Protein Structure Comparison through Fuzzy Contact Maps and the Universal Similarity Metric

David Pelta **Juan Ramón Gonzalez**
Dept. of Computer Science and AI
University of Granada, 18071
Granada,Spain
{dpelta,jrgonzalez}@decsai.ugr.es

Natalio Krasnogor
ASAP Group
University of Nottingham,
Nottingham, NG8 1BB, U.K.
Natalio.Krasnogor@Nottingham.ac.uk

Abstract

Comparing protein structures, either to infer biological functionality or to assess protein structure predictions is an essential component of proteomic research. In this paper we extend our previous work on the use of the Universal Similarity Metric(USM) and Generalized Fuzzy Contact maps. More specifically we compare the impact that generalized fuzzy contact maps representations have on the assessment of protein similarity by means of the Universal Similarity Metric.

Keywords: fuzzy contact maps, protein comparison, universal similarity metric

1 Introduction

The comparison of protein structures is an important problem in bioinformatics [2]. Structure comparison are usefull in a variety of situations like inferring biological functionality of a new structure or assessing the quality of tertiary structure predictors. These, and other roles of protein structural comparison, makes this an important problem in drug development.

Protein structures can be compared in a number of ways, but ultimately, a suitable "encoding" of the three dimensional information is required. In this paper we choose to encode the topological information of protein structures by means of "contact maps". In [5] we used **crisp** contact maps to compute the universal similarity metric (USM), while in [10] we generalized the concept of crisp contact maps to fuzzy contact maps in various ways. In this paper we study the interplay between the clustering capability of the USM and the generalization features of fuzzy contact maps.

The paper is organized as follows: in Section 2, we describe the very basics of protein structure. Section 3 is devoted to the presentation of standard (crisp) and fuzzy contact maps and then in Section 4 we review the main ideas of the universal similarity metric. The experiments and results obtained are described in Section 5. Finally, Section 6 is devoted to the conclusions.

2 Required Notions

A protein is a linear arrangement of amino acids, that is, a polymer. Each amino acid is a multiatom compound. Usually, only the "residue" part of these amino acids are considered when studying protein structures for comparison purposes. Thus a protein's *primary sequence* is usually thought-of as composed of "residues". The primary sequence adopts local motifs called *secondary structure*. The most relevant secondary structure features are helices, sheets and loops. The *native state* or *tertiary structure* of a protein is the three dimensional shape the polymer adopts under certain physiological conditions. That is, the local motifs of the secondary structure agregate further into a higher organization level. In its native state a protein performs its biological function. In some cases, a protein structure may be composed by a set of three dimensional chains structures. Figure 1 graphically shows the previous description[1].

[1] Taken from http://www.accessexcellence.org/RC/VL/GG/

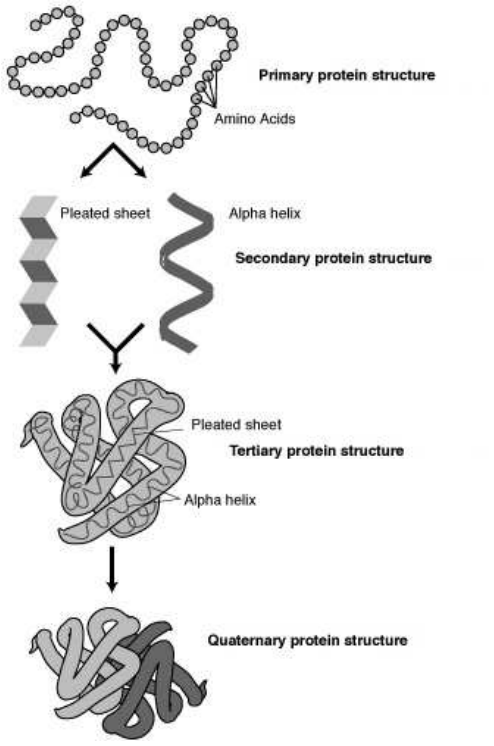

Figure 1: Main protein structures

3 Standard and Fuzzy Contact Map

In this section, we define contact maps and we describe its extension to fuzzy contact maps.

3.1 Standard Contact Map

A contact map [8, 9] is a concise representation of a protein's 3D structure. Formally, a map is specified by a 0-1 matrix S, with entries indexed by pairs of protein residues, such that

$$S_{i,j} = \begin{cases} 1 & \text{if residue } i \text{ and } j \text{ are in contact} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Residues i and j are said to be in "contact" if their Euclidean distance is at most $\Re$ (a threshold measured in Angstroms) in the protein's native fold. Thus, the contact map can be also seen as a white/black image.

In this work, contact maps are calculated by taking into account the distance of the C_α atoms of the residues under consideration. The contact map captures the 3D structure of proteins and certain structural features are conspicuous when the contact map is represented graphically.

3.2 Generalized Fuzzy Contact Maps

In the previous model, a crisp Euclidean distance threshold is used to decide whether two residues are in contact or not. In order to produce a more flexible framework for protein similarity we resort to a richer concept of contact and contact maps that was previously addressed in [10].

We define a *fuzzy contact* as that made by two residues that are *approximately*, rather than exactly, at a distance $\Re$. Formally, a fuzzy contact is defined by:

$$F_{i,j} = \mu(\overline{[i,j]}, \Re) \quad (2)$$

where $\mu()$ is a particular definition of (fuzzy) contact, $\overline{[i,j]}$ stands for the Euclidean distance between residues i, j, and $\Re$ is the threshold as for the crisp contacts. The standard, i.e. crisp, contact map is just a special case of the fuzzy contact map when a user-defined α-cut is specified.

Figure 2 (a),(b) and (c) shows three alternative meanings for "contact" and the corresponding membership functions. Each panel in the figure is a fuzzy contact map in which a dot appears for each pair of residues such that $F_{i,j} > 0$ (i.e. the support of the corresponding fuzzy set).

Fuzzy contact maps are further generalized by removing the constraint (in the original model) of having only one threshold $\Re$ as a reference distance. The formal definition of a General Fuzzy Contact is given by:

$$F_{i,j} = max\{\mu_1(\overline{[i,j]}, \Re_1), \ldots, \mu_m(\overline{[i,j]}, \Re_n)\} \quad (3)$$

with the contact map C defined as:

$$C^{r \times r} = (F_{i,j}) \; with \; 0 \leq i, j \leq r \quad (4)$$

That is, up to n different thresholds and up to m different semantic interpretations of "contact" are used to define the $r \times r$ contact map where r is the number of residues in the protein.

As an example, consider the 2-thresholds, 2-membership functions fuzzy contact map, shown in Fig. 2 (d), that simultaneously highlight *short* and *long* structural patterns. The membership

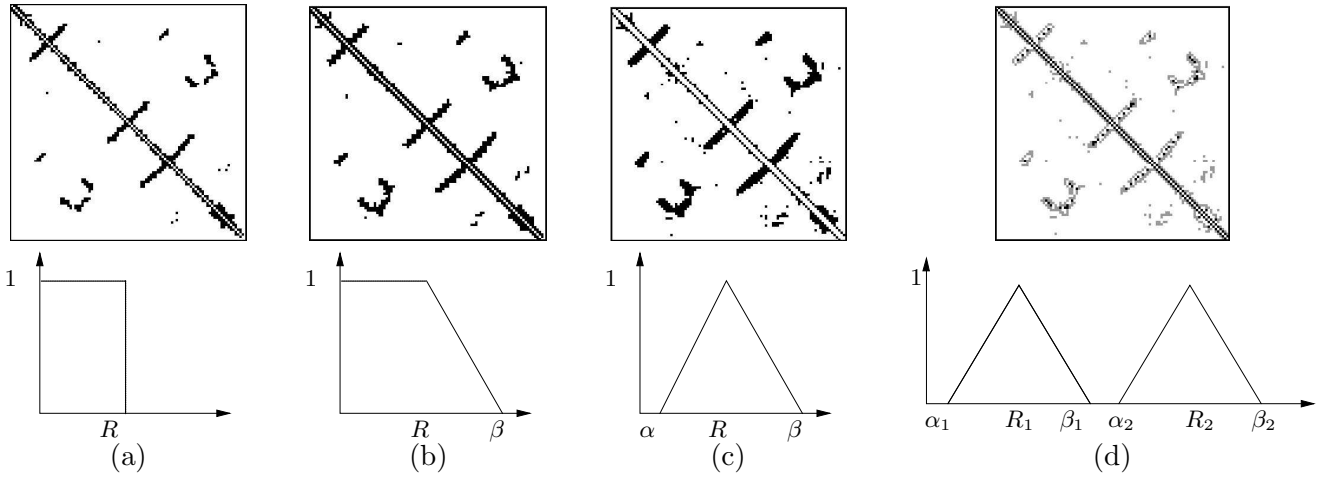

Figure 2: Four examples of contact maps. In (a) the standard model; b) the simplest fuzzy generalization; c) another generalization; d) a 2-threshold, 2 membership functions fuzzy contact map.

functions μ_1, μ_2 for short and long patterns are defined in such a way that they do not overlap and with $\Re_1 < \Re_2$. A simple visual inspection shows that the resulting patterns are different among the maps in the figure.

From an implementation point of view, the upper triangular part of the contact map stores the values $F_{i,j} \in [0,1]$, while the lower triangular part stores the index of the membership function where the maximum for Eq. 3 was achieved. In this example, the "type" of the contact is either 1 (for short) or 2 (for long). When the maximum is reached in two fuzzy sets, then the type assigned corresponds to the leftmost one.

4 The Universal Similarity Metric

The similarity of two given fuzzy contact maps is measured by means of the Universal Similarity Metric (USM). This measure, introduced in [6] and first used in protein structure comparison in [5] approximates every possible similarity metric. At the heart of the USM lies the concept of Kolmogorov Complexity $K(.)$ of an object o, defined as the length of the shortest program for a Universal Turing Machine U that is needed to output o. Following [7] we have:

$$K(o) = \min\{|P|,\ P \textit{ a program and } U(P) = o\} \quad (5)$$

A related measure is the conditional Kolmogorov complexity of o_1 given o_2:

$$K(o_1|o_2) = \min\{|P|,\ P \textit{ a program and } U(P, o_2) = o_1\} \quad (6)$$

Equation 6 measures how much information is needed to produce object 1 if we knew object 2. Then, the Information Distance between two objects, accordingly to [1], is equivalent (up to a logarithmic additive term) to :

$$ID(o_1, o_2) \quad = \quad \max\{K(o_1|o_2), K(o_2|o_1)\} \quad (7)$$

The USM (as appears in [6]) is a proper metric, universal and also normalized. It is defined as:

$$d(o_1, o_2) \quad = \quad \frac{\max\{K(o_1|o_2^*), K(o_2|o_1^*)\}}{\max\{K(o_1), K(o_2)\}} \quad (8)$$

where o_i^* indicates a shortest program for o_i.

In our previous work [5], and following [7], each contact map was represented as a string s and $K(s)$ was approximated by the size (i.e. number of bytes) of the compressed string $zip(s)$, that is, $K(s) \approx |zip(s)|$.

The term $K(o_1|o_2)$ was calculated as $K(o_1 \cdot o_2) - K(o_2)$ where $\cdot$ denotes string concatenation and $K(.)$ was estimated as mentioned above.

In [5], we use the approximation proposed to compute the similarities of standard protein' contact maps. In the next section, we compute the USM

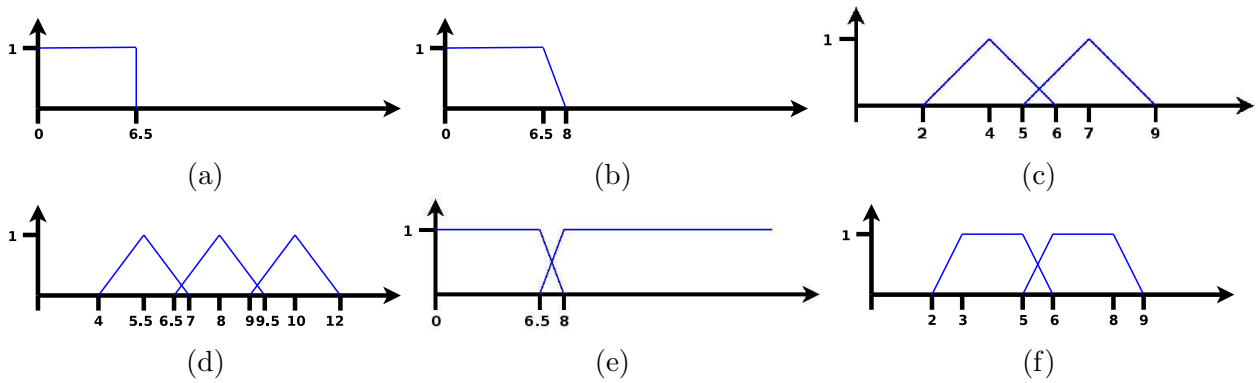

Figure 3: Definitions used to create the fuzzy contact maps.

based on generalized fuzzy contact maps to assess the impact of the encoding in the resulting clusters induced by USM.

5 Experiments and Results

In order to test the benefits of using fuzzy contact maps and the universal similarity metric for protein structure comparison, we use the Chew-Kedem dataset [3] composed by 32 medium size proteins grouped in 5 different families: globins (1eca, 5mbn, 1hlb, 1hlm, 1babA, 1ithA, 1mba, 2hbg, 2lhb, 3sdhA, 1ash, 1flp, 1myt, 1lh2, 2vhb), alpha-beta (1aa9, 1gnp, 6q21, 1ct9, 1qra, 5p21), tim-barrels (6xia, 2mnr, 1chr, 4enl), all beta (1cd8, 1ci5, 1qa9, 1cdb, 1neu, 1qfo) and alpha (1cnp,1jhg). These are extracted from the Protein Data Bank [4].

We constructed 6 different definitions for the fuzzy contact maps (Fig. 3) to check how much information a contact map should have to provide good results. Then, we followed these steps:

1.- From each protein file in the dataset, extract the first chain.
2.- For each chain and membership function definition, produce a fuzzy contact map.
3.- For each pair of protein contact maps c_1, c_2, compute $d(c_1, c_2)$ using Eq. 8 to obtain the similarity between them. The similarities are stored in a matrix.
4.- Apply single linkage clustering over such matrix.

The results obtained using this protocol, are shown as dendrograms in Fig. 4. Each tree corresponds to a particular similarity matrix obtained using a particular definition of contact map. The character between brackets, corresponds to the definitions shown in Fig. 4, so the dendrogram (a) corresponds to the clustering obtained using crisp contact maps. To simplify the analysis, we add a symbol at the right of a protein name, denoting the class where the protein belongs.

Interestingly, some common features arise in every clustering. First, it seems to be easy to perform a good clustering of some classes, namely, globins (denoted with $\sim$), alpha ($+$) and tim barrels ($\#$). The all beta class ($*$) is the hardest one. The class has six proteins and appears separated in two groups in half of the trees (namely a,c,d), having three proteins on each group.

The protein 1ct9, from the alpha-beta class ($\wedge$), always appears within the tim-barrels group ($\#$).

The simplest generalization, namely (b), allows to obtain an almost perfect detection of the classes. Just the protein 1ct9 is misclassified. This grouping can not be achieved with the standard model of contact map.

The dendrogram (e), constructed from contact maps that includes the definition of those used in (b), also achieves an almost perfect clustering but with different ordering than (b). The contact maps produced by definition (e) has more information than (b), because the matrix stores information for every distance occurred, so there is no entry with zero, neither in the membership values nor in the contact types. Most of the values are 1, so it may happen that this novel information can be compressed extremely well so it does not deteriorate the results, as occurs in (d). In other

words, the information gained by the righmost fuzzy set of definition (e) is useless.

The dendrogram (f) is also interesting, because it shows a perfect detection of classes $\#, \wedge$ and $*$. The two members of class + appears separated. One of them, is put alone in one branch of the tree, while the other is coupled within the $*$ class.

Definitions (c) and (f) differ on the type of membership functions used, although they cover the same range of distances. The corresponding contact maps are the same in terms of the type of contacts, but the membership values differ. Such values are greater when definition (f) is applied. In turn, the changes on the fuzzy contact maps, allowed to obtain better results with respect to (c). In the former, just the class + appears separated, while in the later, both + and $*$ are.

So, (not considering protein 1CT9), definitions (a) and (c) leaded to trees where two classes appeared separated. Definitions (d) and (f) divided just one class, while the trees for definitions (b) and (e) obtained a perfect recovering of all of them.

6 Conclusions

In this contribution we focused on the protein structure comparison problem, from the point of view of fuzzy contact maps and the universal similarity metric.

The results obtained show that even the simplest generalization of the standard contact map, can produce similarity values that in turn, allowed to recover the class structure presented in the dataset used.

Also, it seems clear that the shape of the membership function is relevant for the results. This is clearly seen when Figs 4(c) and (f) are compared.

The approach proposed here works fast. Producing the fuzzy contact maps from the corresponding pre-calculated distance matrices, and performing an all against all comparison (through Eq. 8) took around 3 mins on a Pentium IV PC (496 comparisons).

Now, we are focusing in the design and test of the proposed approach on more challenging data sets.

Acknowledgments

This work is supported by Project TIC2002-04242-C03-02 (Spanish Ministry of Science and Technology), and BBSRC grant BB/C511764/1.

References

[1] C. Bennett, P. Gacs, M. Li, P. Vitanyi, and W. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44:4:1407–1423, 1998.

[2] P. Bourne and H. Weissig, editors. *Structural Bioinformatics*. Wiley-Liss, Inc, 2003.

[3] L. Chew and K.Kedem. Finding consensus shape for a protein family. In *18th ACM Symp. on Computational Geometry. Barcelona, Spain*, 2002.

[4] H. B. et.al. The protein data bank. *Nucleic Acids Research*, (28):235–242, 2000. http://www.rscb.org.

[5] N. Krasnogor and D. Pelta. Measuring the similarity of protein structures by means of the universal similarity metric. *Journal of Bioinformatics*, 20(7):1015–1021, May 2005.

[6] M. Li, X. Chen, X. Li, B. Ma, and P.Vitanyi. The similarity metric. In *Proc. of the 14th ACM-SIAM Symp. Discrete Algorithms(SODA) 2003*, 2003.

[7] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 1997.

[8] S. Lifson and C. Sander. Antiparallel and parallel beta-strands differ in amino acid residue preferences. *Nature*, 282:109–11, 1979.

[9] L. Mirny and E. Domany. Protein fold recognition and dynamics in the space of contact maps. *Proteins*, 26:391–410, 1996.

[10] D. Pelta, N. Krasnogor, C. Bousono-Calzon, J. L. Verdegay, J. Hirst, and E. Burke. A fuzzy sets based generalization of contact maps for the overlap of protein structures. *Journal of Fuzzy Sets and Systems*, 152(1):103–123, 2005.

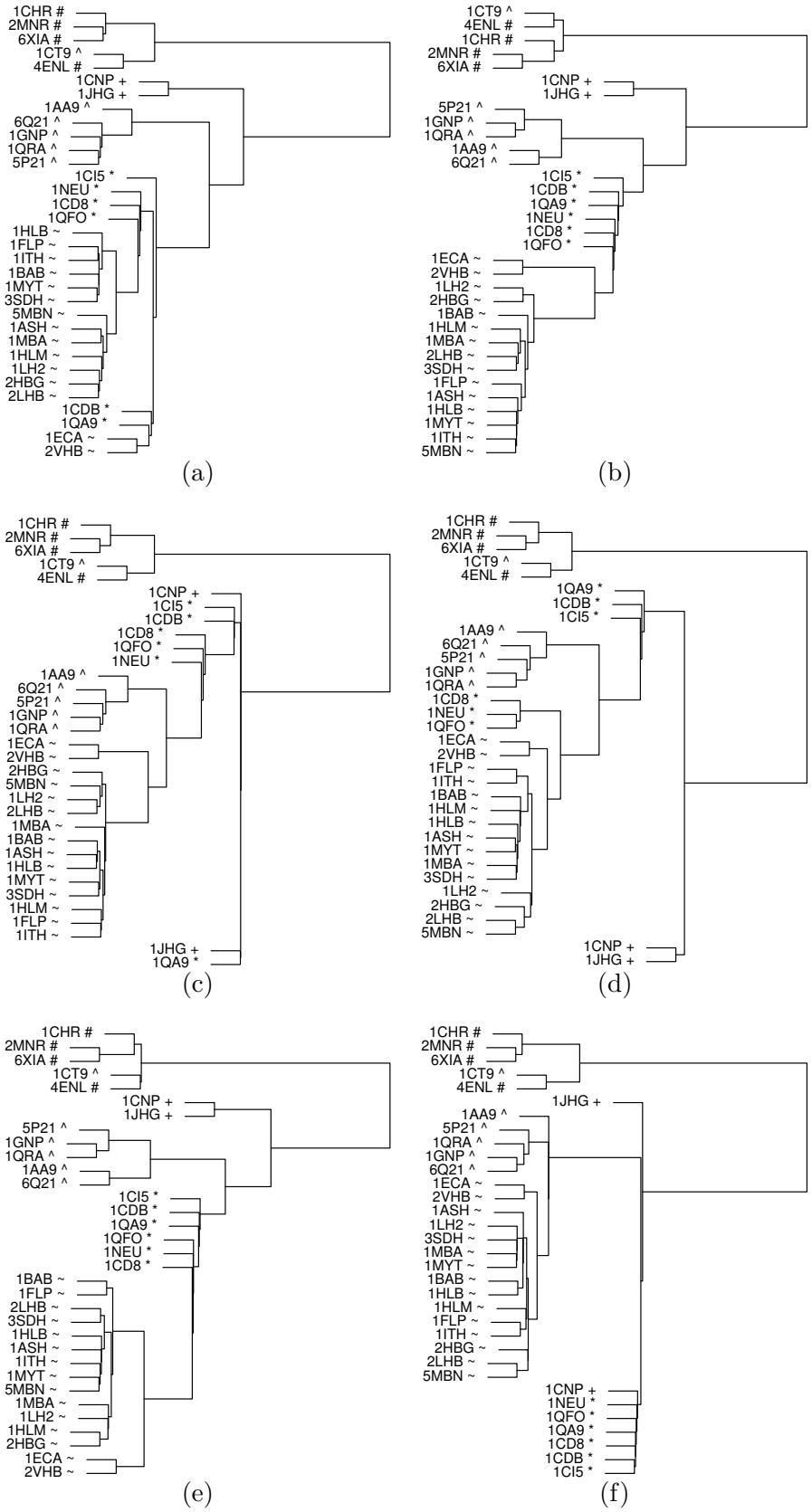

Figure 4: Clustering of the similarities values obtained using different definitions of fuzzy contact maps. The letter between brackets, corresponds to the definitions shown in Fig. 3