A Fuzzy Sets based Generalization of Contact Maps for the Overlap of Protein Structures

David Pelta [a], Natalio Krasnogor [b], Carlos Bousono-Calzon [c], José L. Verdegay [a], Edmund Burke [b]

[a] *Department of Computer Science and Artificial Intelligence E.T.S.I. Informática, Universidad de Granada, 18071, Granada, Spain*

[b] *Automated Scheduling, Optimisation and Planning Research Group University of Nottingham, Nottingham, NG8 1BB, U.K.*

[c] *Universidad Carlos III de Madrid, Avenida de la Universidad 30, Leganes(Madrid), Spain*

Abstract

The comparison of protein structures is an important problem in bioinformatics. As a protein biological role is derived from its three dimensional native state, the comparison of a new protein structure (with unknown function) with other protein structures (with known biological activity) can shed light into the biological role of the former. Consequently, advances in the comparison (and clustering) of proteins accordingly to their three dimensional configurations might also have an impact on drug discovery and other biomedical research that relies on understanding the inter-relations between structure and function in proteins.

The contributions on this paper are: First, we propose a generalization of the Maximum Contact Map Overlap Problem (MAX-CMO) by means of fuzzy sets and systems. The MAX-CMO is a model for protein structure comparison. In our new model, *Generalized Maximum Fuzzy Contact Map Overlap* (GMAX-FCMO), a contact map is defined by means of one (or more) *fuzzy* thresholds and one (or more) *membership functions*. The advantages and limitations of our new model are discussed. Second, we show how a fuzzy sets based metaheuristic can be used to compute protein similarities based on the new model. Finally, we compute the protein structure similarity of real-world proteins and show how our new model correctly measures their (di)similarity.

Key words: Protein Structure Comparison, Protein Structure Alignment, Fuzzy Sets, Maximum Contact Map Overlap, FANS, Universal Similarity Metric

1 Introduction

The comparison of the 3D structures of protein molecules is a challenging problem. The search for effective solution techniques for this problem, is justified because such tools aid scientists in the development of procedures for drug design, in the identification of new types of protein architecture, in the organization of the known universe of protein structures and could assist in the discovery of unexpected evolutionary and functional inter-relations between them [14,15]. Good comparison techniques for protein structures could also be used in the evaluation of *ab-initio, threading or homology modelling* structure predictions. It would be safe to argue that the comparison of proteins' structures, and their clustering accordingly to similarity, is a fundamental aspect of today's biomedical research.

There is no general agreement on which is the best similarity measure to use and what computational method must be harnessed in order to produce the required measurement. Each measure is usually based on a particular biological conception of structural similarity and they generally use different algorithmic strategies. Methodologies based on dynamic programming [27], comparisons of distance matrices [13], maximal common sub-graph detection [1], geometrical matching [28], consensus shapes[9] and consensus structures[23] are but a few of the available tools for structural comparison. In this context we had recently proposed the application of a Universal Similarity Measure to compare protein structures [19] which subsumes (under certain conditions) every other possible similarity concept.

Most of the existing methods implicitly accept that a suitable scoring function can be defined for which optimum values correspond to the best possible structural match between two proteins. It is implicitly assumed that, based on these optimal matches, similarity between protein structures can be captured.

One of the latest approaches for structural matching was introduced in [11] and extended in [6,7,17,21]. This method is based on the maximum overlap (also called alignment) of contact maps. Although the problem of maximizing the overlap between two contact maps was shown to be NP-hard [10,11,16] good approximate algorithms exist. The current state of the art to obtain single (sub)optimal maximum contact map overlaps is reported in [5] while

Email addresses: dpelta@decsai.ugr.es (David Pelta), Natalio.Krasnogor@Nottingham.ac.uk (Natalio Krasnogor), cbousono@uc3m.edu.es (Carlos Bousono-Calzon), verdegay@decsai.ugr.es (José L. Verdegay), ekb@nott.cs.ac.uk (Edmund Burke).

[1] Research supported in part by Project TIC2002-04242-C03-02 from Spanish Ministry of Science and Technology and the British BBSRC/EPSRC Bioinformatics initiative (42/BIO14458).

multiple (sub)optimal overlaps can be obtained with a memetic algorithm [17].

In [4], Bourne and Shindyallov say:

> *Consider a spectrum that at one end maximizes the geometric relationship between two proteins and at the other provides the maximum amount of biological significance in the alignment. Depending on the task at hand, you may wish to be at one end of the spectrum or the other, or in the middle*

and they go on saying:

> *An important consideration when using any structural alignment method is to consider the nature of the problem you are trying to solve and to experiment with a variety of methods*

It is with this spirit in mind that we extended the standard contact map definition and the associate MAX-CMO problem with the aid of fuzzy sets and systems.

Maximum contact map overlaps in their basic definition are "sanitized" mathematical constructs that may, sometimes, leave out important features of protein topological fingerprints for the sake of mathematical solvability. As an example consider the uncertainties derived from errors in the determination of the atomic Cartesian coordinates by X-Ray Crystallography or NMR. Experimental errors range from 0.01 $\mathring{A}$ to 1.27 $\mathring{A}$ which is close the value of some co-valent bonds [22]. This type of uncertainties cannot be handled with the standard model. Moreover, crisp contact maps are computed based on only one preferred threshold value and they loose all the information related to contacts at alternative thresholds. That is, contact maps can only provide a coarse approximation to protein's true topological features.

In order to solve the problems associated with standard contact maps we adopted a fuzzified version that includes two distinct thresholds: one to capture short-distance features while the other one represents long-distance patterns. Moreover, unlike MAX-CMO our similarity distance is normalized. Fuzzy contact maps lead to the formulation of a more general combinatorial problem called *General Maximum Fuzzy Contact Map Overlap Problem*.

This new model and its associated combinatorial problem allow the end-user, i.e. the biologist, to capture a variety of protein's topological features and the uncertainties related to their experimental determination. The biologist can choose on what region of the Bourne and Shindyallov spectrum he/she wants to work: at one end of the spectrum where exact and (almost) optimal solutions can be computed from a **mathematical** viewpoint (i.e. crisp MAX-CMO) all the way up to the other end of the spectrum where more **biologically** meaningful solutions could obtained.

This article is organized as follows: First we describe the standard (i.e. crisp) model of contact maps and the Maximum Contact Map Overlap Problem in the context of protein structure comparison. Our contributions start in section 3 with a generalization of contact maps and the introduction in section 4 of the GMAX-FCMO problem. Section 5 describes a fuzzy based metaheuristic, *FANS*, which can solve GMAX-FCMO. In section 6 we show how GMAX-FCMO and *FANS* are used to measure protein similarity. The paper concludes in section 7 where future work is suggested.

2 Protein Structure Comparison by Contact Map Overlaps

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 though-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. This is illustrated in figures 1(a) and (b). The proximity relation between residues in a protein is captured by a mathematical construct called a contact map. In (c) the contact map for the proximity relation in (b) is depicted as a graph.

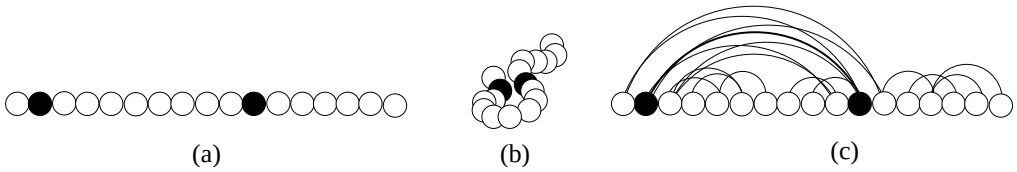


Fig. 1. Illustrations of (a) an unfolded protein, (b) a protein after folding, and (c) a contact map graph of the final folding.

2.1 The Standard (crisp) Contact Map

The contact map [24,8,25] is a concise representation of a protein's native three-dimensional 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 $\Re$ (measured in Angstroms) in the protein's native fold. Oftentimes $\Re$ is called the "threshold" of the contact map (usual values for $\Re$ are between 2 and 9 $\mathring{A}$). The graphical representation of the contact map for the fold in Fig.1(b) is shown also in Fig.2 as a dot-matrix representation.

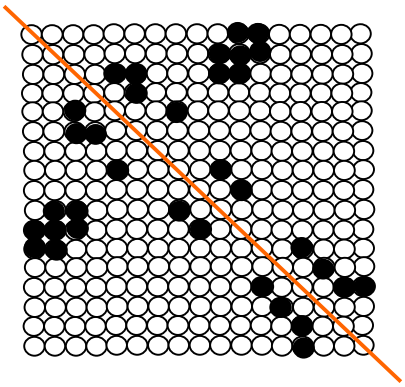


Fig. 2. A dot-matrix representation for the contact map of the fold in Fig.1(b).

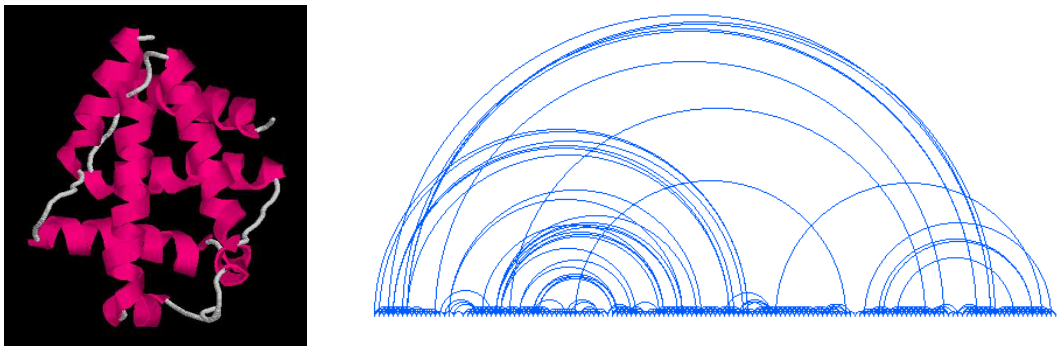
Contact maps might be calculated by taking into account the distance of the C_α atoms of the residues under consideration, or the minimum distance between *any* two atoms belonging to those residues. In some cases, contact maps are computed based on the distances between the centers of mass of the side chains of residues. The contact map captures the three dimensional structure of proteins and certain structural features are conspicuous when the contact map is represented graphically. Consider for example a contact map represented as a white-black dot-matrix. If the protein structure associated to a given contact map contained $\alpha-$helices, then wide bands on its main diagonal will be visible. On the other hand, if $\beta - sheets$ were present in the protein structure, these will show as bands parallel or perpendicular to the diagonal.

2.2 Maximum (crisp) Contact Map Overlap Formulation

Protein similarity can be computed by aligning the two contact maps of a pair of proteins. An alignment of two proteins is a pairing of amino acids between them. For example, Figures 3(a) and 4(a) show the structures of two related proteins taken from the Protein Data Bank (PDB)[2]. These proteins share a 6 helices structural motif. Figures 3(b) and 4(b) display the contact map of the proteins as a graph in which each contact between two residues corresponds

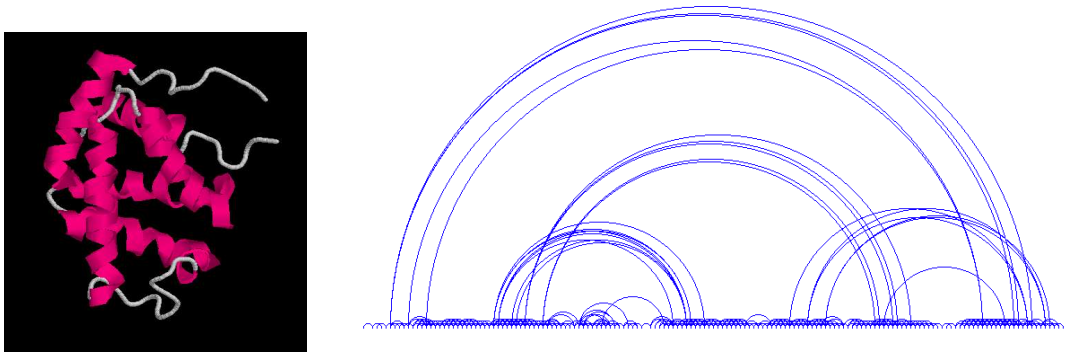
to an edge. Figure 5 shows a candidate alignment between the contact maps of these protein structures.

The *alignment* between two contact maps is an assignment of residues in the first contact map to residues on the second contact map. Residues that are thus aligned are considered equivalent. Further, consider a pair of contacts, one from each protein. We say that such a pair of contacts is equivalent if the pairs of residues that define the end-points of these contacts are equivalent. In the *Max CMO problem*, the value of an alignment between a pair of proteins is the number of equivalent contacts between these proteins. This number is called the *overlap* of the contact maps and the goal is to maximize this value. That is, the larger that value the more similar the two proteins are considered.



(a) (b)

Fig. 3. Native structure (a) for protein 1ash taken from the PDB [2] and its contact map (b).



(a) (b)

Fig. 4. Native structure (a) for protein 1hlm taken from the PDB [2] and its contact map (b).

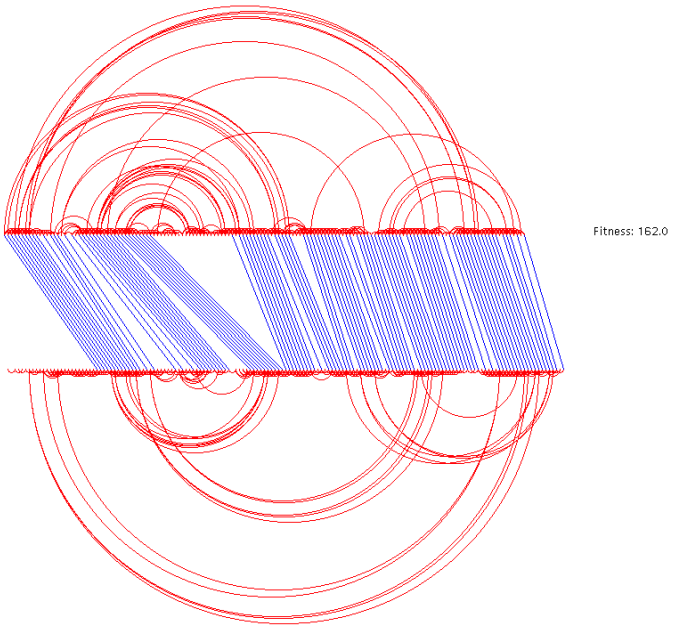


Fig. 5. A potential alignment (of value 162) for the contact maps of similar proteins 1ash and 1hlm. The optimal alignment based on the crisp model has a value of 279.

3 The New Generalized Fuzzy Contact Maps

Under the standard 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. These are described next.

3.1 Introducing Fuzzy Contact Maps

Using a membership function $\mu()$ we define *fuzzy contacts* as those made by two residues that are "roughly" at a distance $\Re$. An example of a membership function $\mu()$, which establishes the level of contact between two residues, is depicted in Fig.6. Formally, a fuzzy contact is defined by:

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

where $\overline{[i,j]}$ stands for the Euclidean distance between residues i and 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. In the example in Fig. 6, setting $\alpha = 1$ turns the fuzzy contact map into the standard crisp contact map.

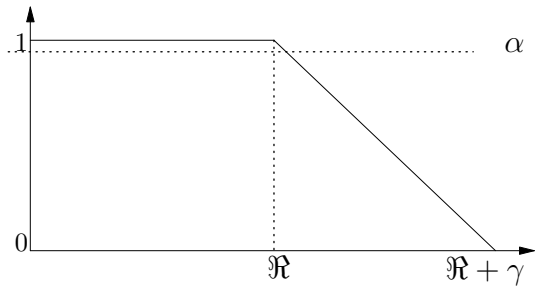


Fig. 6. Membership function for the fuzzy set of residues whose inter distance is "at most $\Re$ plus tolerance γ Å". An α-cut is also displayed. When $\alpha = 1$ a standard (i.e. crisp) contact map is obtained.

3.1.1 Discussion on Fuzzy Contact Maps

The reader must note that different biological structural features might be better captured by changing the meaning of "contact". While this cannot be done under the standard model, the generalized fuzzy version allows the user to do precisely that by readily changing the shape and parameters of the membership function $\mu()$.

Figure 7 shows three alternative meanings for "contact" as realized by three different membership functions. Recall that each panel in the figure is a fuzzy contact map in which a colored dot appears for each pair of residues such that $F_{i,j} > 0$ (i.e. the support of the corresponding fuzzy set). The leftmost membership function is simply one that corresponds to the standard contact map. The rightmost $\mu()$ defines a contact map that contains those same contacts that appear under the standard model plus some new contacts. The new contacts are those formed by residues that are located at a slightly bigger distance than the preferred threshold $\Re$. In turn, the membership function in the middle panel defines a new contact map where two residues are deemed to be in contact if they are located "at $\Re$ Å away with symmetric tolerance γ", that is, at a slightly smaller or slightly bigger distance.

3.2 Generalizing Fuzzy Contact Maps

The fuzzy contact maps defined above can be further generalized by removing the constraint (in the original model) of having only one threshold $\Re$ as a reference distance. That is, we can extend our model in two ways:

- 1-threshold fuzzy contact maps are generalized to $n-$threshold fuzzy contact maps: this is motivated by the fact that different contact patterns may arise simultaneously at different euclidean distances, e.g., the average inter-residue distances within α-helices are different than those present in a $\beta-$sheet or loop.

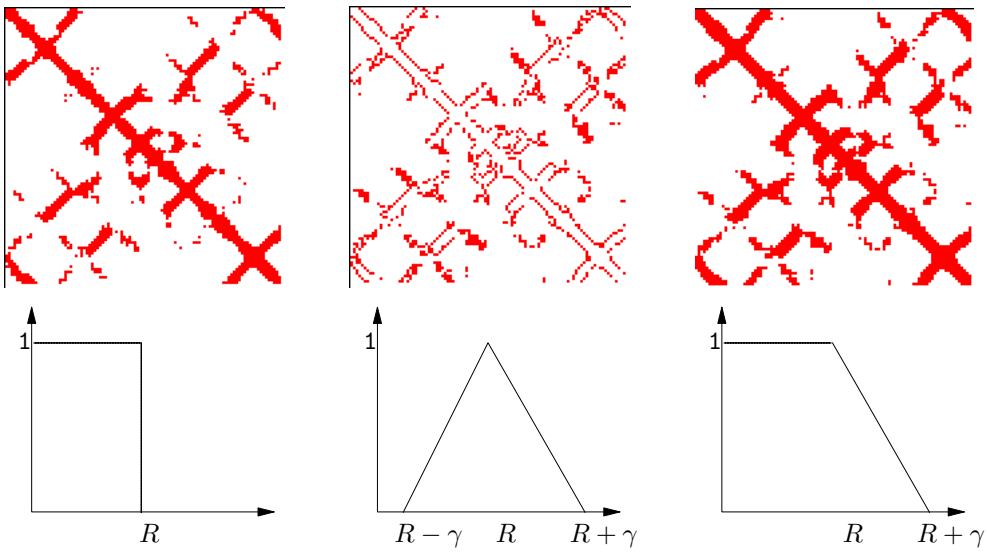


Fig. 7. Three different meaning for "contact". In the figures $R = 10$Å, $\gamma = 1.2$Å. The fuzzy contact maps displayed are for protein 1AA9.

- 1-membership function fuzzy contact maps are generalized to $m-$membership function contact maps: the rationale behind this extension to the model is that different *meanings* for "contact" could be simultaneously needed by the biologist. A 1-membership function contact map (either crisp or fuzzy) cannot capture this need.

Under *General Fuzzy Contact Maps*, the standard crisp contact maps (with one threshold and a default membership function) are readily included at one extreme of the spectrum of potential maps.

The formal definition of a General Fuzzy Contact is given by:

$$F_{i,j} = max\{\mu_1(\overline{[i,j]}, \Re_1), \mu_2(\overline{[i,j]}, \Re_2), \ldots, \mu_m(\overline{[i,j]}, \Re_n)\} \tag{3}$$

with the contact map C defined as:

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

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 being r the number of residues in the protein. Please note that in general n and m could form any order relationship, that is, $n < m, n \leq m, n = m, n \geq m$ or $n > m$.

3.3 A 2−Threshold, 2−Membership Functions Fuzzy Contact Map Example

As a particular example consider a 2-thresholds, 2-membership functions fuzzy contact map intended to simultaneously highlight *short* and *long* structural patterns. The membership functions μ_1, μ_2 for short and long patterns are defined in such a way that they do not overlap and with $\Re_1 < \Re_2$. Each entry in the fuzzy contact map will be either of type short or long with $F_{i,j} \in [0, 1]$ indicating just how short or how long (with respect to the corresponding thresholds) that entry is.

(a) (b) (c)

Fig. 8. Different contact map models for protein 1AAG. In (a) standard contact map with $\Re = 6.5\mathring{A}$, (b) 1-threshold Fuzzy Contact Map with $\Re = 6.5\mathring{A}$ and $\mu()$ as in 7(c), and (c) 2-thresholds Fuzzy Contact Map with $\Re_1 = 6.5\mathring{A}$, $\Re_2 = 10.0\mathring{A}$,$\gamma = 1.2\mathring{A}$ and $\mu_1(), \mu_2()$ as in 7(c)

We computed three different contact maps for protein 1AA9. These are shown in Fig. 8 where the parameters used to create them are also described. The first contact map uses the standard model, the second is a generalized contact map with one threshold and one membership function, and the third is a 2-thresholds, 2-membership functions fuzzy contact map. A simple visual inspection shows that the resulting patterns are different across the models and that the more general model in 8(c) presents a richer set of features. Moreover, the *identity* of the contact (i.e. from which membership function it arises) is color coded adding another information dimension to our model.

3.4 Discussion on the Generalized Fuzzy Contact Maps

The $n-$threshold $m-$membership functions fuzzy contact maps are a powerful tool to exploit models of contact maps that span the Bourne and Shindyallov spectrum. In the simplest case of a 1-threshold with 1-membership function having $\alpha - cut = 1.0$, the standard crisp contact map model is obtained. On the other hand, as different thresholds are added and various "meanings" are

attached to the *contact* concept (by means of a variety of membership functions) richer and richer models are obtained. The proposed model empowers the end user with the ability to select how much "biology" to include in the mathematical construct. Moreover, this is done accordingly to the task at hand and not to some arbitrary mathematical constraint.

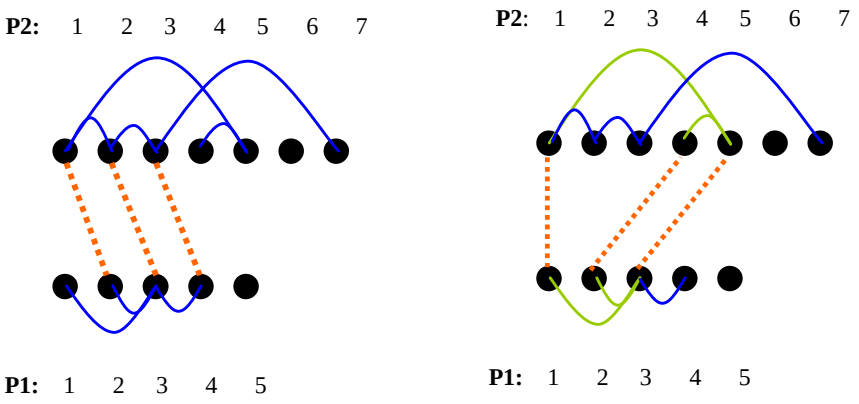
4 The Generalized Maximum Fuzzy Contact Map Overlap Problem

The new contact map model introduced in section 3 is not only a good tool to visually inspect protein structures but, it can also be used to measure the similarity between a pair of proteins. In order to achieve this, we extend the maximum (crisp) contact map overlap problem in such a way that the additional information contained in the generalized fuzzy contact maps could be used. When using general fuzzy contact maps, protein similarity is measured by solving the *Generalized Maximum Fuzzy Contact Map Overlap Problem* (GMAX-FCMO).

Recall that a contact map $C^{r\times r}$ can be graphically represented either as a dot-matrix (like in Fig. 8) or as graphs (like in Fig. 3). When working and reasoning in terms of "overlaps" it is easier to use the later representation. Under the standard model (i.e. MAX-CMO) the value of an overlap gives a measure of the similarity of two proteins. Please note that in this case the overlap value is not a *normalized* similarity measure. That is, if for example the overlap value of proteins A and B is 100 and that of C and D is also 100, it does not follow from that information that A is as similar to B as C is to D. The reason for that is that A and B could be 120 residues long each while C and D 200 and 400 residues long respectively. In that case we could certainly argue that an overlap value of 100 in the A, B case is a rather accurate assessment of their similarity while for C and D it is not.

Figure 9(a) shows an overlap between the contact maps of protein $P1$ and $P2$. The overlap or alignment is represented by the red lines. The value of that alignment is 2. This number is reached by counting the number of size 4 cycles made up by a contact in P1 (in blue), a red edge , a contact in P2 (in blue) and a second red edge.

In Fig.9(b) the same connectivity structure is maintained but the contacts arise from two different membership functions, i.e. the *meaning* of contact differ. The contacts arising from the first membership function are drawn in green and the ones arising from the second membership function in blue. Under the generalized overlap model, aligning residues in $P1$ to residues in $P2$ as in Fig.9(a) would constitute a "semantic mismatch" as green and blue



(a) (b)

Fig. 9. An example of an overlap between two contact maps. In (a) the contact maps (represented as graphs rather than dot-matrices) are crisp. In (b) the contact maps are generalized fuzzy ones, the different edge colors represent different membership functions.

contacts come from different membership functions and hence they should not be aligned. On the other hand, the alignment in (b) preserves the semantic meaning of the contacts. That is, under the GMAX-FCMO problem we need not only to maximize the overlap value (i.e. number of size 4 cycles) but it is also necessary to preserve the semantic meaning of the contacts that are thus aligned. We formally define the new problem next.

Given:

- two proteins P_1, P_2 with r_1, r_2 residues each and (without loss of generality) $r_1 \leq r_2$.
- two 2-threshold, 2-membership functions contact maps $C_1 = (c^1_{i,j}), C_2 = (c^2_{i,j})$
- two matrices $T_1 = (t^1_{i,j}), T_2 = (t^2_{i,j})$ indicating the type or meaning of each contact [2].

we define an overlap as a (partial) ordered mapping $\sigma : r_1 \mapsto r_2$ such that if $i, j \in r_1, i < j$ then $\sigma(i) < \sigma(j)$.

A cycle s is defined as a quaternion $(i, j, \sigma(i), \sigma(j))$. The *contribution* of the

[2] T^i matrices are used to simplify the explanation. From an implementation viewpoint all the information, i.e. membership values and types, could be stored in the same contact map.

cycle s to the value of the overlap is given by:

$$P(s) = (c^1_{i,j} * c^2_{\sigma(i),\sigma(j)}) * (t^1_{i,j} \circledast t^2_{\sigma(i),\sigma(j)})$$

The operator $a \circledast b$ returns 1 iff $a = b$ and -1 otherwise. The first term of P is simply the product of the membership value of a contact in the first protein structure by the membership value of a contact in the second protein. The second term is the *compatibility* term which measures whether the contacts arise from the same membership function or not (i.e. whether their semantic meanings match). In the GMAX-FCMO we seek to maximize the sum of the contributions of every *compatible* cycle s induced by the overlap σ. Please note that these definitions are easy to use with $m-$membership functions and $n-$thresholds fuzzy contact maps.

The Generalized Maximum Fuzzy Contact Map Overlap thus requires that a σ be found such that it maximizes the sum of the contributions of the *compatible* cycles.

Once an optimal overlap has been found, it must then be normalized. That is, if the optimal overlap for proteins P_1, P_2 is *opt* the *similarity* is defined as:

$$SIM(P_1, P_2) = \frac{opt}{MAX\{selfSim(P_1), selfSim(P_2)\}} \tag{5}$$

where $SelfSim(P_k) = \sum_{i=1}^{r_k-1} \sum_{j=i+1}^{r_k} (C^k_{i,j})^2$ stands for the self-similarity of a protein measured through the corresponding fuzzy contact map (i.e. the sum of the squared fuzzy contact map entries in the matrix's top triangle). The idea behind this calculation is to overlap a protein with itself [3] to obtain an upper bound for the value of similarity. The normalization thus is done using the maximum of the self-similarities of fuzzy contact maps C^1 and C^2. The reader should note that were the two contact maps crisp maps, then the normalization could be simply done based on the maximum of the sizes of the two contact maps.

As the compatibility term in $P(s)$ can take negative values, it is unknown at present where the approach in [5] can be use to solve GMAX-FCMO. That is why we present next a fuzzy sets based metaheuristic that can reliably compute $SIM(P1, P2)$.

[3] In this case, we consider the function σ as the identity function.

5 Fuzzy Adaptive Neighborhood Search for GMAX-FCMO

The Fuzzy Adaptive Neighborhood Search (*FANS*) [3,26,18] metaheuristic is a fuzzy sets based extension to the classical Variable Neighborhood Search (VNS)[12]. Both FANS and VNS are local search methods in which the neighborhood used to sample the solution space (e.g. the space of all possible overlaps or alignments) is systematically and dynamically adjusted during the search process. The motivation for the systematic change of neighborhood is to allow the local search to proceed beyond a local optimum. As local optimae are function of the neighborhood employed, vigorously changing the move operator allows to bypass poor optimae.

In contrast with VNS, FANS provides a second method to escape local optimae and continue the search in promising regions of the search space. To achieve this a fuzzy objective function is used. The objective function defines which of the neighboring solutions to the current best are deemed "acceptable" for further exploration. In the current implementation of *FANS* neighboring solutions are explored one at a time and as soon as an acceptable one is found the search continues from that one.

FANS has been extensively tested on a variety of domains (e.g. [3,26]) and found particularly efficient in the protein structure prediction problem [18]. We briefly describe next the basic components of *FANS* for solving the GMAX-FCMO problem.

5.1 Solution Representation

An overlap for GMAX-FCMO is represented as an integer vector σ of size r_1, where r_1 is the size of the shortest protein. A value $\sigma[i] = i'$ means that residue i in the first protein is aligned with residue i' in the second one. If σ is a partial mapping then some of the positions in the vector might be undefined. An undefined alignment is represented as -1.

5.2 Neighborhood Operators

A neighborhood operator assigns to a given overlap σ a set of alternative overlaps: $N(\sigma) = \{\sigma', \textit{such that } \sigma' \textit{ is a valid overlap}\}$. *FANS* defines three neighborhood operators which are applied on a randomized basis (with equal chances) to the current best solution:

- $N_1(\sigma)$: inserts a random alignment into the overlap σ.

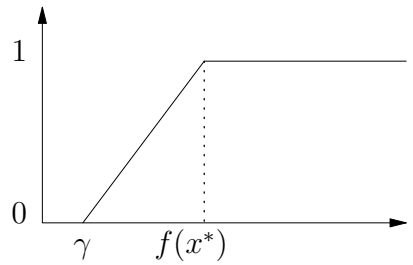


Fig. 10. Fuzzy Valuation used in *FANS*. The cost of a reference solution $f(x^*)$ provides a measure to control the acceptance of neighborhood solutions.

- $N_2(\sigma)$: inserts two random alignments into the overlap σ.
- $N_3(\sigma)$: changes the alignments in σ to the left or right.

Once one of the above neighborhoods is selected it is applied k times to a solution. Initially $k = 3$.

5.3 Adapting k

The length of the search done with each neighborhood is governed by the parameter k. If after a while the three neighborhoods fail to return an acceptable solution then k is decremented by 1. As the search progresses, the value of k decreases, turning the search into a more exploitative one. In this way, the search starts doing more exploration (performing several changes in the current solution) then exploitation.

5.4 Fuzzy Objective Function Valuation

The fuzzy valuation used here is based on the overlap value of the current solution. The "acceptability" of a neighbor solution is measured in terms of the relation between its cost and that of the current solution. In this particular implementation we employ the fuzzy objective function as in Fig. 10. The reader must note that our definitions allow the algorithm to transit to solutions with worst overlap values than the current reference solution.

6 Using GMAX-FCMO and *FANS* to Measure the Similarity of Protein Structures

The formulation of GMAX-FCMO and its solution with *FANS* allow us to compute the similarities that exist among a given set of protein structures. In

this section we show the robustness of the proposed method by testing it on three well known data sets. In [20] we used these same data sets to test the so called "Universal Similarity Metric" for protein structures.

The Chew-Kedem Data Set:

This data set was used in [9] to evaluate a new method for measuring consensus shapes. These are 32 medium size proteins of 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).

Skolnick Data Set: This data set was used in various recent papers related to structural comparison of proteins[17,6,7,21]. We selected here only 32 of these proteins: 1ntr, 1nat, 1qmp, 1rn1, 3chy, 4tmy, 1bo0, 1dbw,1byo, 1baw, 1kdi, 1nin, 1pla, 2b3i, 2pcy, 2plt,3ypi, 8tim, 1tmh, 1tre, 1tri, 1ydv, 1hti, 1amk, 1awz, 1b9b, 1btm, 1bcf, 1b7i, 1dps, 1fha, 1rcd, 1ier.

Leluk-Konieczny-Roterman data set: This is a small data set recently employed in [23] to test a new similarity measure based on geometric parameters of polypeptide chains: 1aat,1azx,2ach,7api,1ova,2ant.

Our new results will be assessed based on these three sets and in reference to results previously reported in [20].

6.1 The GMAX-FCMO Similarity Measure Protocol

The following protocol was used to compute the similarity of protein structures based on the proposed model and algorithm. The reader must note that this protocol is generic in the sense that it can be applied not only to the data sets used here but to any set of proteins.

(1) Extract from each pdb file the first chain. If other than chain 'A' is used from the pdb file this is shown in the text as pdb accession number and a letter, e.g. 1bab**B**. A script to extract the first chain from a given pdb can be found in http://www.cs.nott.ac.uk/˜nxk/protocol.html .
(2) Produce a generalized fuzzy contact map for each of the pdb files in the dataset. The thresholds are fixed to $R_1 = 6.5$ and $R_2 = 10$, using $\gamma = 1.5$. The membership function employed appears in Fig. 7(b). The distances are measured from the C_α atoms.
(3) For each pair of generalized fuzzy contact maps c_1, c_2 compute its similarity, $SIM(c_1, c_2)$, through solving the Generalized Maximum Fuzzy Contact Map Overlap Problem as per Eq.5.

(4) Eq.5 is computed with *FANS*. The algorithm is executed with three different random seeds, leading to three values of similarity for every pair.
(5) With the resulting pairwise similarity matrix apply a clustering technique to visualize the data.

6.2 Normalized Similarity Matrix and Cluster

We applied the protocol described above to the Chew-Kedem data set first. For each one of the three runs of *FANS* we obtained a normalized similarity matrix for the data set. Two of the three similarity matrices are shown in tables 1, 2, 3 and 4 where the first two contain the best similarities obtained with the randomized search algorithm. In turn, the third and fourth similarity tables show the smallest values found with the *FANS* and they are included here for comparison purposes (see below).

We fed these similarity matrices obtained after solving the GMAX-FCMO to an "off-the-shelf" clustering method (step five in the protocol) to visually inspect the results. More specifically, we run the clustering server located in http://www2.biology.ualberta.ca/jbrzusto/cluster.php . The web-server executes a combinatorial hierarchical clustering process that begins with each structure in a cluster. When more than one cluster exist then they are combined in a pairwise fashion, i.e., the two closest cluster are combined into a new one. Then an inter-cluster distance is calculated between the new cluster and the pre-existing ones. The inter-cluster distance was calculated as the unweighted arithmetic average distance (i.e. GMAX-FCMO distances) between a protein structure in one cluster and a protein structure in a second cluster.

The clusters obtained are shown in Fig. 11. Two different trees are displayed, the first corresponding to tables 1 and 2 and the second to tables 3 and 4. Symbols were attached to the protein names to show more clearly how GMAX-FCMO and *FANS* can correctly classify the proteins accordingly to their families. Moreover, for the Chew-Kedem data set our results are of comparable quality to those described in [17,20].

We conducted additional experiments of our new approach with the Skolnick and Leluk-Konieczny-Roterman data sets. The results obtained here are in close agreement to those given by state of the art structural comparison techniques[20,17,5]. Due to space limitations we include here only the comparison between the Universal Similarity Metric and the GMAX-FCMO Metric on the Skolnick data sets (the smaller Leluk-Konieczny-Roterman cluster is not shown). The clusters obtained are jointly shown in Fig. 12. A detailed analysis of the two clusters shows the correct assessment of similarity that our new method achieves. That is, proteins are almost perfectly clustered to-

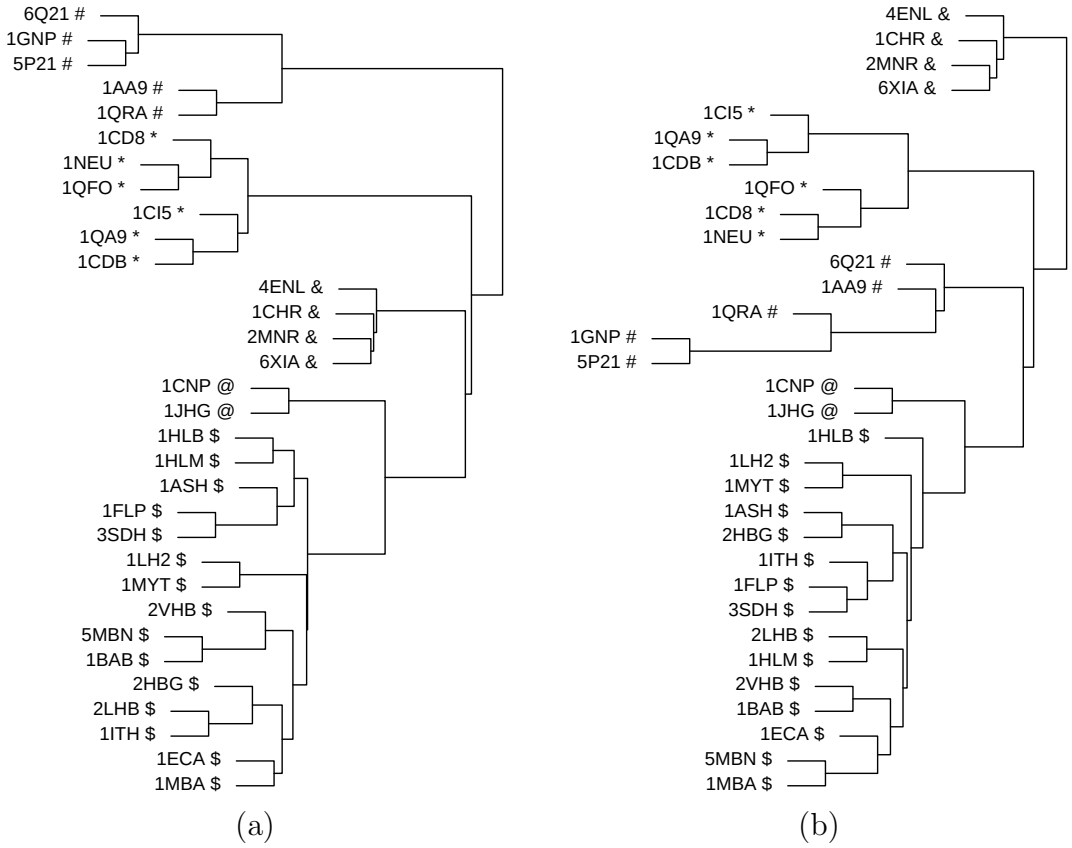


Fig. 11. Clustering Chew-Kedem data set proteins based on their similarity: using tables 1 and 2 in (a) and tables 3 and 4 in (b)

gether accordingly to the families to which they belong: (1)Flavodoxin-like CheY-related(1ntr, 1nat, 1qmp, 1rn1, 3chy, 4tmy, 1bo0, 1dbw), (2) Plastocyanin (1byo, 1baw, 1kdi, 1nin, 1pla, 2b3i, 2pcy, 2plt), (3) TIM-Barrel (3ypi, 8tim, 1tmh, 1tre, 1tri, 1ydv, 1hti, 1amk, 1awz, 1b9b, 1btm) and (4) Ferratin like (1bcf, 1b7i, 1dps, 1fha, 1rcd, 1ier). The whole set of results is publicly available at http://decsai.ugr.es/˜dpelta/GMAXFCMO/index.html

These results clearly indicate that solving the GMAX-FCMO, even with a simple heuristic like *FANS*, allows to obtain a correct *biological* ranking of similarities between protein structures. The clusters thus obtained replicate correctly the composition of the families in the Chew-Kedem, Skolnick and Leluk-Konieczy-Roterman data sets.

It is also important to remark that although *FANS* is a stochastic algorithm, the normalized measures of similarities we obtain are robust. That is, although the similarity tables arise from 3 different runs of the algorithm, the resulting clusters are very similar.

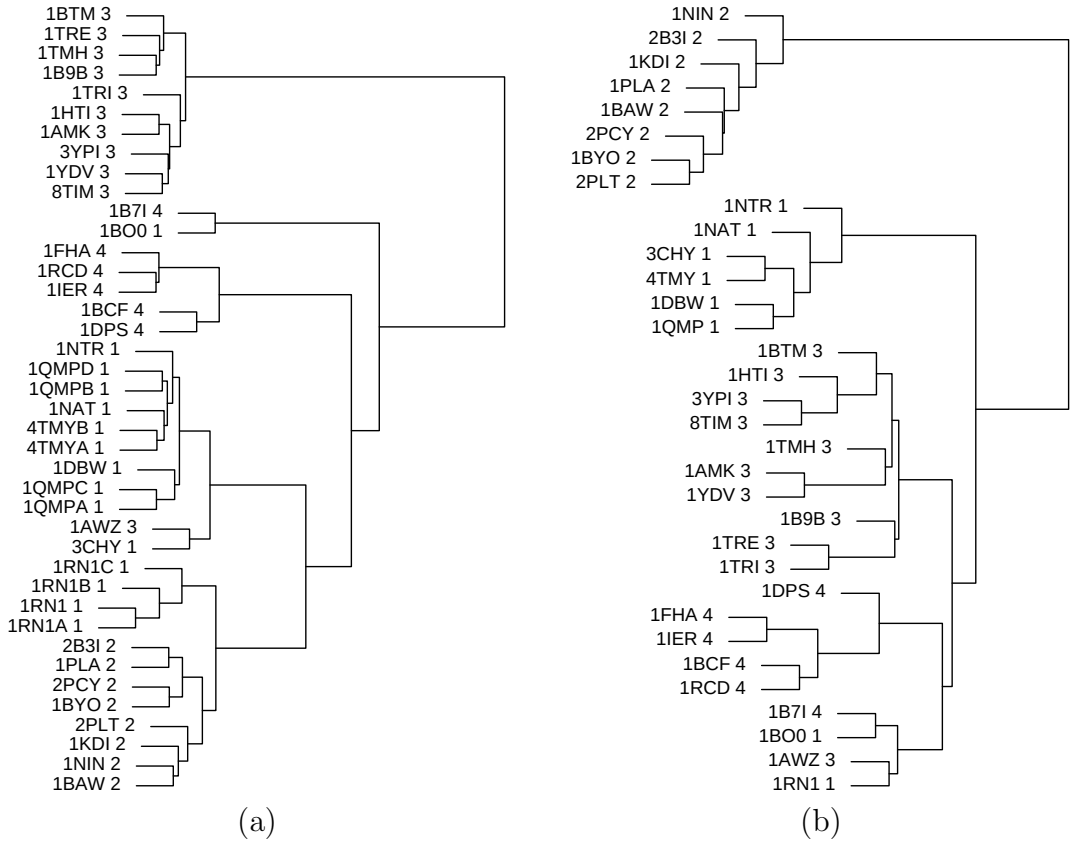


Fig. 12. Comparison of the clusters obtained with the Universal Similarity Metric of [20] (a) and GMAX-FCMO (b) for the Skolnick data set.

7 Conclusions

The improvement of modelling and algorithmic techniques for the comparison of protein structures is a worth-while exercise. The correct assessment of protein structure similarities could have impact on a large number of proteomic activities.

In this paper we propose, first, a generalization of contact maps and we define the *Generalized Fuzzy Contact Map*. In this generalized version one (or more) *fuzzy* thresholds and one (or more) *membership functions* are used to specify the cut-off distances needed to compute the map and also the meaning of contact. This contribution by itself is a step forward in our modelling capabilities as it allows the biologist to include as little or as much domain knowledge as he/she may want.

Next we extended the Maximum Contact Map Overlap Problem (MAX-CMO) by means of fuzzy sets and systems. Our extended setting, *Generalized Maximum Fuzzy Contact Map Overlap Problem* (GMAX-FCMO), allows for a more *biological* formulation of the optimisation problem which is ultimately

used to compute a normalized similarity measure (the original MAX-CMO is not normalized). We also discussed the advantages and limitations of our new models.

We also show that a simple and efficient fuzzy sets based metaheuristic (*FANS*) can be used to solve GMAX-FCMO.

The paper's last contribution was to show how (using GMAX-FCMO and *FANS* through a step-by-step protocol) to correctly measure the similarity between proteins of well known data sets. In turn, these similarity matrices induce correct clusterings of the protein structures.

7.1 Future Work

One of the main avenues for future work would be to investigate whether the approach described by [5], which was originally defined for the standard crisp MAX-CMO model, could be extended to the GMAX-FCMO model. Although there is strong evidence that would sugest that this model is also NP-hard, its formal complexity needs to be evaluated. We are currently running a larger set of experiments to try to assess any weaknesses in the method which could have gone undetected. We are implementing a public web-server with all the tools described in this paper. The software used to compute the similarity in Eq. 5 is available from the authors.

References

[1] P. Artimiuk, A. Poirrette, D. Rice, and P. Willett. The use of graph theoretical methods for the comparison of the structure of biological macromolecules. *Topics of Current Chemistry*, 174:73–103, 1995.

[2] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissing, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[3] A. Blanco, D. Pelta, and J. Verdegay. A fuzzy valuation-based local search framework for combinatorial problems. *Journal of Fuzzy Optimization and Decision Making*, 1(2):177–193, 2002.

[4] P. Bourne and I. Shindyallov. *Structural Bioinformatics*, chapter Structure Comparison and Alignment. Wiley-Liss, Inc, 2003.

[5] 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 overlpa. *Journal of Computational Biology*, 11(1):27–52, 2004.

[6] A. Caprara and G. Lancia. Structural alignment of large-size proteins via lagrangian relaxation. In *Proceedings of RECOMB 2002*. ACM, 2002.

[7] 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.

[8] H. Chan and K. Dill. Origins of structure in globular proteins. *Proc. National Academy of Science,USA*, 97:6388–6392, 1990.

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

[10] D. Goldman. Phd thesis. *Department of Computer Sciences, UC Berkeley*, 2000.

[11] D. Goldman, S. Istrail, and C. Papadimitriou. Algorithmic aspects of protein structure similarity. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, pages 512–522, 1999.

[12] P. Hansen and N. Mladenovic. Variable neighborhood search: Principles and applications. *European Journal of Operational Research*, (130):449–467, 2001.

[13] L. Holm and C. Sander. Protein structure comparison by alignament of distance matrices. *Journal of Molecular Biology*, pages 123–138, 1993.

[14] L. Holm and C. Sander. Mapping the protein universe. *Science*, 273:595–602, 1996.

[15] P. Koehl. Protein structure similarities. *Current Opinion in Structural Biology*, 11:348–353, 2001.

[16] N. Krasnogor. *Studies on the Theory and Design Space of Memetic Algorithms*. Ph.D. Thesis, Faculty of Computing, Mathematics and Engineering, University of the West of England, Bristol, United Kingdom., 2002.

[17] N. Krasnogor. Self-generating metaheuristics in bioinformatics: The proteins structure comparison case. *Genetic Programming and Evolvable Machines*, 5(2), 2004.

[18] N. Krasnogor and D. Pelta. Fuzzy memes in multimeme algorithms: a fuzzy-evolutionary hybrid. In J. Verdegay, editor, *Book chapter in "Fuzzy Sets based Heuristics for Optimization"*. Springer, 2002.

[19] N. Krasnogor and D. Pelta. *Fuzzy Sets based Heuristics for Optimization*, chapter Fuzzy Memes in Multimeme Algorithms: a Fuzzy-Evolutionary Hybrid. Studies in Fuzziness and Soft Computing. Physica-Verlag, 2003. to appear.

[20] N. Krasnogor and D. Pelta. Measuring the similarity of protein structures by means of the universal similarity metric. *Bioinformatics*, (7):1015–1021, 2004.

[21] G. Lancia, R. Carr, B. Walenz, and S. Istrail. 101 optimal pdb structure alignments: a branch-and-cut algorithm for the maximum contact map overlap problem. *Proceedings of The Fifth Annual International Conference on Computational Molecular Biology, RECOMB 2001*, 2001.

[22] R. A. Laskowski. *Structural Bioinformatics*, chapter Structural Quality Assurance. Wiley-Liss, Inc, 2003.

[23] J. Leluk, L. Konieczny, and I. Roterman. Search for structural similarity in proteins. *Bioinformatics*, 19(1):117–124, 2003.

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

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

[26] D. Pelta, A. Blanco, and J. L. Verdegay. *Fuzzy Sets based Heuristics for Optimization*, chapter Fuzzy Adaptive Neighborhood Search: Examples of Application. Studies in Fuzziness and Soft Computing. Physica-Verlag, 2003. to appear.

[27] W. Taylor. Protein structure comparison using iterated double dynamic programming. *Protein Science*, 8:654–665, 1999.

[28] T. Wu, S. Schmidler, T. Hastie, and D. Brutlag. Regresion analysis of multiple protein structures. *Journal of Computational Biology*, 5:585–595, 1998.

	1AA9	1GNP	6Q21	1QRA	5P21	1CD8	1NEU	1QA9	1CDB	1CI5	1QFO	1ECA	2LHB	5MBN	2VHB	1CNP
1AA9	1.0000	0.2844	0.5828	0.5513	0.2826	0.0728	0.0627	0.0729	0.0626	0.0691	0.0809	0.1149	0.1202	0.1064	0.1131	0.0826
1GNP	0.2844	1.0000	0.7662	0.7141	0.8944	0.0871	0.0683	0.0702	0.0702	0.0751	0.0862	0.1265	0.1062	0.1322	0.1153	0.0740
6Q21	0.5828	0.7662	1.0000	0.5961	0.8254	0.0742	0.0716	0.0815	0.0790	0.0722	0.0909	0.1331	0.0970	0.1313	0.1246	0.1085
1QRA	0.5513	0.7141	0.5961	1.0000	0.3013	0.0980	0.0642	0.0704	0.0797	0.0766	0.0927	0.1114	0.1397	0.1221	0.0942	0.0821
5P21	0.2826	0.8944	0.8254	0.3013	1.0000	0.0831	0.0769	0.0828	0.0725	0.0711	0.1119	0.1039	0.0973	0.1406	0.0876	0.0848
1CD8	0.0728	0.0871	0.0742	0.0980	0.0831	1.0000	0.4449	0.3824	0.3197	0.2804	0.5673	0.1120	0.1038	0.0966	0.1249	0.1056
1NEU	0.0627	0.0683	0.0716	0.0642	0.0769	0.4449	1.0000	0.5287	0.4445	0.5470	0.5893	0.1026	0.0959	0.0913	0.0996	0.0946
1QA9	0.0729	0.0702	0.0815	0.0704	0.0828	0.3824	0.5287	1.0000	0.5451	0.4230	0.5237	0.1017	0.1141	0.1023	0.1177	0.1176
1CDB	0.0626	0.0702	0.0790	0.0797	0.0725	0.3197	0.4445	0.5451	1.0000	0.4141	0.3006	0.0985	0.0920	0.0939	0.1041	0.1160
1CI5	0.0691	0.0751	0.0722	0.0766	0.0711	0.2804	0.5470	0.4230	0.4141	1.0000	0.3780	0.1172	0.0916	0.0949	0.0986	0.1159
1QFO	0.0809	0.0862	0.0909	0.0927	0.1119	0.5673	0.5893	0.5237	0.3006	0.3780	1.0000	0.1138	0.1136	0.0962	0.1175	0.1140
1ECA	0.1149	0.1265	0.1331	0.1114	0.1039	0.1120	0.1026	0.1017	0.0985	0.1172	0.1138	1.0000	0.2845	0.3076	0.2958	0.1385
2LHB	0.1202	0.1062	0.0970	0.1397	0.0973	0.1038	0.0959	0.1141	0.0920	0.0916	0.1136	0.2845	1.0000	0.3367	0.4086	0.1287
5MBN	0.1064	0.1322	0.1313	0.1221	0.1406	0.0966	0.0913	0.1023	0.0939	0.0949	0.0962	0.3076	0.3367	1.0000	0.4401	0.1256
2VHB	0.1131	0.1153	0.1246	0.0942	0.0876	0.1249	0.0996	0.1177	0.1041	0.0986	0.1175	0.2958	0.4086	0.4401	1.0000	0.1270
1CNP	0.0826	0.0740	0.1085	0.0821	0.0848	0.1056	0.0946	0.1176	0.1160	0.1159	0.1140	0.1385	0.1287	0.1256	0.1270	1.0000
1JHG	0.0950	0.0945	0.0903	0.0970	0.0876	0.0923	0.0901	0.1325	0.1214	0.1092	0.1070	0.1572	0.1525	0.1698	0.1966	0.2795
1CHR	0.0394	0.0590	0.0522	0.0476	0.0641	0.0298	0.0303	0.0276	0.0251	0.0275	0.0319	0.0475	0.0475	0.0470	0.0352	0.0286
2MNR	0.0538	0.0519	0.0533	0.0508	0.0450	0.0300	0.0299	0.0329	0.0271	0.0311	0.0340	0.0520	0.0507	0.0600	0.0510	0.0295
4ENL	0.0444	0.0473	0.0416	0.0341	0.0323	0.0242	0.0206	0.0205	0.0190	0.0237	0.0238	0.0350	0.0496	0.0449	0.0451	0.0193
6XIA	0.0603	0.0458	0.0631	0.0734	0.0635	0.0375	0.0307	0.0296	0.0316	0.0280	0.0414	0.0533	0.0731	0.0570	0.0669	0.0321
1ASH	0.1195	0.1241	0.1188	0.1462	0.1237	0.0994	0.1039	0.1079	0.0945	0.1039	0.1069	0.3327	0.2465	0.2127	0.1968	0.1371
1BAB	0.0948	0.1094	0.1142	0.1385	0.1252	0.0973	0.1038	0.1001	0.0826	0.1041	0.1162	0.2851	0.2271	0.5545	0.3222	0.1398
1FLP	0.1067	0.1141	0.1786	0.1685	0.1308	0.1103	0.0911	0.1163	0.0975	0.1145	0.1006	0.3071	0.3430	0.2607	0.2985	0.1555
1HLB	0.1009	0.1427	0.1213	0.1077	0.1279	0.1028	0.0952	0.1113	0.0966	0.0999	0.1231	0.3313	0.3650	0.1962	0.2559	0.1088
1HLM	0.1077	0.1518	0.1141	0.1350	0.1752	0.0922	0.1027	0.1107	0.1036	0.1441	0.1256	0.1905	0.2990	0.2451	0.2136	0.1326
1ITH	0.1430	0.1046	0.1291	0.1300	0.1476	0.1137	0.1002	0.1003	0.1013	0.1066	0.1224	0.3634	0.5222	0.5036	0.2903	0.1285
1LH2	0.1231	0.1219	0.1530	0.1200	0.1161	0.0968	0.0980	0.0953	0.0863	0.0955	0.0967	0.2801	0.3127	0.2946	0.2541	0.1276
1MBA	0.1117	0.1332	0.1182	0.1213	0.1617	0.0946	0.0910	0.0905	0.0857	0.1115	0.1067	0.3275	0.4084	0.3678	0.2229	0.1527
1MYT	0.1120	0.1198	0.1174	0.1501	0.1284	0.1053	0.0879	0.0972	0.0999	0.0925	0.1158	0.3282	0.1930	0.2887	0.2623	0.1268
2HBG	0.1293	0.1157	0.1154	0.1044	0.1240	0.1004	0.0989	0.1019	0.1043	0.0848	0.1061	0.3068	0.3590	0.3354	0.2368	0.1120
3SDH	0.0947	0.1366	0.1169	0.1131	0.0961	0.1116	0.1040	0.1180	0.0875	0.0986	0.1085	0.2414	0.2475	0.2577	0.2667	0.1410

Table 1. Maximum Similarities(I)

	1JHG	1CHR	2MNR	4ENL	6XIA	1ASH	1BAB	1FLP	1HLB	1HLM	1ITH	1LH2	1MBA	1MYT	2HBG	3SDH
1AA9	0.0950	0.0394	0.0538	0.0444	0.0603	0.1195	0.0948	0.1067	0.1009	0.1077	0.1430	0.1231	0.1117	0.1120	0.1293	0.0947
1GNP	0.0945	0.0590	0.0519	0.0473	0.0458	0.1241	0.1094	0.1141	0.1427	0.1518	0.1046	0.1219	0.1332	0.1198	0.1157	0.1366
6Q21	0.0903	0.0522	0.0533	0.0416	0.0631	0.1188	0.1142	0.1786	0.1213	0.1141	0.1291	0.1530	0.1182	0.1174	0.1154	0.1169
1QRA	0.0970	0.0476	0.0508	0.0341	0.0734	0.1462	0.1385	0.1685	0.1077	0.1350	0.1300	0.1200	0.1213	0.1501	0.1044	0.1131
5P21	0.0876	0.0641	0.0450	0.0323	0.0635	0.1237	0.1252	0.1308	0.1279	0.1752	0.1476	0.1161	0.1617	0.1284	0.1240	0.0961
1CD8	0.0923	0.0298	0.0300	0.0242	0.0375	0.0994	0.0973	0.1103	0.1028	0.0922	0.1137	0.0968	0.0946	0.1053	0.1004	0.1116
1NEU	0.0901	0.0303	0.0299	0.0206	0.0307	0.1039	0.1038	0.0911	0.0952	0.1027	0.1002	0.0980	0.0910	0.0879	0.0989	0.1040
1QA9	0.1325	0.0276	0.0329	0.0205	0.0296	0.1079	0.1001	0.1163	0.1113	0.1107	0.1003	0.0953	0.0905	0.0972	0.1019	0.1180
1CDB	0.1214	0.0251	0.0271	0.0190	0.0316	0.0945	0.0826	0.0975	0.0966	0.1036	0.1013	0.0863	0.0857	0.0999	0.1043	0.0875
1CI5	0.1092	0.0275	0.0311	0.0237	0.0280	0.1039	0.1041	0.1145	0.0999	0.1441	0.1066	0.0955	0.1115	0.0925	0.0848	0.0986
1QFO	0.1070	0.0319	0.0340	0.0238	0.0414	0.1069	0.1162	0.1006	0.1231	0.1256	0.1224	0.0967	0.1067	0.1158	0.1061	0.1085
1ECA	0.1572	0.0475	0.0520	0.0350	0.0533	0.3327	0.2851	0.3071	0.3313	0.1905	0.3634	0.2801	0.3275	0.3282	0.3068	0.2414
2LHB	0.1525	0.0475	0.0507	0.0496	0.0731	0.2465	0.2271	0.3430	0.3650	0.2990	0.5222	0.3127	0.4084	0.1930	0.3590	0.2475
5MBN	0.1698	0.0470	0.0600	0.0449	0.0570	0.2127	0.5545	0.2607	0.1962	0.2451	0.5036	0.2946	0.3678	0.2887	0.3354	0.2577
2VHB	0.1966	0.0352	0.0510	0.0451	0.0669	0.1968	0.3222	0.2985	0.2559	0.2136	0.2903	0.2541	0.2229	0.2623	0.2368	0.2667
1CNP	0.2795	0.0286	0.0295	0.0193	0.0321	0.1371	0.1398	0.1555	0.1088	0.1326	0.1285	0.1276	0.1527	0.1268	0.1120	0.1410
1JHG	1.0000	0.0281	0.0386	0.0263	0.0573	0.1439	0.1672	0.1704	0.1715	0.1447	0.1536	0.1496	0.1694	0.1963	0.1625	0.1444
1CHR	0.0281	1.0000	0.0573	0.0537	0.0629	0.0438	0.0402	0.0525	0.0627	0.0439	0.0526	0.0422	0.0465	0.0534	0.0662	0.0492
2MNR	0.0386	0.0573	1.0000	0.0607	0.0668	0.0448	0.0613	0.0585	0.0480	0.0466	0.0427	0.0574	0.0464	0.0537	0.0466	0.0668
4ENL	0.0263	0.0537	0.0607	1.0000	0.0457	0.0427	0.0458	0.0378	0.0397	0.0463	0.0373	0.0353	0.0457	0.0305	0.0525	0.0492
6XIA	0.0573	0.0629	0.0668	0.0457	1.0000	0.0521	0.0564	0.0629	0.0615	0.0666	0.0532	0.0749	0.0484	0.0706	0.0517	0.0592
1ASH	0.1439	0.0438	0.0448	0.0427	0.0521	1.0000	0.3050	0.3096	0.1999	0.2765	0.3151	0.2469	0.3252	0.3124	0.3796	0.3588
1BAB	0.1672	0.0402	0.0613	0.0458	0.0564	0.3050	1.0000	0.2338	0.2635	0.2932	0.2706	0.2442	0.3454	0.1965	0.2857	0.3773
1FLP	0.1704	0.0525	0.0585	0.0378	0.0629	0.3096	0.2338	1.0000	0.2607	0.2994	0.3302	0.2270	0.2899	0.3074	0.2646	0.4849
1HLB	0.1715	0.0627	0.0480	0.0397	0.0615	0.1999	0.2635	0.2607	1.0000	0.3366	0.2685	0.1833	0.3322	0.2585	0.2070	0.3232
1HLM	0.1447	0.0439	0.0466	0.0463	0.0666	0.2765	0.2932	0.2994	0.3366	1.0000	0.2769	0.3061	0.3087	0.2734	0.3042	0.3577
1ITH	0.1536	0.0526	0.0427	0.0373	0.0532	0.3151	0.2706	0.3302	0.2685	0.2769	1.0000	0.3034	0.2985	0.2633	0.4733	0.2731
1LH2	0.1496	0.0422	0.0574	0.0353	0.0749	0.2469	0.2442	0.2270	0.1833	0.3061	0.3034	1.0000	0.2719	0.4323	0.2126	0.2090
1MBA	0.1694	0.0465	0.0464	0.0457	0.0484	0.3252	0.3454	0.2899	0.3322	0.3087	0.2985	0.2719	1.0000	0.3554	0.2972	0.2616
1MYT	0.1963	0.0534	0.0537	0.0305	0.0706	0.3124	0.1965	0.3074	0.2585	0.2734	0.2633	0.4323	0.3554	1.0000	0.4029	0.2623
2HBG	0.1625	0.0662	0.0466	0.0525	0.0517	0.3796	0.2857	0.2646	0.2070	0.3042	0.4733	0.2126	0.2972	0.4029	1.0000	0.2156
3SDH	0.1444	0.0492	0.0668	0.0492	0.0592	0.3588	0.3773	0.4849	0.3232	0.3577	0.2731	0.2090	0.2616	0.2623	0.2156	1.0000

Table 2. Maximum Similarities(II)

	1AA9	1GNP	6Q21	1QRA	5P21	1CD8	1NEU	1QA9	1CDB	1CI5	1QFO	1ECA	2LHB	5MBN	2VHB	1CNP
1AA9	1.0000	0.1215	0.0913	0.2688	0.1736	0.0656	0.0511	0.0600	0.0562	0.0594	0.0620	0.0884	0.0867	0.0952	0.0873	0.0704
1GNP	0.1215	1.0000	0.2111	0.5110	0.5490	0.0706	0.0597	0.0639	0.0576	0.0600	0.0730	0.1156	0.0913	0.0693	0.0916	0.0536
6Q21	0.0913	0.2111	1.0000	0.1789	0.1518	0.0718	0.0638	0.0667	0.0674	0.0674	0.0577	0.0879	0.0865	0.0921	0.1092	0.0626
1QRA	0.2688	0.5110	0.1789	1.0000	0.1373	0.0710	0.0558	0.0572	0.0631	0.0625	0.0733	0.0877	0.1191	0.0888	0.0904	0.0694
5P21	0.1736	0.5490	0.1518	0.1373	1.0000	0.0584	0.0580	0.0609	0.0649	0.0626	0.0670	0.0959	0.0823	0.1110	0.0716	0.0718
1CD8	0.0656	0.0706	0.0718	0.0710	0.0584	1.0000	0.3219	0.1617	0.1682	0.0965	0.2172	0.0938	0.0781	0.0796	0.0888	0.0970
1NEU	0.0511	0.0597	0.0638	0.0558	0.0580	0.3219	1.0000	0.3563	0.2821	0.2380	0.2864	0.0944	0.0820	0.0687	0.0691	0.0843
1QA9	0.0600	0.0639	0.0667	0.0572	0.0609	0.1617	0.3563	1.0000	0.3717	0.2623	0.1495	0.0956	0.0929	0.0886	0.0896	0.1097
1CDB	0.0562	0.0576	0.0674	0.0631	0.0649	0.1682	0.2821	0.3717	1.0000	0.3680	0.1106	0.0943	0.0866	0.0847	0.0971	0.1027
1CI5	0.0594	0.0600	0.0674	0.0625	0.0626	0.0965	0.2380	0.2623	0.3680	1.0000	0.1784	0.0964	0.0856	0.0732	0.0814	0.1017
1QFO	0.0620	0.0730	0.0577	0.0733	0.0670	0.2172	0.2864	0.1495	0.1106	0.1784	1.0000	0.0948	0.0780	0.0794	0.1059	0.0979
1ECA	0.0884	0.1156	0.0879	0.0877	0.0959	0.0938	0.0944	0.0956	0.0943	0.0964	0.0948	1.0000	0.2051	0.2193	0.2357	0.1302
2LHB	0.0867	0.0913	0.0865	0.1191	0.0823	0.0781	0.0820	0.0929	0.0866	0.0856	0.0780	0.2051	1.0000	0.2073	0.2229	0.1042
5MBN	0.0952	0.0693	0.0921	0.0888	0.1110	0.0796	0.0687	0.0886	0.0847	0.0732	0.0794	0.2193	0.2073	1.0000	0.2371	0.1205
2VHB	0.0873	0.0916	0.1092	0.0904	0.0716	0.0888	0.0691	0.0896	0.0971	0.0814	0.1059	0.2357	0.2229	0.2371	1.0000	0.1235
1CNP	0.0704	0.0536	0.0626	0.0694	0.0718	0.0970	0.0843	0.1097	0.1027	0.1017	0.0979	0.1302	0.1042	0.1205	0.1235	1.0000
1JHG	0.0805	0.0773	0.0829	0.0773	0.0681	0.0861	0.0861	0.0986	0.1065	0.0970	0.0813	0.1556	0.1113	0.1402	0.1370	0.1922
1CHR	0.0348	0.0356	0.0338	0.0336	0.0418	0.0247	0.0250	0.0250	0.0237	0.0262	0.0294	0.0418	0.0331	0.0328	0.0317	0.0219
2MNR	0.0514	0.0388	0.0345	0.0465	0.0363	0.0245	0.0237	0.0267	0.0238	0.0256	0.0241	0.0412	0.0423	0.0515	0.0407	0.0269
4ENL	0.0293	0.0258	0.0275	0.0278	0.0282	0.0186	0.0185	0.0162	0.0166	0.0158	0.0202	0.0271	0.0309	0.0359	0.0338	0.0180
6XIA	0.0474	0.0398	0.0420	0.0378	0.0375	0.0289	0.0242	0.0279	0.0258	0.0243	0.0246	0.0333	0.0560	0.0397	0.0500	0.0287
1ASH	0.0975	0.1028	0.0957	0.1102	0.1145	0.0705	0.0849	0.0843	0.0829	0.0762	0.0951	0.1787	0.1787	0.1596	0.1701	0.0927
1BAB	0.0748	0.1018	0.0839	0.1008	0.1162	0.0828	0.0834	0.0779	0.0786	0.0710	0.0874	0.1318	0.1787	0.2102	0.2555	0.1191
1FLP	0.0836	0.0945	0.0829	0.0936	0.0846	0.1080	0.0710	0.0943	0.0760	0.0787	0.0871	0.2224	0.2125	0.1636	0.2004	0.1034
1HLB	0.0959	0.1103	0.1007	0.0760	0.0939	0.0985	0.0716	0.0973	0.0866	0.0769	0.0908	0.1499	0.1736	0.1669	0.1399	0.0903
1HLM	0.0770	0.0872	0.0959	0.1179	0.0899	0.0821	0.0849	0.0762	0.0759	0.0795	0.0946	0.1247	0.2382	0.1637	0.1718	0.1178
1ITH	0.0818	0.0989	0.1080	0.0987	0.0768	0.0942	0.0746	0.0853	0.0883	0.0846	0.0856	0.1935	0.1707	0.1633	0.1905	0.0980
1LH2	0.0997	0.0931	0.1244	0.0848	0.0935	0.0697	0.0638	0.0817	0.0747	0.0651	0.0905	0.1776	0.1419	0.1518	0.1411	0.1126
1MBA	0.0870	0.0821	0.0862	0.0899	0.0902	0.0849	0.0771	0.0798	0.0651	0.0794	0.0828	0.2247	0.1265	0.2935	0.1768	0.1232
1MYT	0.0785	0.0781	0.1047	0.1199	0.0855	0.0780	0.0814	0.0904	0.0873	0.0842	0.1066	0.1872	0.1486	0.1781	0.1691	0.1126
2HBG	0.0972	0.0847	0.1030	0.0852	0.1004	0.0730	0.0797	0.0924	0.0898	0.0704	0.0894	0.1774	0.1453	0.1994	0.1435	0.0991
3SDH	0.0875	0.0869	0.0845	0.0947	0.0766	0.0876	0.0885	0.0852	0.0753	0.0833	0.0936	0.1748	0.1923	0.1646	0.1973	0.1115

Table 3. Minimum Similarities (I)

	1JHG	1CHR	2MNR	4ENL	6XIA	1ASH	1BAB	1FLP	1HLB	1HLM	1ITH	1LH2	1MBA	1MYT	2HBG	3SDH
1AA9	0.0805	0.0348	0.0514	0.0293	0.0474	0.0975	0.0748	0.0836	0.0959	0.0770	0.0818	0.0997	0.0870	0.0785	0.0972	0.0875
1GNP	0.0773	0.0356	0.0388	0.0258	0.0398	0.1028	0.1018	0.0945	0.1103	0.0872	0.0989	0.0931	0.0821	0.0781	0.0847	0.0869
6Q21	0.0829	0.0338	0.0345	0.0275	0.0420	0.0957	0.0839	0.0829	0.1007	0.0959	0.1080	0.1244	0.0862	0.1047	0.1030	0.0845
1QRA	0.0773	0.0336	0.0465	0.0278	0.0378	0.1102	0.1008	0.0936	0.0760	0.1179	0.0987	0.0848	0.0899	0.1199	0.0852	0.0947
5P21	0.0681	0.0418	0.0363	0.0282	0.0375	0.1145	0.1162	0.0846	0.0939	0.0899	0.0768	0.0935	0.0902	0.0855	0.1004	0.0766
1CD8	0.0861	0.0247	0.0245	0.0186	0.0289	0.0705	0.0828	0.1080	0.0985	0.0821	0.0942	0.0697	0.0849	0.0780	0.0730	0.0876
1NEU	0.0861	0.0250	0.0237	0.0185	0.0242	0.0849	0.0834	0.0710	0.0716	0.0849	0.0746	0.0638	0.0771	0.0814	0.0797	0.0885
1QA9	0.0986	0.0250	0.0267	0.0162	0.0279	0.0843	0.0779	0.0943	0.0973	0.0762	0.0853	0.0817	0.0798	0.0904	0.0924	0.0852
1CDB	0.1065	0.0237	0.0238	0.0166	0.0258	0.0829	0.0786	0.0760	0.0866	0.0759	0.0883	0.0747	0.0651	0.0873	0.0898	0.0753
1CI5	0.0970	0.0262	0.0256	0.0158	0.0243	0.0762	0.0710	0.0787	0.0769	0.0795	0.0846	0.0651	0.0794	0.0842	0.0704	0.0833
1QFO	0.0813	0.0294	0.0241	0.0202	0.0246	0.0951	0.0874	0.0871	0.0908	0.0946	0.0856	0.0905	0.0828	0.1066	0.0894	0.0936
1ECA	0.1556	0.0418	0.0412	0.0271	0.0333	0.1787	0.1318	0.2224	0.1499	0.1247	0.1935	0.1776	0.2247	0.1872	0.1774	0.1748
2LHB	0.1113	0.0331	0.0423	0.0309	0.0560	0.1787	0.1787	0.2125	0.1736	0.2382	0.1707	0.1419	0.1265	0.1486	0.1453	0.1923
5MBN	0.1402	0.0328	0.0515	0.0359	0.0397	0.1596	0.2102	0.1636	0.1669	0.1637	0.1633	0.1518	0.2935	0.1781	0.1994	0.1646
2VHB	0.1370	0.0317	0.0407	0.0338	0.0500	0.1701	0.2555	0.2004	0.1399	0.1718	0.1905	0.1411	0.1768	0.1691	0.1435	0.1973
1CNP	0.1922	0.0219	0.0269	0.0180	0.0287	0.0927	0.1191	0.1034	0.0903	0.1178	0.0980	0.1126	0.1232	0.1126	0.0991	0.1115
1JHG	1.0000	0.0225	0.0246	0.0208	0.0343	0.1205	0.1442	0.1188	0.1146	0.0977	0.1219	0.1013	0.1047	0.1408	0.1130	0.1166
1CHR	0.0225	1.0000	0.0476	0.0374	0.0427	0.0303	0.0310	0.0461	0.0491	0.0385	0.0351	0.0360	0.0357	0.0312	0.0445	0.0356
2MNR	0.0246	0.0476	1.0000	0.0359	0.0549	0.0372	0.0326	0.0353	0.0261	0.0340	0.0296	0.0291	0.0397	0.0320	0.0339	0.0308
4ENL	0.0208	0.0374	0.0359	1.0000	0.0342	0.0298	0.0318	0.0286	0.0312	0.0407	0.0247	0.0313	0.0329	0.0286	0.0277	0.0268
6XIA	0.0343	0.0427	0.0549	0.0342	1.0000	0.0430	0.0447	0.0451	0.0457	0.0298	0.0416	0.0338	0.0464	0.0433	0.0432	0.0411
1ASH	0.1205	0.0303	0.0372	0.0298	0.0430	1.0000	0.1933	0.2483	0.0990	0.1473	0.1803	0.1480	0.1568	0.1917	0.2646	0.2168
1BAB	0.1442	0.0310	0.0326	0.0318	0.0447	0.1933	1.0000	0.1813	0.1801	0.1852	0.1866	0.2203	0.2159	0.1765	0.1921	0.1542
1FLP	0.1188	0.0461	0.0353	0.0286	0.0451	0.2483	0.1813	1.0000	0.1443	0.1986	0.2492	0.1511	0.2185	0.1691	0.2312	0.2614
1HLB	0.1146	0.0491	0.0261	0.0312	0.0457	0.0990	0.1801	0.1443	1.0000	0.1542	0.1783	0.1433	0.1444	0.1755	0.1485	0.2495
1HLM	0.0977	0.0385	0.0340	0.0407	0.0298	0.1473	0.1852	0.1986	0.1542	1.0000	0.2187	0.1890	0.2640	0.1751	0.1667	0.1940
1ITH	0.1219	0.0351	0.0296	0.0247	0.0416	0.1803	0.1866	0.2492	0.1783	0.2187	1.0000	0.2073	0.1755	0.2099	0.1487	0.2128
1LH2	0.1013	0.0360	0.0291	0.0313	0.0338	0.1480	0.2203	0.1511	0.1433	0.1890	0.2073	1.0000	0.1926	0.2656	0.1227	0.1631
1MBA	0.1047	0.0357	0.0397	0.0329	0.0464	0.1568	0.2159	0.2185	0.1444	0.2640	0.1755	0.1926	1.0000	0.2167	0.1871	0.1666
1MYT	0.1408	0.0312	0.0320	0.0286	0.0433	0.1917	0.1765	0.1691	0.1755	0.1751	0.2099	0.2656	0.2167	1.0000	0.2003	0.1837
2HBG	0.1130	0.0445	0.0339	0.0277	0.0432	0.2646	0.1921	0.2312	0.1485	0.1667	0.1487	0.1227	0.1871	0.2003	1.0000	0.1688
3SDH	0.1166	0.0356	0.0308	0.0268	0.0411	0.2168	0.1542	0.2614	0.2495	0.1940	0.2128	0.1631	0.1666	0.1837	0.1688	1.0000

Table 4. Minimum Similarities (II)