On Using Fuzzy Contact Maps for Protein Structure Comparison: A Methodological and Classification Study

Lluvia Morales[1] **Juan Ramón Gonzalez**[2] **David Pelta**[2]

Dept. of Computer Science and AI, University of Granada, 18071 Granada,Spain
[1] lluviamorales@ugr.es
[2] {jrgonzalez,dpelta}@decsai.ugr.es

Abstract

The comparison of protein structures is an important problem in Bioinformatics, and Soft Computing techniques were recently introduced for achieving a better representation and potentially, for getting better solving strategies. In this paper we work over the Generalized Maximum Fuzzy Contact Map Overlap model for analyzing the impact of different cycle contributions and normalizations, in order to obtain better solutions besides clearly and quality over the comparison.

Keywords: Fuzzy Contact Maps, Protein Comparison.

1 INTRODUCTION

A protein is a complex molecule composed by a linear arrangement of amino acids. Each amino acid is a multi-atom 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". Under specific physiological conditions, the linear arrangement of residues will *fold* and adopt a complex three dimensional shape. The shape thus adopted is called the *native state* (or tertiary structure) of the protein. In its native state, residues that are far away along the linear arrangement may come into proximity in three dimensional space in a fashion similar to what occurs with the extremes of a sheet of paper when used to produce complex origami shapes. The proximity relation between residues in a protein can be captured by a mathematical construct called a "contact map".

A contact map [9, 8] 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.

The comparison of proteins through their contact maps is equivalent to solving the maximum contact map overlap problem MAX-CMO [2, 1] (when the maps are crisp) or the generalized fuzzy contact map overlap problem GMAX-FCMO [4] (when the maps are fuzzy). Any of them belongs to the NP-Hard class of complexity.

In previous work we addressed the comparison of fuzzy contact maps against crisp contact maps[6] were was shown that that if we first solved the problem trough GMAX-FCMO, and then such solutions were measured as in MAX-CMO, the results obtained were better than those obtained when MAX-CMO is solved directly.

In this paper we extend the previous work by doing different kinds of analysis on the overlap computation from FMAX-FCMO and on the normalization made to classify the proteins according to them sizes.

The paper is organized as follows: In Section 2, fuzzy contact maps as well as their comparison on the MAX-CMO and GMAX-FCMO models is presented. Section 3 describes the experiments and results obtained. Finally, Section 4 is devoted to the conlutions and future work.

2 FUZZY CONTACT MAPS MODEL DESCRIPTION

Fuzzy contact maps were introduced in [4] with two aims: a) to take into account potential measurements errors in atom coordinates, and b) to allow highlighting features that occurs at different thresholds.

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:

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

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 1 (a), (b) and (c) shows three alternative definitions for "contact". Each panel in the Figure is a fuzzy contact map where a dot appears for each pair of residues having $F_{i,j} > 0$ (i.e. the support of the corresponding fuzzy set). Althought, in the same FCM can exist more than one contact type

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:

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

with the contact map C defined as:

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

i.e. up to n different thresholds and up to n different semantic interpretations of "contact" are used to define the $r \times r$ contact map being r the number of residues in the protein.

2.1 Contact Maps Comparison

A solution for the comparison of two contact maps under the Maximum Contact Map Overlap model consists of an alignment or pairing between the nodes of the two contact maps. The value of an alignment is the number of cycles of length four that appear between the corresponding graphs after the pairings. This value is called the overlap (fitness) of the two contact maps and the goal is to maximize it. For example, Figure 2 shows a sample solution for the comparison of two

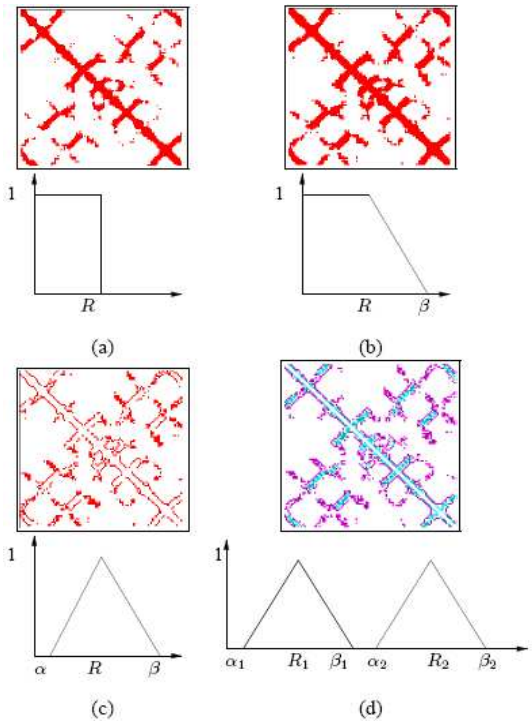

Figure 1: Four examples of contact maps. In (a) the standard model; (b) the simplest fuzzy generalization; (c) another generalization; (d) a two threshold, two membership functions fuzzy contact maps.

contact maps of 5 and 7 residues respectively. Three residues are paired (first index corresponds to the bottom graph): $1 \leftrightarrow 1$, $2 \leftrightarrow 4$ and $3 \leftrightarrow 5$, and the overlap value is two because there two cycles of length four. The first cycle is formed by arcs $1 \leftrightarrow 1$ and $3 \leftrightarrow 5$; while the second one have arcs $2 \leftrightarrow 4$ and $3 \leftrightarrow 5$ connecting the cycle.

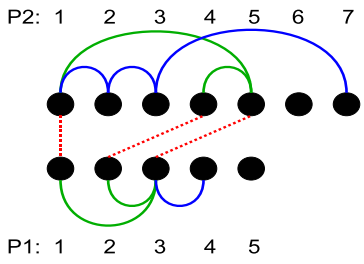

Figure 2: Two levels of contacts in a fuzzy contacts graph

Solutions for the GMAX-FCMO model have exactly the same structure as in the crisp MAX-CMO, but the overlap value is computed differently. Instead of just counting the cycles of length four in the alignment, the contribution of each cycle to the global fitness is

calculated as a function of the membership values of the contacts involved and their types. More specifically, we multiply both the membership values and then, if both contacts have the same type, then the contribution is added to the fitness. Otherwise, when contacts of different type are involved in a cycle, the contribution is subtracted from the fitness.

So, as the contact maps in Figure 2 are fuzzy, the contribution of one cycle to the fitness is the product of the membership value of the contact $1 \leftrightarrow 3$ of the bottom protein and that of the contact $1 \leftrightarrow 4$ of the upper protein, with positive sign becuse the two contacts are of the same type; and, this value is added to the product of the membership of contact $2 \leftrightarrow 3$ and the membership of contact $4 \leftrightarrow 5$.

3 EXPERIMENTS AND RESULTS

For solving GMAX-FCMO, we adapted a multi-start Variable Neighborhood Search (VNS) metaheuristic developed for MAX-CMO that was presented in [7]. It is also publicly available online as one of the methods used in the ProCKSi server [5]. This algorithm follows a standard VNS algorithm structure with just a few changes: there is an extra multi-start loop to better explore the solution space and reduce the big influence of a single initial solution; and a "simplify" function that is used after every local search to removes pairings that do not contribute to the solution fitness (it helps to avoid the saturation of the solution with useless pairings). The algorithm also uses reduced solution evaluation to recompute the cost of a neighbor solution considering only the changes from the current one, thus significantly reducing the computational time needed.

Further, we applied the GMAX-FCMO problem using the membership function described in Figure 3 which have a threshold of 8Å and where a decayng slope reduces the level of membership of a contact as the distance aproaches from 6.5 to 8.

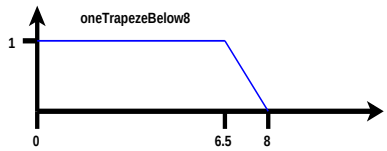

Figure 3: Experimental fuzzy function

The algorithm was proved over 150 selected protein structures from the Nh3D v3.0 test dataset[3]. This dataset has been compiled by selecting well resolved representatives from the Topology level of CATH database. The database has 806 topology representatives belonging to 40 architectures. For each architecture we selected the average structure in terms of residues and number of contacts, plus another one randomly selected. After removing dulicates, we selected 10 structures that constitues the query set. Eac query was then compared against every structure in the test dataset.

As stated before, we perform different kinds of experiments and analysis, that will be explained in the following subsections.

3.1 THE OVERLAP CONTRIBUTION

To calculate the overlap of two proteins we considered, not only the product of memberships to compute the contribution but also, we experiment with the min, max and, average value between them in order to get adapted fitness values.

Table 1: Contacts between proteins of equal architecture

	Fitness	Norm1	Norm2	Norm3	Normfuzzy
PRODUCT	0,565	0,468	0,625	0,542	0,622
MIN	0,571	0,479	0,637	0,552	0,631
MAX	0,565	0,470	0,628	0,546	0,623
MEAN	0,569	0,476	0,636	0,551	0,629

Table 2: Contacts between proteins of equal class

	Fitness	Norm1	Norm2	Norm3	Normfuzzy
PRODUCT	0,569	0,419	0,553	0,491	0,580
MIN	0,573	0,426	0,557	0,498	0,584
MAX	0,574	0,420	0,561	0,494	0,585
MEAN	0,575	0,426	0,561	0,497	0,586

As we can see on column 1 from Tables 1 and 2, the fitness value not changed significantly for any of the proposed contributions. However, the experiment realized in the next section in addition to this one confirm us relevant assumptions.

3.2 NORMALIZATION ALTERNATIVES

Altough the analysis from an optimization point of view is relevant, it is also interesting to check the quality of our method as a protein structure classifier. Moreover, we supose that overlap values are not adequate *per se* for classification because such values depend of the size of the proteins being compared. Indeed a normalization scheme should be applied and

we ilustrate that this may play a crucial role in protein classification. But, after that, we must define the *self-similarity* of a protein mesured through the corresponding fuzzy contact map as

$$selfSim\,(P_k) = \sum_{i=1}^{r_k-1} \sum_{j=i+1}^{r_k} \left(C_{i,j}^{k}\right)^2 \qquad (5)$$

Then, four alternatives on how to do normalization were used in our experiment:

1. $Norm1(P_i, P_j) =$
   $overlap(P_i, P_j)\ /\ min(contactsP_i,\ contactsP_j)$
2. $Norm2(P_i, P_j) =$
   $2\ *\ overlap(P_i, P_j)/(contactsP_i\ +\ contactsP_j)$
3. $Norm3(P_i, P_j) =$
   $\begin{cases} 0 & \text{if the contacts difference is greater than 75\%} \\ norm1(P_i, P_j) & \text{otherwise} \end{cases}$
4. $NormFuzzy(P_i, P_j) =$
   $overlap(P_i, P_j)\ /\ max(selfSim(P_i),\ selfSim(P_j))$

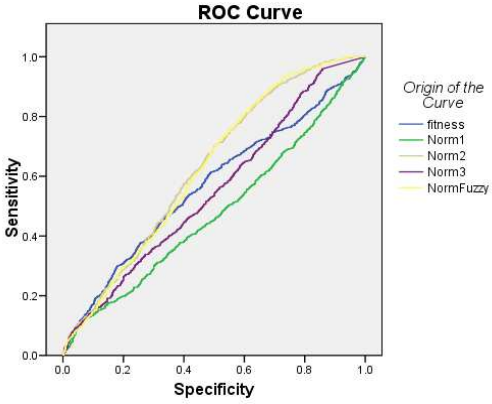

Figure 4: ROC curves for the same arquitecture of proteins and all kinds of contributions

As could be supoused by the areas described in Tables 1 and 2 for different normalization kinds. The futher analysis of the results using ROC curves, shows that *Norm2* and *NormFuzzy* normalizations are the best options for post processing the overlap in order to clasificate the proteins set studied. This statement could be proven in 4 for proteins with the same architecture and in 5 for proteins with the same class.

4 CONCLUSIONS

Acknowledgements

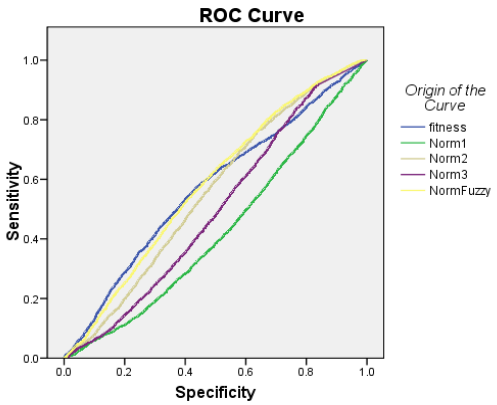

Figure 5: ROC curves for the same class of proteins and all kinds of contributions

This work is supported by the Mexican National Council on Science and Technology (CONACYT) and ...

References

[1] A. Caprara, R. Carr, S. Istrail, G. Lancia, and B. Walenz, "1001 optimal pdb structure alignments: integer programming methods for finding the maximum contact map overlap." *J Comput Biol*, vol. 11, no. 1, pp. 27–52, 2004.

[2] B. Carr, W. Hart, N. Krasnogor, E. Burke, J. Hirst, and J. Smith, "Alignment of protein structures with a memetic evolutionary algorithm," in *GECCO-2002: Proceedings of the Genetic and Evolutionary Computation Conference*. Morgan Kaufman, 2002.

[3] B. Thiruv, G. Quon, S. A. Saldanha, and B. Steipe, "Nh3D: A Reference Dataset of Non-homologous protein structures ," in *BMC Structural Biology 2005 5(12)*.

[4] 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*, vol. 152, no. 1, pp. 103–123, 2005.

[5] J. B. D. Barthel, J. D. Hirst and N. Krasnogor, "Procksi: A meta-server for protein comparison using kolmogorov and other similarity measures," in *5th European Conference on Computational Biology (ECCB06), Israel*, 2007.

[6] J. R. González and D. A. Pelta, "On Using Fuzzy contact Maps for Protein structure Comparison," in *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International.*

[7] J. R. González, D. A. Pelta, and J. M. Moreno-Vega, "Multistart VNS for the Maximum Contact Map Overlap Problem," in *Proceedings of 18th MECVNS, Tenerife, Spain*, 2005.

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

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