Ordering protein contact matrices

Chuan Xu, Guillaume Bouvier, Benjamin Bardiaux, Michael Nilges, Thérèse Malliavin, Abdel Lisser

PII: S2001-0370(17)30105-8
DOI: doi:10.1016/j.csbj.2018.03.001
Reference: CSBJ 216

To appear in: *Computational and Structural Biotechnology Journal*

Received date: 1 October 2017
Revised date: 28 February 2018
Accepted date: 1 March 2018

Please cite this article as: Xu Chuan, Bouvier Guillaume, Bardiaux Benjamin, Nilges Michael, Malliavin Thérèse, Lisser Abdel, Ordering protein contact matrices, *Computational and Structural Biotechnology Journal* (2018), doi:10.1016/j.csbj.2018.03.001

Ordering protein contact matrices

Chuan Xu[a], Guillaume Bouvier[b,c], Benjamin Bardiaux[b,c], Michael Nilges[b,c], Thérèse Malliavin[b,c], Abdel Lisser[a]

[a] *Laboratoire de Recherche en Informatique, Université Paris-Sud and CNRS UMR8623*
[b] *Unité de Bioinformatique Structurale, Institut Pasteur and CNRS UMR3528*
[c] *Centre de Bioinformatique, Biostatistique et Biologie Intégrative, Institut Pasteur and CNRS USR3756*

Abstract

Numerous biophysical approaches provide information about residues spatial proximity in proteins. However, correct assignment of the protein fold from this proximity information is not straightforward if the spatially close protein residues are not assigned to residues in the primary sequence. Here, we propose an algorithm to assign such residue numbers by ordering the columns and lines of the raw protein contact matrix directly obtained from proximity information between unassigned amino acids. The ordering problem is formatted as the search of a trail within a graph connecting protein residues through the nonzero contact values. The algorithm performs in two steps: (i) finding the longest trail of the graph using an original dynamic programming algorithm, (ii) clustering the individual ordered matrices using a self-organizing map (SOM) approach. The combination of the dynamic programming and self-organizing map approaches constitutes a quite innovative point of the present work. The algorithm was validated on a set of about 900 proteins, representative of the sizes and proportions of secondary structures observed in the Protein Data Bank. The algorithm was revealed to be efficient for noise levels up to 40%, obtaining average gaps of about 20% at maximum between ordered and initial matrices. The proposed approach paves the ways toward a method of fold prediction from noisy proximity information, as TM scores larger than 0.5 have been obtained for ten randomly chosen proteins, in the case of a noise level of 10%. The methods has been also validated on two experimental cases, on which it performed satisfactorily.

Keywords:
protein contact matrix, fold prediction, graph theory, dynamic programming, self-organizing map