [12] Kumar M, *Proposed Enhanced Proteins Classification Databases, International Journal for Pharmaceutical Research Scholars*, 2013, 2(4), 160-163.
- [13] Sung, W.K, (2010). *Algorithms in Bioinformatics: A practical introduction*. Boca Raton: Chapman & Hall/CRC Press.

Author Profile



Manish Kumar is pursuing PhD in Bioinformatics, from Shri Venkateshwara University, Uttar Pradesh. He has also completed M. Sc (Bioinformatics) and B.Sc (Biosciences) from Jamia Millia Islamia University, New Delhi. He has three years of teaching and research experience. He has been earlier associated with Guru Nanak Dev University, Amritsar, in area of Computer Aided Drug Design and Sequence Analysis. He has published number of research papers in national and international journals. He has also attended number of conferences, workshops and refresher course within India. His areas of interest are Computer Aided Drug Design, Sequence Analysis and Computational & Structural Biology.