www.elsevier.com/locate/yjtbi

Efficient Dynamic Programming Algorithm with Prior Knowledge for Protein β-strand Alignment

Mostafa Sabzekar, Mahmoud Naghibzadeh, Javad Sadri

PII: S0022-5193(17)30018-8
DOI: http://dx.doi.org/10.1016/j.jtbi.2017.01.018
Reference: YJTBI8928

To appear in: *Journal of Theoretical Biology*

Received date: 17 May 2016
Revised date: 11 November 2016
Accepted date: 12 January 2017

Cite this article as: Mostafa Sabzekar, Mahmoud Naghibzadeh and Javad Sadri Efficient Dynamic Programming Algorithm with Prior Knowledge for Protein β strand Alignment, *Journal of Theoretical Biology* http://dx.doi.org/10.1016/j.jtbi.2017.01.018

Efficient Dynamic Programming Algorithm with Prior Knowledge for Protein β-strand Alignment

Mostafa Sabzekar[1], Mahmoud Naghibzadeh[1*], Javad Sadri[2]

[1]Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

[2]Department of Computer Science & Software Engineering, Concordia University, Canada

[*]Corresponding author. Tel.: +985138615100. E-mail address: naghibzadeh@um.ac.ir.

ABSTRACT

One of the main tasks towards the prediction of protein β-sheet structure is to predict the native alignment of β-strands. The alignment of two β-strands defines similar regions that may reflect functional, structural, or evolutionary relationships between them. Therefore, any improvement in β-strands alignment not only reduces the computational search space but also improves β-sheet structure prediction accuracy. To define the alignment scores, previous studies utilized predicted residue-residue contacts (contact maps). However, there are two serious problems using them. First, the precision of contact map prediction techniques, especially for long-range contacts (i.e., β-residues), is still not satisfactory. Second, the residue-residue contact predictors usually utilize general properties of amino acids and disregard the structural features of β-residues. In this paper, we consider β-structure information, which is estimated from protein β-sheet data sets, as alignment scores. However, the predicted contact maps are used as a prior knowledge about residues. They are used for strengthening or weakening the alignment scores in our algorithm. Thus, we can utilize both β-residues and β-structure information in alignment of β-strands. The structure of dynamic programming of the alignment algorithm is changed in order to work with our prior knowledge. Moreover, the Four Russians method is applied to the proposed alignment algorithm in order to reduce the time complexity of the problem. For evaluating the proposed method, we applied it to the state-of-the-art β-sheet structure prediction methods. The experimental results on the BetaSheet916 data set showed significant improvements in the execution time, the accuracy of β-strands' alignment and consequently β-sheet structure prediction accuracy. The results are available at http://conceptsgate.com/BetaSheet.

Keywords: Protein β-sheet, Strand alignment, Contact map, Structural information, Dynamic programming, The Four Russians method

1 INTRODUCTION

Prediction of three dimensional structure of protein from its amino acid sequence, also known as protein structure prediction (PSP), is the focus of interest of various research communities. The prominence of this field can be investigated in two directions. One is its close association with protein functions and another includes its role in narrowing the gap between known protein sequences and structures. However, PSP is one of the most complex and significant open problems in computational biology. One aspect of this complexity arises from the fact that the conformational space of protein structure increases exponentially with the number of amino acids in protein chains. A promising approach to overcome this challenge is to divide the PSP problem into a number of sub-problems such as prediction of contact map, secondary structure elements and their arrangements, and local structures. Although each of them has a certain level of inaccuracy, they can provide vital information regarding the spatial structure of proteins.

An important intermediate step towards the prediction of protein structure is known as β-sheet structure prediction problem. This step is essential not only for reducing the search space of PSP methods [1], [2], but also for elucidating folding pathways [3], [4], designing new proteins [5], [6] and studying its effect on many human diseases ranging from cancer and AIDS to Alzheimer's and mad cow [7]. β-sheets and α-helices are predominant local structures found in proteins. More than 80% of the proteins in the Protein Data Bank (PDB) [8] contain β-sheets [9]. They are formed by pairwise interaction of β-strands which are stabilized by non-local hydrogen bonds running in parallel or antiparallel direction. In this type of interaction, adjacent β-strands bring distant amino acids into close contact with one another.

Accurate prediction of protein β-sheets is regarded as the primary bottleneck in prediction of 3D structure of protein. The main challenge in this phase is the presence of non-local hydrogen bonds in proteins. In contrary to α-helices that are stabilized through local interactions, β-sheets are more complex which are formed by residues at distant positions in the protein chain. The nature of such a long distance interactions are still a challenging problem and leave room for further improvements in β-sheet structure prediction approaches. Prediction of a protein β-sheet, when its secondary structure is known, can be regarded as the prediction of (i) β-strand pairings, (ii) their interaction types (i.e., parallel or antiparallel) and (iii) β-residue interactions (contact map) [10]. Several methods have been proposed to predict each sub-solution and also there are many efforts underway to improve the full identification of β-sheet structure (all three levels). Each of these predictions has a certain level of inaccuracy so it is necessary to propose more and more precise predictors to avoid inaccuracy propagation. In addition, it should be noted that the assumption of knowing the correct secondary structure for protein β-sheet structure prediction (as it is considered by [2], [9], [10]) may be unrealistic. One solution is to use secondary structure predictors but the problem is that they have some inaccuracies and are not always reliable. To overcome this problem, for example, the authors in [11] utilize PSIPRED [12] to predict the secondary structures and then try to reduce the inaccuracies of the predictions.

The first step in prediction of protein β-sheet structure is to predict the native alignment of β-strands. Thus, for any pair of β-strands, the optimum alignment is determined in both parallel and antiparallel directions. The alignment scores, which reflect the tendency of strands for pairing, are then used for the next steps of prediction. The alignment of two β-strands identifies similar regions that may be as a result of functional, structural, or evolutionary relationships between them [13]. Also, correct β-strand alignment would reduce the computational searching space and significantly improve the accuracy of β-sheet prediction algorithms. In almost all previous studies, residue contact statistics are used to define the alignment scores. Predicted contact maps of a