Selecting protein fuzzy contact maps through information and structure measures

Carlos Bousoño-Calzón
Signal Processing and Communication Dpt.
Univ. Carlos III de Madrid
Avda. de la Universidad, 30
28911 Leganés (Madrid) Spain
cbousono@tsc.uc3m.es

Alicia Oropesa-García
Signal Processing and Communication Dpt.
Univ. Carlos III de Madrid
Avda. de la Universidad, 30
28911 Leganés (Madrid) Spain
aoropesa@tsc.uc3m.es

Natalio Krasnogor
Automated Scheduling, Optimisation and Planning Research Group
Univ. of Notingham
NG8 1BB, UK
nxk@cs.cs.ac.uk

Abstract

Protein contact maps are representations of the proteins three dimensional folding topology. A fuzzy generalization of contact maps (FGCM) provides to the researcher flexibility not present in standard (i.e. crisp) contact maps but it also changes the information content of the data. To aid in the rationale –rather than ad-hoc- selection of generalized contact maps parameters we introduce some universal measures of information. We discuss this in the paper and show its impact on FGCM.

Keywords: Protein contact maps, fuzzy membership functions, entropy, symmetry, complexity.

1 Introduction

"When we look at a cell through a microscope or analyse its electrical or biochemical activity, we are, in essence, observing proteins". Protein's biological functions depend essentially on their three dimensional (3D) shape: this determines the molecules they are to bind with; this constrains the movements allowed for a protein to act as a mechanical machine [1], etc.

Whereas it is possible to predict a protein secondary structure starting with the amino acid sequence, in general, it is presently not possible to reliably predict *ab-initio* the 3D structure (i.e. its native state). However, some proteins are amenable to crystallographic analysis; X-ray diffraction (with a wavelength of 0.1 nm.) patterns are measured and by suitable interpretation of the raw data it is possible to determine the position of all the non-hydrogen atoms in the native state of the molecule reliably. Nuclear Magnetic Resonance (NMR), a complementary technique, is used to determine 3D structures for small proteins (specially if they resist crystallization or it is necessary to monitor changes in the conformation). As both X-ray crystallography and NMR are empirical techniques they introduce experimental errors in the measurements which sometimes are amplified by the pattern interpretation technique used to decode the atomic coordinates.

Bioinformatics techniques must therefore cope with empirical errors in the 3D determination and with efficient comparison between different (protein) 3D data. In a previous work [2], a fuzzy generalisation for the representation of the 3D protein structure was proposed to cope with some of these errors. The introduction of fuzzy membership functions also gives the researcher flexibility to define "contacts" in the 3D protein for different biological features (for example, selecting α-helices or β-sheets by distance properties).

Departing from the subjective tuning of these "fuzzy" degrees of freedom, in this paper we address a preliminary discussion of the information changes induced in the protein data through "universal" measures (such as entropy or symmetry, as introduced in [3]) in order to select, or constrain, the use of membership functions.

2 Protein contact map and the fuzzy generalisation

We consider the basic 3D protein structure to be given by a distance matrix. For a particular protein composed of a sequence of N amino acids, the distance matrix is a symmetric matrix of dimension N x N. Rows and columns represent this amino acid sequence, and each matrix entry stands for the Euclidean distance in the 3D space between the amino acids referred by the row and column. This explains the matrix symmetry as well as a zero diagonal. An example of this matrix, which corresponds to alpha-beta 1AA9 protein with 172 amino acid sequence in the PDB data set (see [2] for further reference), can be observed in Fig. 1, where distance is mapped to image intensity.

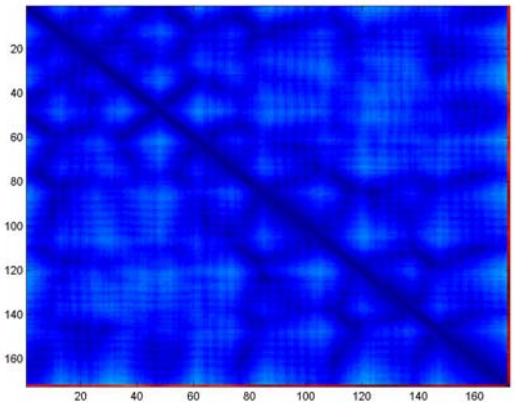

Fig. 1: Protein distance matrix example

The standard (crisp) contact map is a matrix with 0-1 entries computed out of the distance matrix by applying the function in Fig.2:

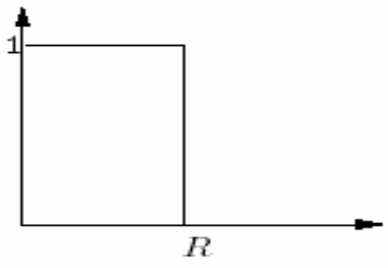

Fig.2: Crisp function for the contact map

The fuzzy generalisation, as discussed in this paper (see [2] for the complete generalisation), can be stated as the change of the crisp function in Fig. 2 by the membership function of Fig. 3:

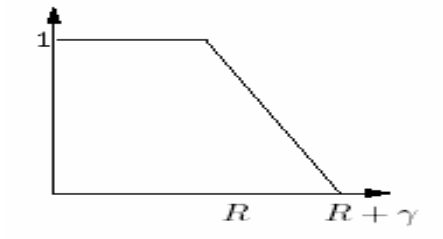

Fig.3: Membership function for the fuzzy contact map

Note that this membership function depends of two parameters (R,γ) instead of only the crisp threshold R.

The contact map is an essential tool in formulating the protein overlap problem to compare 3D structures[10] thus the generalised version may help to increase the symmetry of this formulation as discussed later. Symmetry can help to simplify the complexity of the protein overlap problem or design algorithms for this purpose, but this issue is out of the scope of this preliminary discussion (see for example [11]).

3 Information and structure measures for graphs

3.1 Entropy measure

Information Theory (IT) has been applied to measure information (whose actual meaning depends on the particular use of these concepts) conveyed by biological sequences such as proteins [4]. Entropy is the most important measure from IT: Given a discrete random variable (r.v.) by its probability distribution $\{p_1,\ldots,p_n\}$, its entropy is defined as

$$e = \sum_{k=1}^{n} p_k \log(\frac{1}{p_k}) \qquad (1)$$

However, to directly apply this measure to the distance matrix in Fig.1 by considering each of its entries as r.v. realisations does not provide any clue about the 3D protein structure. Fig. 4 shows an image with similar entropy generated by a gaussian r.v. with the same mean and variance of Fig.1:

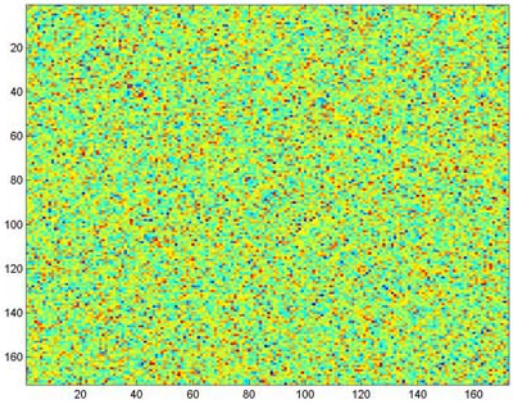

Fig. 4: Random distance matrix

In order to introduce structure in the entropy formalism, the usual approach is to apply Markov models. However, this approach is rather complex, specially if the only objective is to measure large structures as those in proteins. We alternatively use a measure of symmetry, firstly introduced in [3], as a computational efficient measure for structure. The rationale behind this proposal is the increasingly important role that symmetry concepts are playing in understanding structure [5] and complex dynamics [6].

3.2 Symmetry measure

As for the symmetry in a distance matrix or relationship graph, it is known that if a graph is highly symmetric its eigenvalues are highly degenerated [7]. This is not proved to be a *if and only if relation* so that we define a similarity measure heuristic which accounts for inverse of the pair-wise difference of the eigenvalues of a graph. This is formalized as follows:

Let λ be the vector of eigenvalues (ordered from the largest to the shortest, and normalised) of the distance matrix, the contact map or the fuzzy contact map, G (n x n dimension). Note that all these matrices have real eigenvalues since they coincide with their transposes. Let $\delta\lambda$ be the vector of the differences between each element of λ with its successor, and $\delta\lambda_k$ its k component. The measure of symmetry is defined as:

$$\mu(G) = \sum_{k=0}^{n-1} \exp\left(-2\,\delta\lambda_k\right) \qquad (2)$$

Since this definition is heuristic and its validity does not rest upon any theoretical development, we provide its values for Figs. 1 and 4 which are 131.36 and 7.67, respectively. Additional evidence of its usefulness has been provided in [3]. Although the entropy for these two images is essentially similar, their symmetry measures are two orders of magnitude different. Further check can be brought from the contact map as calculated from data in Fig. 1 by applying the crisp function (Fig.2) with R=10 (Angstroms) and shown in Fig. 5. The following table illustrate the discriminative power of these measures on these figures.

Table 1.: Measures comparison

Fig.	Ent. Measure (bits)	Sym. measure	Kolm. Compl. (Kbits)
1	2.64	131.36	77
4	3.89	7.67	277
5	0.52	141.84	43

Table 1 relates the entropy and symmetry measures to images attaching a subjective meaning to them as raw information and structure degree respectively. Note that Fig. 5 essentially preserves the same structure as Fig. 1, however the quantisation of the matrix entries reduces the whole information content.

3.3 Kolmogorov Complexity measure

Complementarity of (first order) entropy and this symmetry measure, as discussed above, is reminiscent of Kolmogorov complexity (see [8] for a formal definition). The Kolmogorov complexity of an object gives the length of the shortest program that can generate it and is a quantity closely related to the degree of compression which can be achieved for a sequence encoding the original object. To achieve such a compression an explanatory algorithm is needed for the encoding sequence. The size of such an algorithm (which is essentially invariant to any specific encoding) comprises the Kolmogorov complexity of, in our application domain, the 3D structure encoded by (generalised) contact maps.

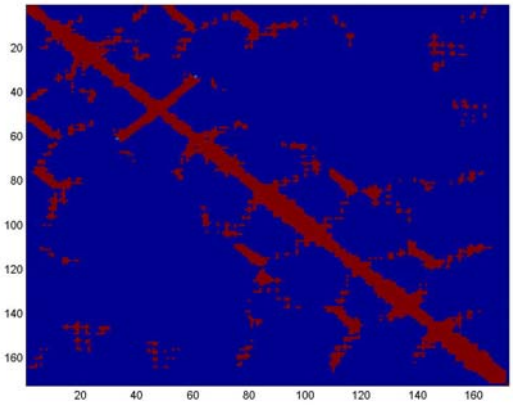

Fig.5 Crisp contact map for 1AA9

An easy approximation to the Kolmogorov complexity can be borrowed from available compression algorithms [9]. In our case, the images are *winzipped* JPEG (essentially suppressing JPEG headers), so to comply with [9]. The fourth column in Table 1 shows the consistency of the entropy and symmetry measures: as the entropy decreases and/or the symmetry increases, data can be more compressed.

4 Selecting membership function parameters: some experiments

The use of fuzzy membership functions in the definition of protein (generalised) contact maps not only allows a researcher to interpret protein data in different ways but also changes the information content in the data. The use of the entropy and symmetry measures introduced in previous sections allow us to control these issues in order to select membership functions or their parameters with a view of maximum information gain. In what follows, we discuss how the parameters (R,γ) of the membership function in Fig. 3 affect these measures as applied to the distance function in Fig. 1. We do not intend to provide an optimisation framework in order to set these parameters since different criteria may be equally feasible, but simply to provide an experimental illustration.

In a previous work, [2], we had set (R,γ)=(10, 1.2) for the same data and membership function. Figs. 6 and 7 gives the entropy and symmetry measures for an exploring range of parameters γ and R around (R,γ)=(10, 1.2), respectively.

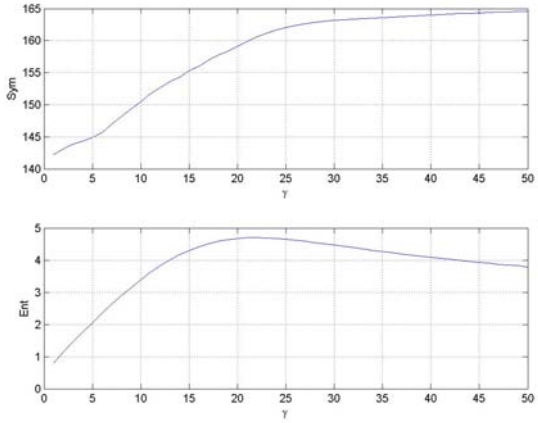

Fig. 6: Entropy and symmetry for R=10.

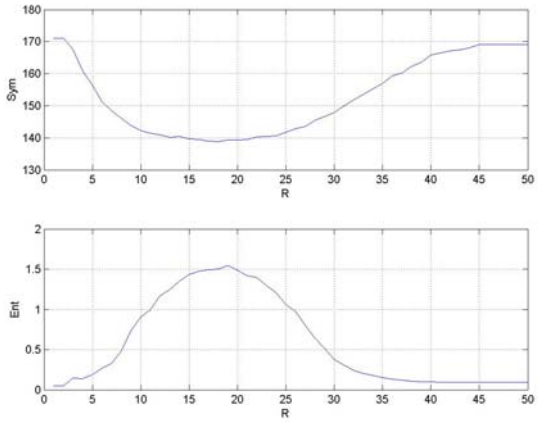

Fig. 7: Entropy and symmetry for γ=1.2.

Despite local irregularities, there are consistent extrema in these plots with very different behaviours for the two parameters. In order to appraise these results, Figs. 8, 9 and 10 give the fuzzy contact maps for (10,20), (10,50) and (18,1.2), respectively. Fig. 11 provides the reference fuzzy contact map, (10, 1.2).

The inspection of these figures illustrates the effect of modifying entropy and symmetry through the fuzzy contact maps. For instance, the difference between Figs 8. and 9 is explained by gaining in symmetry at a cost in entropy; so Fig. 9 has less information than Fig. 8 but the interpretation of the 3D shape of the protein has more regularities. Of course, the final selection of parameters must be driven by the researcher's objectives.

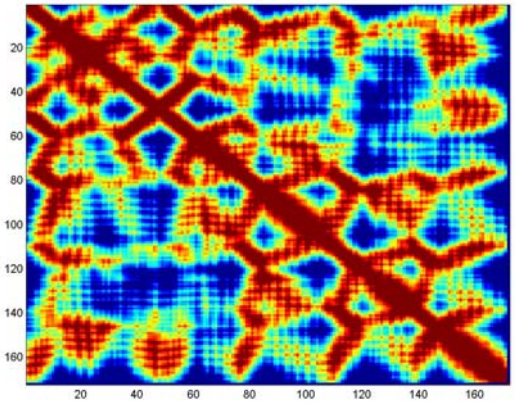

Fig. 8: Fuzzy contact map (R,γ) = (10,20)

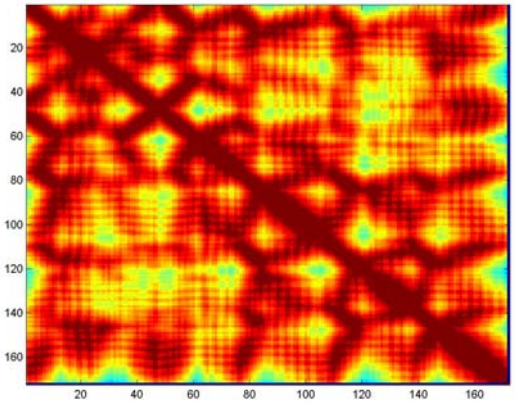

Fig. 9: Fuzzy contact map (R,γ) = (10,50)

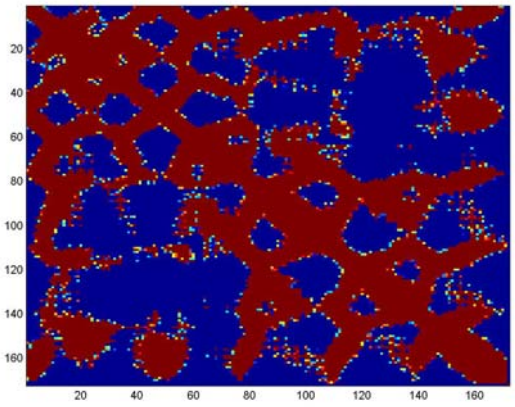

Fig. 10: Fuzzy contact map (R,γ) = (18,1.2)

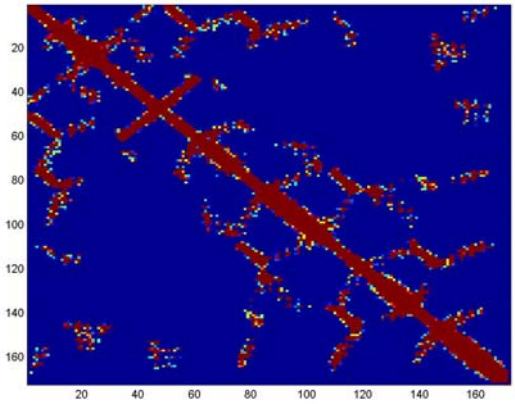

Fig. 11: Fuzzy contact map (R,γ) = (10,1.2)

5 Conclusions

A protein's 3D structure is a central building block in the life sciences thus it becomes necessary to have flexible and efficient tools for its handling and processing. Fuzzy membership functions provides interesting degrees of freedom which also changes the informational content of the data.

We have introduced different measures of information which address entropy and structure in the data, and discuss their consistency. Images of the real protein data help to interpret the significance of such a measures in selecting membership parameters.

References

[1] B. Alberts et al., *Molecular biology of the cell*, 4th ed., NY: Garland, 2002.

[2] D. Pelta, N. Krasnogor, C. Bousoño-Calzón, J. L. Verdegay, E. Burke, "A Fuzzy generalisation of contact maps for the overlap of protein structures," *Journal of Fuzzy Sets and Systems*, in press.

[3] A. Oropesa-García and C. Bousoño-Calzón, "A Structural symmetry measure for complex systems: a neural network case," *Proc. CSIMTA'04*, Cherbourg (France), 19-22 September 2004, pp: 119-122.

[4] H. P. Yockey, *Information theory and molecular biology*, NY: Cambridge Univ.Press, 1992.

[5] D. Goodsell, A. Olson, "Structural symmetry and protein function", *Annal Review of Biophysics and Biomolecular Structure*, vol. 29, pp: 105-153, June 2000.

[6] M. Golubitsky and I. Stewart, *The symmetry perspective*, Berlin: Birkhäuser, 2002.

[7] N. Biggs, *Algebraic graph theory*, UK: Cambridge, 1996.

[8] T. Cover, J. Thomas, *Elements of information theory*, NY: Wiley, 1991.

[9] N. Krasnogor and D.A. Pelta, "Measuring the Similarity of Protein Structures by Means of the Universal Similarity Metric". *Bioinformatics* 20 (7), 2004

[10] R.D. Carr, W.E. Hart, N.Krasnogor, J.D. Hirst, E.K. Burke, "Alignment of Protein Structures with a Memetic Evolutionary Algorithm," *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2002, New York, USA 200., 2002*

[11] J. Navaza et al., "On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation", *Biological Crystallography, Acta Cryst. D58*, pp. 1820±1825, 2002.