Applied Soft Computing 14 (2014) 357–362

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Predicting helix pair structure from fuzzy contact maps

Tony C.Y. Kuo*, Janice Glasgow

Queen's University, School of Computing, Kingston, Ontario, Canada

ARTICLE INFO

Article history:
Received 30 April 2012
Received in revised form 21 August 2013
Accepted 8 October 2013
Available online 16 October 2013

Keywords:
Helix
Protein structure
Distance constraints
Contact map
Fuzzy contact map

ABSTRACT

One approach to protein structure prediction is to first predict from sequence, a thresholded and binary 2D representation of a protein's topology known as a contact map. The predicted contact map can be used as distance constraints to construct a 3D structure. We focus on the latter half of the process for helix pairs and present an approach that aims to obtain a set of non-binary distance constraints from contacts maps. We extend the definition of "in contact" by incorporating fuzzy logic to construct fuzzy contact maps. Then, template-based retrieval and distance geometry bound smoothing were applied to obtain distance constraints in the form of a distance map. From the distance map, we can calculate the helix pair structure. Our experimental results indicate that distance constraints close to the true distance map could be predicted at various noise levels and the resulting structure was highly correlated to the predicted distance map.

1. Introduction

Proteins are a class of organic molecules that dictate all biochemical reactions in a living cell. The determination of a protein's three-dimensional structure is an important step towards the understanding of protein function. Methods currently exist for experimentally determining structure using X-ray crystallography and Nuclear Magnetic Resonance (NMR) Spectroscopy [1–4]. However, they are complex and time consuming. With the completion of the human genome and the subsequent advances in protein sequencing technology, experimental determination of protein structure cannot keep pace. Thus, there is great interest in methods for predicting protein structure from its sequence.

Methods for predicting protein structure typically fall into one of three categories: homology modeling, fold-recognition, and *ab initio* prediction. In homology modeling, a prediction of the protein structure can be obtained if the target protein sequence has significant homology to another protein of known three-dimensional structure. This approach uses sequence similarity to the target protein to select templates from the set of known protein structures [5]. A sequence alignment between the target protein and the template protein [6,7] is then used to transfer coordinates from the templates, which acts as a foundation for predicting the rest of the protein structure. Typically, many models are generated from many templates which are then assessed using statistical, energy-based, or machine learning methods [8–10].

There has been recognition of the fact that proteins often have similar folds despite no significant sequence similarity [11]. Due to a lack of sequence similarity, recognition of similar folding patterns cannot be detected until the three-dimensional structure has been determined. However, structure classification databases are growing [12,13] and can be useful for predicting proteins that lack significant homologues. Fold-recognition methods make use of structure classifications to find alignments using structure rather than sequence and is thus useful when sequence similarity to the target is low. This approach is based on the regularities of secondary structure arrangement and the topology of polypeptide chains to search for folds that are compatible with a particular sequence by determining how well a fold will fit a sequence [11]. Threading is variation of the fold recognition method [14,15]. Rather than a score based on sequence, this approach builds many rough models and evaluates which model is most likely to be correct based on structural properties [16]. Current fold recognition approaches typically use both threading and sequence alignments.

Protein structure prediction by *ab initio* approaches require a representation of protein geometry as an atomic resolution model of a protein and its solvent environment are computationally expensive. Thus, approximations are used where only one or a few atoms represent each residue and the solvent is implicit [17]. A potential energy function and other parameters based on physical–chemical properties must then be determined [18,19]. The protein conformation space is then searched in order to generate a model, usually by searching an energy surface using stochastic methods such as simulated annealing or Monte Carlo simulation [20–22]. Advances and knowledge gained from *ab initio* approaches are important as they aid all other protein structure prediction approaches in terms of model assessment and optimization.

* Corresponding author. Tel.: +886 978226119.
E-mail addresses: kuo@cs.queensu.ca, tony.cy.kuo@gmail.com (T.C.Y. Kuo), janice@cs.queensu.ca (J. Glasgow).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.10.009

In addition to the above protein structure prediction approaches, there is a growing body of research on the prediction of contact maps and the recovery of the protein structure from contact maps. There exist several methods for predicting contact maps from sequence and they have been shown to be useful in the protein structure prediction [23–25] and protein structure comparison by the Contact Map Overlap metric [26]. The viability of these approaches have been acknowledged by their inclusion in Critical Assessment of Techniques for Protein Structure Prediction (CASP), a biannual blind experiment for assessing protein structure prediction capabilities. The reconstruction of protein structure typically involves using contact maps as distance constraints for a protein's three-dimensional structure [27]. In particular, Walsh et al. [28] experimented with multi-class contact maps and observed an improvement in protein structure reconstruction. This provides motivation to examine how fuzzy contact maps may benefit protein structure prediction. Existing works on fuzzy contact maps focus on its suitability as a representation for protein structure comparison [29,30]. These works constructed fuzzy contact maps from observed distance maps and performed protein structure classification to show structure comparison was improved by the fuzzy contact map representation.

In this study, we present an approach to obtain non-binary distance constraints from fuzzy contact maps to be used in calculating protein structure. Our focus is on alpha-helices, one of the regular local configurations of a protein's backbone known as protein secondary structure. Specifically, we are concerned with the structure of helix pairs. To the best of our knowledge, this is the first study to reconstruct from fuzzy contact maps.

Our approach incorporates fuzzy logic to construct fuzzy contact maps from a set of binary contact maps at various thresholds and extends the binary term "in contact" to a fuzzy membership function. This representation provides additional detail as well as robustness. The fuzzy contact map construction methodology in this study is aimed at being applicable to predicted contact maps from sequence. Furthermore, we aim to leverage the knowledge contained in fuzzy contact maps to predict the three-dimensional structure.

We implement a process that retrieves similar distance map regions to a query from a database of experimentally solved structures using fuzzy contact map similarity. This retrieval process uses an exhaustive search for the best pair-wise alignment between two fuzzy contact maps and returns a set of similar fuzzy contact map regions that corresponds to distance map regions similar to a query. From the retrieved distance maps regions, we adapt a distance map using distance geometry bound smoothing [31]. The three-dimensional coordinates of the helix pair are then calculated with the EMBED algorithm [32].

The organization of this article is as follows. In Section 2 we introduce the method to construct fuzzy contact maps from a set of binary contact maps at various distance thresholds. In Section 3 we make use of fuzzy contact maps and present a retrieval method to obtain a set of the distance map regions similar to a query based on fuzzy contact map similarity. Section 4 describes the adaptation of the retrieved distance maps regions into a distance map through the use of bound smoothing. We present the experimental setup and results in Section 5. Finally, in Section 6 we discuss of our results.

2. Fuzzy contact maps

The distance map of a protein can capture to a large extent, its three-dimensional structure. A distance map, D, for a protein of n residues is a $n \times n$ symmetrical matrix where $d_{i,j}$ represents the distance between residue i and residue j. The contact map is a binary version of the distance map whose values are 1 or 0 for a contact and non-contact respectively. A contact is said to exist when the distance between two residues is at or below a given distance threshold and can be defined by distances between C_α atoms, C_β atoms of amino acid residues, or atoms on the side chains.

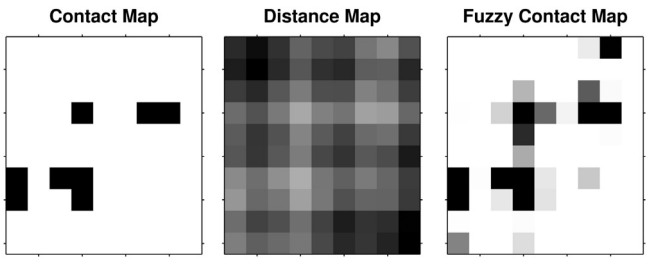

Fig. 1. An image analogy example of a helix-pair distance map and its corresponding contact map and fuzzy contact map at 8 Å.

Fuzzy contact maps are an extension of standard contact maps that use a fuzzy membership function rather than a singular value to define the threshold. This results in a degree of truth for two residues being "in contact". Typically, distance maps are converted into fuzzy contact maps by mapping the distance value to a degree of truth in the range [0, 1] via a membership function such that

$$F_D = f(d_{ij}), \quad 1 \le i \le n,\ 1 \le j \le m; \tag{1}$$

where F_D is the fuzzy contact map constructed from the distance map, f is a function describing the shape of the membership function, and d_{ij} is the distance between residues i and j.

Using an image analogy (Fig. 1) where the distance map is a gray-scale image, the contact map would be a thresholded image while the fuzzy contact map would be a bounded normalized image. It seems obvious that the fuzzy contact map would contain more detail and be more representative of the distance map.

Our aim is to construct fuzzy contact maps from contact maps so that it can be applicable to predicted contact maps. To that end, a set of contact maps across a range of threshold values can be considered a slice of a distance map from which a fuzzy contact map could be constructed. Since contact maps can presumably be predicted at any arbitrary threshold, this would be extensible to predicted contact maps.

A set of contact maps will encompass a small range of threshold distances and they may contain errors. Thus, the fuzzy contact map conveniently provides the following:

- A representation that is compact for a set of contact maps.
- A representation that allows for a measure of similarity to be applied between distance maps and a set of contact maps.
- A representation that allows for the handling of errors in the set of contact maps.

The method to construct a fuzzy contact map from a set of contact maps is as follows. In the general case, the number of maps in the set of contact maps, n, is equal to the number of distance thresholds in set T such that

$$T = \{t_1, t_2, \ldots, t_n\}, \tag{2}$$

where $t_i < t_{i+1}$ for $1 \le i < n$. The set of contact maps, C, is then

$$C = \{c_1, c_2, \ldots, c_n\}, \tag{3}$$

where c_i is a contact map with a corresponding threshold of t_i. The set of distance values which represent each contact map in the set is

$$D = \{d_1, d_2, d_3, \ldots, d_n\}, \tag{4}$$

where $d_1 = t_1$ and $d_i = (t_i + t_{i-1})/2$, $1 < i \le n$.