An Image-Matching Approach to Protein Similarity Analysis

Fernando Fernandes Jr. Carlos E. R. Lopes Raquel C. de Melo Marcelo M. Santoro
Rodrigo L. Carceroni Wagner Meira Jr. Arnaldo de A. Araújo Carlos H. Silveira

Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627, Belo Horizonte, MG, Brazil
{junior, dlopes, raquelcm, carceron, meira, arnaldo, silveira}@dcc.ufmg.br
santoro@icb.ufmg.br

Abstract

In this work we model the problem of identifying how close structurally two proteins are as a problem of measuring the similarity between color images that represent their contact maps, where the chromatic information encodes the chemical nature of the contacts. We study two conceptually distinct methods to measure the similarity between such contact maps: a content-based image retrieval one and another based on image registration. In experiments with contact maps constructed from the PDB, the image registration approach was able to identify with 100% precision 8 instances of a protein class mixed with 28 proteins of other classes. The content-based image retrieval approach had an accuracy only a little worse than that.

1. Introduction

Bioinformatics is an emerging field undergoing rapid growth. This has mainly been fueled by advances in DNA[1] sequencing and mapping techniques. The Human Genome Project has resulted in an exponentially growing database of genetics sequences, while the Structural Genomics Initiative is doing the same for the Protein Data Bank (PDB)[2]. One of the most active research areas in bioinformatics is protein structure analysis.

Proteins are the most versatile macromolecules in living systems and serve crucial functions in all biological processes. They function as catalysts, transporters and store other molecules such as oxygen, provide mechanical support and immune protection, generate movement, transmit nerve impulses and control growth and differentiation [1].

1 Deoxy-ribonucleic Acid

Proteins are composed by a sequence of amino acids. This sequence is called *primary structure*. Different regions of the sequence form regular *secondary structures* such as α-helices or β-strands. The *terciary structure*, which is the three-dimensional structure of the protein, is formed by packing such structural elements into one or several compact globular units called domains.

The functional properties of proteins depend upon their three-dimensional structures. These structures arise because a particular sequence of amino acids folds to generate, from linear chain, domains with specific three-dimensional structure. It is known that the chain of amino acids completely determines the structure of a protein. However, there can be many proteins with the same structure, and the same function, but with very different sequences. Hence, the protein structure study is important because it determines protein function.

In Figure 1, we present four proteins of three different structural motifs. *Myoglobin*, the oxygen carrier in the muscles, is an α-protein, composed mostly by α-helices. *Troponin-C*, a calcium-binding from cardiac muscle, is an $\alpha\beta$-protein, a mixture of all kinds of secondary structures. *Plastocyanin*, an electron carrier, is a β-protein composed mostly by β-sheets.

By april 1^{st} 2004, the PDB had on its archives approximately 25.000 proteins, and this number has been increasing year after year. Just on the first four months of 2004, 1.521 proteins were added to that database. Even though the PDB is now a huge data set, significant knowledge still has to be extracted.

In this work we use an image-matching approach to analyze the similarity of proteins. We apply this approach to an image database that contains information about the chemical interactions in proteins. Each image is the contact map of a protein, where each color represents one type of chemical interaction between two amino acids. Hence, these con-

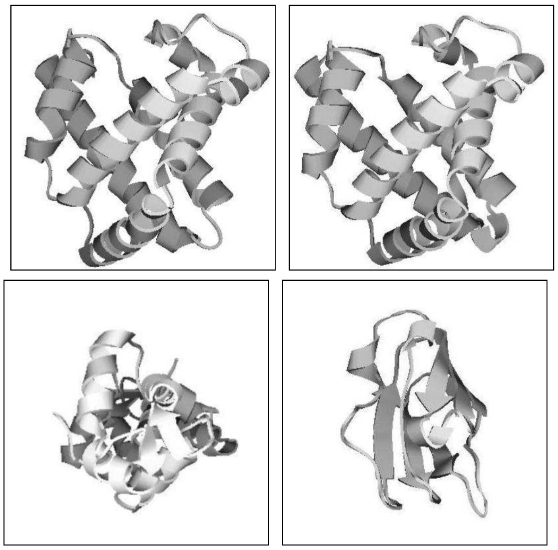

Figure 1. Proteins with different structural motifs. The top ones are α-proteins, more specificaly the, Human and Sea Hare *Myoglobins*. The lower leftmost is the *Troponin-C*, an $\alpha\beta$-protein. The lower rightmost one is a β-protein, the *Plastocyanin*.

tact maps are a two-dimensional representation of the protein structure. Contect-based image retrieval (CBIR) and image registration techniques are used to measure the similarity between contact maps making possible to measure the similarity between protein structures.

In fact, a chain folds in a three-dimensional structure because of chemical interactions between its amino acids. These interactions are also indispensable for the action of the proteins. In the enzymes, for example, they are responsible for the binding and unbinding of the substrate, and they are involved in the catalysis as well. Thus, it is very interesting to study proteins similarity based on their internal chemical interactions.

The three most important kinds of interactions are *hydrophobic*, *electrostatic* and *hydrogen bonds*:

- The *hydrophobic interactions* consist of the attraction between hydrophobic amino acids because of their water aversion. This makes water-soluble proteins fold in a hydrophobic core and a hydrophilic surface.
- The *electrostatic interactions* are the attraction between amino acids with different charges. Most of them are located at the surface of the proteins.
- The *hydrogen bonds* are strong, short-distance interactions between amino acids that share a common hydrogen atom. They are very important to stabilize the α-helices and the β-strands, structures that are important to keep the fold of the protein [4].

In this paper, we used the spatial location of these three types of interactions as the features from which protein structure is identified.

The remaing of this document is organized as follows: Section 2 characterizes the problem that we are dealing with; Section 3 presents related works on protein structure analysis; Section 4 provides an overview of the proposed approach; in Section 5 we explain about the creation of the image database; and Sections 6 and 7 show some experimental results and concluding remarks, respectively.

2. Problem characterization

Contact maps are useful tools to study protein structural motifs. Each contact map consists of a colored image that represents all chemical interactions (Electrostatic, Hydrophobic and Hydrogen Bonds in this work) between the amino acids of a protein. As each protein has a unique contact map, these maps are a good representation of protein structure. Our database is composed by contact map images with $n \times n$ pixels, where n is the number of amino acids of the protein sequence. We define the color of each pixel $[i, j]$ as: **white** if there is no interaction between amino acids i and j; **blue** if there is a hydrophobic interaction; **green** if there is an electrostatic interaction and **red** if there is a hydrogen bond.

These maps have a restriction. Two amino acids can make more than one kind of interaction, which can not be expressed in those two dimensional images. So, we decided to prioritize the electrostatic interactions then, the hydrogen bonds and finally, the hydrophobic interactions. This is the increasing frequency order of these types of contact. Because of the high frequency of the hydrophobic interactions, we believe that ignoring a few of them will not affect the general distribution of these contacts.

In Figure 2, we present plots of protein chemical interactions as they are in a contact map image that compose our database. These graphics came from the proteins listed in Figure 1. We present plots instead of the original contact map images to better visualization.

With the increasing growing rates of the public protein databases, as observed in the PDB, retrieving relevant proteins from a given query is becoming a difficult task. Current database search is performed by text query on protein IDs or names.

Using contact maps, as visual representation of proteins, allow us to use an image-matching methodology (described in Section 4) in order to analyse structural similarity between some of them. Hence, we expect to, given a specific protein, select the maps of proteins that have similar interaction patterns and thus similar structure.

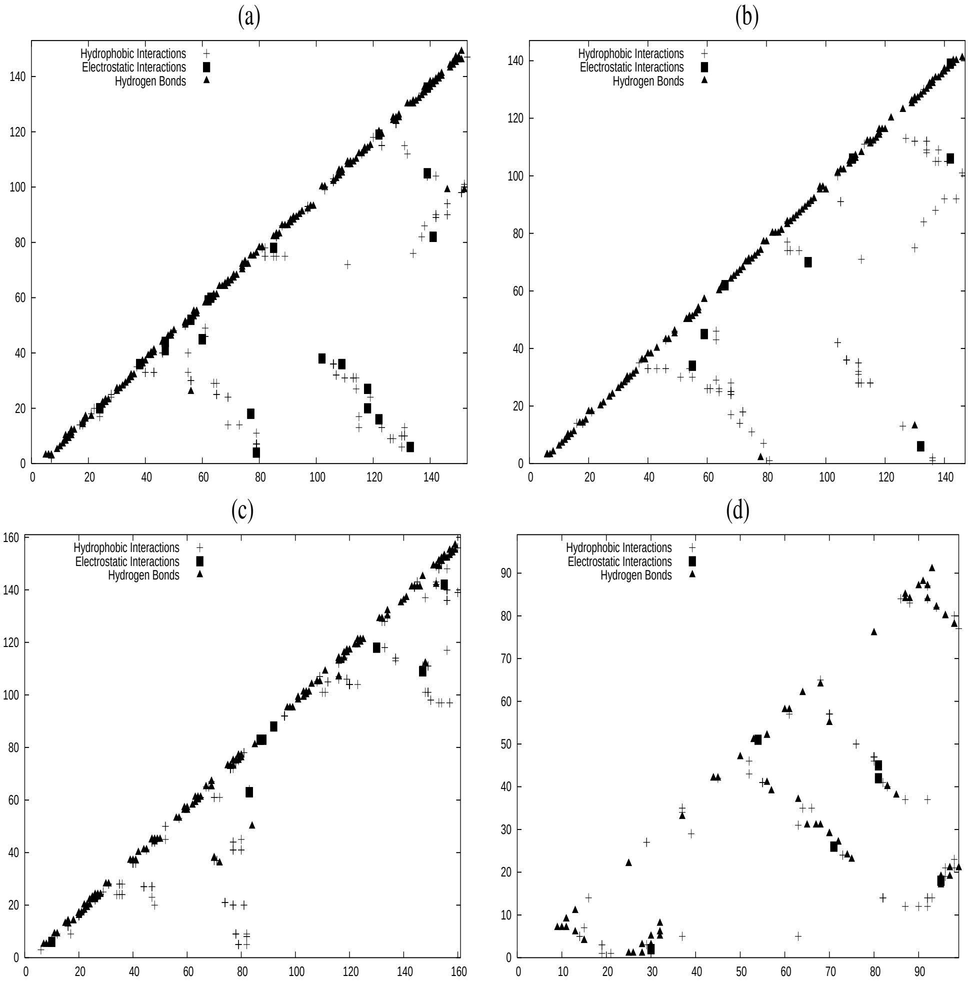

Figure 2. Examples of proteins contact maps: (a) Human *Myoglobin*, (b) Sea Hare *Myoglobin*, (c) *Troponin-C* and (d) *Plastocyanin*

3. Related works

Protein topology has been a topic of great interest in the last years. TOPS [19] is one of the best web sites destinated for protein topology studies. It presents a protein topology atlas composed by cartoons representing the topologies of proteins. These are graphical forms of two-dimensional schematic diagrams which represent a fold as a sequence of secondary structures and their relative orientation and spatial position. TOPS also performs protein classification based on pattern searching using a string-graph algorithm.

Another method presented for comparison of protein structures is TOPSCAN [18]. The authors used secondary structures and their relative direction, proximity, accessibility and length. This method uses the *Needleman* and *Wunsch* dynamic programming algorithm [12].

The CATH [14] database is a hierarchical domain classification of proteins. Structures are grouped into fold families depending on the shape and connectivity of the secondary structures. This is done using the structure comparison algorithm SSAP [15]. Parameters for clustering domains into fold families were determined by empirical trials throughout the database.

SCOP [11] is a web site of hierarchy of proteins. It was created by manual inspection and abetted by automated methods, aiming to provide a description of the structural and evolutionary relationships between proteins whose structures are known.

All these works use information mostly about the secondary structures of proteins to analyize their structural similarities. Our approach deals with the chemical interactions in the molecules which is an orginal approach. We believe that each family of proteins presents a specific structural signature, that can be extracted from contact map images. With our methodology we can analyze the interactions similarity between molecules.

Another advantage of our methodology is that it can also deal with comparison of proteins with the same structure. Proteins with the same structure can have very different sequences and so, different amino acids interactions. In the other methods, proteins structurally similar are considered equal. Because we use dissimilarity metrics that are well-defined mathematically and continuously-valued, we can measure the amount by which multiple proteins from a single family differ from each other.

4. Image-matching methodology

In this section we show how different image-matching approaches may be used to compute measures of structural dissimilarity between two arbitrary proteins from their contact maps. In particular, we consider two conceptually distinct ways of treating this problem: as a *Content-Based Image Retrieval* (CBIR) problem or as a problem of image registration.

Content-Based Image Retrieval is a scientific discipline largely based on the notion that it is ideally possible to perform some form of semantics-preserving compression [16] of each image in a database into a signature vector that should be as small as possible to maximize the efficiency of image-based queries to the database later on. Usually, such signature vectors are computed from low-level primitives and their perceptual groupings [10], *i.e.*, from attributes that can be measured directly from images such as color, texture and geometric primitives (lines, segments, curves, boundaries, junctions, etc), and their spatial relationships in the image, which convey higher-level semantic cues.

A strong motivation to apply CBIR techniques to the protein classification problem is the growing size of protein databases such as the PDB. Even though indexing such large databases can be a costly operation, it may be done incrementally and, once it is finished, queries to the database are answered very efficiently. On the other hand, in spite of the fact that semantics-preserving encoding of complex protein structural properties into very small vectors is possible (that's what the *primary structure* is), it is not clear that performing such encoding directly from the contact maps is a computationally feasible problem.

Thus, alternatively, we propose a way of measuring how dissimilar any two proteins are based on the cost of registering the images formed by their contact maps. The image registration paradigm [5] is often used to match multiple images of a single object that suffers non-rigid deformations [8]. A cost is attributed to each deformation that the object may suffer and the image-to-image dissimilarity is computed by finding the minimum-cost deformation that maps one image onto the other.

A motivation to apply this idea to contact maps is that distinct proteins did evolve from common ancestral molecules, and their phylogenetic distances (Figure 5) are strongly correlated to structural dissimilarity. Thus, if we can somehow model the "deformations" needed to "warp" a contact map into another as a sequence of simple transformations that mimic the effects of evolutionary changes in proteic structure, the structural dissimilarity between any two proteins can be computed by finding the minimum-cost sequence of such transformations between their contact maps.

The trade-offs in the choice between these two paradigms are similar to those in the choice between feature-based and direct methods in Computer Vision: like feature-based methods, CBIR techniques tend to be more efficient on very large data sets, but on the other hand, like direct methods, image registration techniques tend to be more accurate, at least in terms of matching pairs of images that are indeed closely related.

4.1. The content-based image retrieval approach

In order to fully specify a Content-Based Image Retrieval algorithm, it is necessary to define how the signature vector of each possible image is generated and how the similarity between two arbitrary vectors is computed [9]. Figure 3 presents a schematic representation of the CBIR algorithm proposed in this paper.

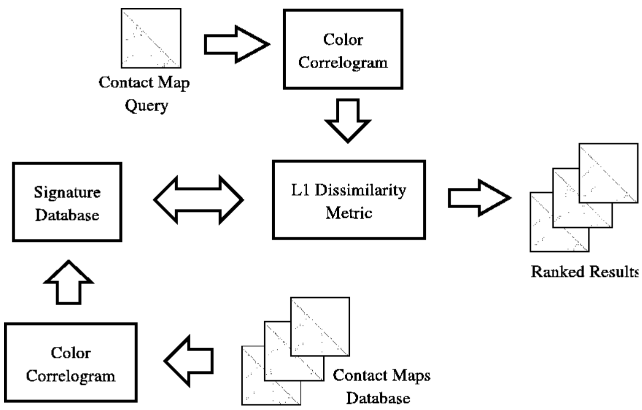

Figure 3. Content-Based Image Retrieval algorithm for protein similarity analysis.

We use the *Color Correlogram* [6] as the image signature, the d_1 distance measure for similarity analysis and image-based queries as input to the system.

The Color Correlogram expresses how the spatial correlation of pairs of colors changes with distance. It specifies the probability of finding a pixel of color j at distance k from a given pixel of color i. Let I be an $n \times n$ image with a color space quantized into m colors $c_1, \ldots, c_m$. Let also a distance $d \leq n$ be fixed *a priori*. Then, the correlogram of I is defined for $i\ ,\ j \in [m]$, $k \in [d]$ as

$$\gamma^{(k)}_{c_i,c_j}(I) \triangleq \underset{p_1 \in I_{c_i}, p_2 \in I}{Prob} \left[p_2 \in I_{c_j} \mid |p_1 - p_2| = k \right], \quad (1)$$

where the notation $p_1 \in I_{c_i}$ means that the color of pixel p_1 in image I is c_i, *i.e.*, that $p_1 \in I, I(p_1) = c_i$.

To compute the *correlogram* we have to evaluate the following equation:

$$\gamma^{(k)}_{c_i,c_j}(I) = \frac{\Gamma^{(k)}_{c_i,c_j}(I)}{h_{c_i} \cdot 8k}, \quad (2)$$

where h_{c_i} is the color histogram value of c_i and

$$\Gamma^{(k)}_{c_i,c_j} \triangleq \left| \left\{ p_1 \in I_{c_i}, p_2 \in I_{c_j} \mid |p_1 - p_2| = k \right\} \right|. \quad (3)$$

The naive algorithm to compute the formula above takes $O\left(n^2 d^2\right)$ time. Instead, we use a dynamic programming version, also proposed in [6], which takes only $O\left(n^2 d\right)$ in the overall running time.[2]

The d_1 measure is known to be relatively insensitive to the contents of individual vector elements. Instead, it corresponds to a weighted of average of discrepancy across the entire set of features in the image signatures. In the case of the correlograms of two images I and I', these weights are inversely proportional to $\gamma^{(k)}_{c_i,c_j}(I) + \gamma^{(k)}_{c_i,c_j}(I')$, *i.e.*, the larger this term is, the smaller is the influence of the pair of colors (c_i, c_j) in the overall measure. More specifically, the d_1 measure for the correlograms of images I and I' is

$$|I - I'|_{\gamma,d_1} \triangleq \sum_{\substack{i,j \in [m], \\ k \in [d]}} \frac{|\gamma^{(k)}_{c_i,c_j}(I) - \gamma^{(k)}_{c_i,c_j}(I')|}{1 + \gamma^{(k)}_{c_i,c_j}(I) + \gamma^{(k)}_{c_i,c_j}(I')}, \quad (4)$$

where the 1 in the denominator prevents division by zero. Importantly, once the Color Correlograms of two images have been built, the metric above takes only $O(d)$ time to compute, which means that queries on large databases are answered efficiently.

2 Since the number of colors in our application is very small, we do not analyze the computational complexity with respect to it.

4.2. The image registration approach

Our image registration methodology is loosely inspired on the *Approximate Stereo* work of Kutulakos [7], which introduced an algorithm to match multiple images in a way that is invariant to a class of transformations called *shuffle transforms*. A shuffle transform is a geometric transformation that causes a repositioning of individual pixels bounded by a *dispersion radius*, r. More specifically, two images I and I' are related by an r-*shuffle* if, and only if, for every pixel in I there is a pixel of identical color within a disk of radius r in I'.

The use of this kind of transformation in the analysis of structural protein dissimilarity is attractive because its spatially-localized nature preserves high-level geometric features, much as evolutionarily-feasible changes in a protein's primary structure do. However, the notion of dispersion radius as stated above is not appropriate for our application, because it is a worst-case global property, *i.e.*, if even one pixel in image I does not have an r-radius neighbor on I', then I and I' are *not* related by a r-*shuffle*.

Here, instead, we define the concept of *Average Dispersion Radius*, $\hat{r}_{\text{disp}}$, between two images as the average Euclidean distance between pixels in one image and the closest pixels with the same color in the other image. More formally, the Average Dispersion Radius between two $n \times n$ images is defined as

$$\hat{r}_{\text{disp}}\left(I, I'\right) \triangleq \frac{1}{2\,n^2} \sum_{i,j \in [n]} r(I, I', i, j) + r(I', I, i, j), \quad (5)$$

where

$$r(I, I', i, j) \triangleq \min_{\substack{x,y \in [n], \\ I(i,j) = I'(x,y)}} \left[\sqrt{(x-i)^2 + (y-j)^2} \right]. \quad (6)$$

The naive computation of the formulas above takes $O\left(n^4\right)$ time. However, we improve this to $O\left(n^2\right)$ by pre-computing, for each color $c_i, i \in [m]$, the distance transform relative to the pixels of image I that have color c_i, using the so-called *chamfer algorithm* [3] (which is $O\left(n^2\right)$), and then repeating this operation on image I'. After this preprocessing step, each term $r(I, I', i, j)$ in Eq. (5) can be computed in $O(1)$, simply by looking up the value at position (i, j) in the distance transform relative to the pixels of I' that have color $I(i, j)$.

In practice, in order to be able to use the Average Dispersion Radius as a reliable and efficiently-computable measure of protein structural dissimilarity, we exclude from the computation of Eq. (5) – both in the numerator and in the denominator – all **white** pixels, *i.e.*, those that do not represent actual contacts. As it can be observed in Figure 2, if no-contact pixels are disregarded, contact maps are sparse matrices with only $O(n)$ elements. By creating auxiliary lists of these elements, queries on the dissimilarity of two pre-processed contact maps may actually be answered in only $O(n)$ time.

Finally, it should be observed that two arbitrary contact maps seldom have exactly the same size. We deal with this technical difficulty by rescaling pixel coordinates in all contact maps to a pre-defined image size (1000×1000 in our current implementation) and then computing the dissimilarity metric defined in Eq. (5) with the optimizations described above.

5. Contact maps database

In the image database used in this work, we can distinguish among different kinds of chemical interactions through the different colors in the images. This is interesting because each type of interaction has a different rule in the folding and in the activity of the proteins.

The used contact maps contain information about the hydrophobic and electrostatic interactions and about the hydrogen bonds, the main interactions that govern the folding of a protein. The cut-off distances between amino acids used to select the contacts presented in the map were [2.0, 3.8] Å[3] for hydrophobic contacts, [2.0, 6.0] Å for the electrostatic interactions and [2.0, 3.2] Å for the hydrogen bonds. These values are biologically appropriate for selecting each type of interaction. There are other kinds of contacts but we intend to analyze them in future work.

All biological data used in this work comes from the PDB. The contact maps were generated using a module of STING [13], a software developed by EMBRAPA[4].

Our database contains 36 proteins of 20 different structural motifs. The used proteins were selected from SCOP[11]. In our selection we have α, β and $\alpha\beta$ proteins contact map images. The α class consists of proteins composed mostly by α-helices, the β proteins are composed mostly by β-sheets and the $\alpha\beta$ proteins are a mixture of these secondary structures.

To better evaluate our methodology, we included in our test database contact map images from:

- proteins with very different structure;
- proteins with similar structure but which are different proteins (with different functions);
- proteins of different animal species but that have the same structure (and function).

We selected 9 different *Myoglobins* from different animal species such as Whale, Horse, Elephant, Turtle, Sea Hare, Seal, Pig, Human and Tuna. The other 28 contact maps were chosen from varied α, β and $\alpha\beta$-proteins of different sizes.

3 1 Å (angstrom) = 10^{-10} meters

4 Empresa Brasileira de Pesquisa Agropecuária.

6. Experimental results

In this section, we present results from two experiments that we have done. The first one is used to evaluate the accuracy of the approach and the last one illustrates an alternative utility derived from the methodology.

6.1. Accuracy evaluation

To evaluate and compare the retrieval accuracy of the Color Correlogram and of the Avarage Dispersion Radius Metric, we have used a sample of images from the database composed by proteins with well know structural motifs.

Hence, we selected from our database 9 *Myoglobin* contact maps from different animal species. All *Myoglobins* have the same structure. We also selected 28 varied proteins of different structural motifs. In this set, we have α, β and $\alpha\beta$-proteins of different sizes. The image query was Human *Myoglobin*.

The optimum result that should ideally be obtained by the algorithms is a rank of proteins, based on the dissimilarity indexes, in which the whole *Myoglobin* set would be presented before any of the other proteins. In Figure 4, we present the results from the two analyzed algorithms. Each box represents one of the proteins from our sample and the value in the y axis is a normalized dissimilarity index. The normalization was necessary to compare the results presented by the different algorithms. The colors of the boxes indicate the type of the protein. The *Myoglobins* are presented in black. The darker the box color, the more similar to *Myoglobin* the protein is.

The Color Correlogram presented good results. From the 8 most similar proteins in the set of 36 used in the experiment, 6 were on the top of the similarity rank. All of them are proteins from mammals such as Whale, Pig, Elephant, Horse and Seal. The others considered not so similar were from Seal, Turtle, Fish and Sea Hare. These animals are more distant in the evolution line from the Human Being than the top-rank ones, as can be seen in Figure 5. The Average Dispersion Radius Metric presented better results as all the *Myoglobins* were presented in the top of the rank, wich implies in a 100% retrieval precision.

6.2. Similarity analysis of proteins of same structure

As shown above, our methodology presented good results when comparing proteins with different structures. An interesting feature of the usage of image-matching techniques is the possibility of dealing with proteins of same structural motif. This possibility arises because the algorithms supply a rank of the proteins ordered by the dissimilarity index. Hence, we can supply as input a database composed only by proteins with the same structure but from different animal species, for example, and obtain these proteins ordered by the dissimilarity index.

In Figure 6, we present the result of a similarity analysis of 8 *Myoglobins* from different animal species compared to the Human *Myoglobin*. Looking at the *Myoglobin* evolution hierarchy in the Figure 5, we can see that the nearer to Human *Myoglobin* are the *Myoglobins* from Elephant, Horse and Pig, with approximately the same distance in the evolution line, followed by the Whale, Seal, Turtle, Tuna and Sea Hare.

Note that, with the Color Correlogram, the lower dissimilarities values are assigned to mammals such as Whale, Pig, Elephant, Horse and Seal, in this order. The larger values occur with Turtle, Tuna and Sea Hare. With the Average Dispersion Radius Metric, the rank was Pig, Horse, Elephant, Seal, Turtle, Whale, Tuna and Sea Hare.

If we consider the weighted distances to Human *Myoglobin* in Figure 5 as a vector, and the ranks of the two algorithms as other vectors, all of them ordered by the animal species, we can calculate the vector correlations between the ranks produced with our methodology and the real hierarchy vector. These two correlation values, 0.9595 for Color Correlogram and 0.9768 for Average Dispersion Radius Metric, are a measure of how similar the results of the algorithms are with to real hierarchy.

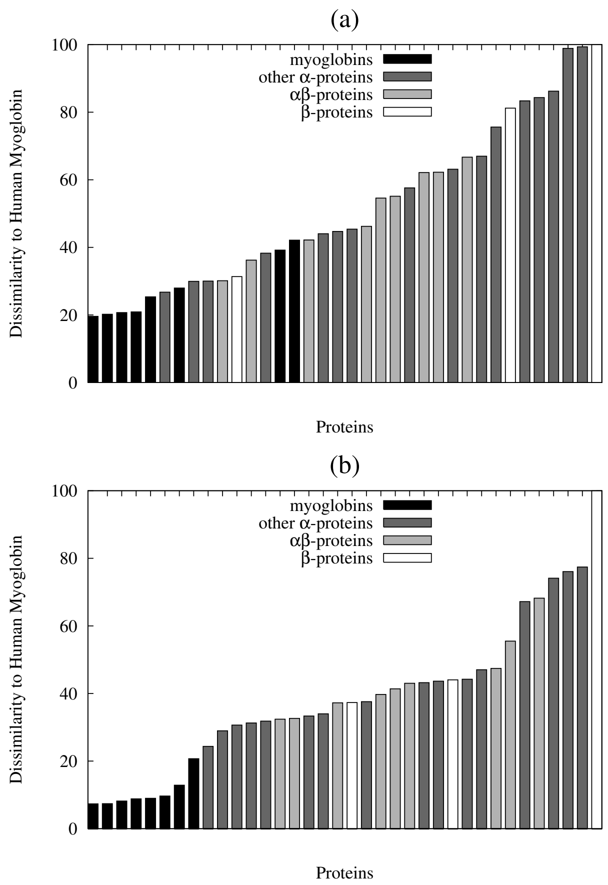

Figure 4. Dissimilarity to Human *Myoglobin* using the: (a) Color Correlogram and (b) the Average Dispersion Radius Metric

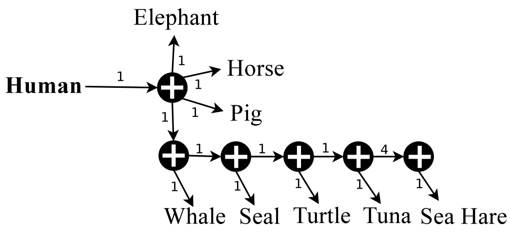

Figure 5. *Myoblobin* evolution sequence hierarchy - the greater the sum of weights between two animal species, the greater the sequence dissimilarity of their *Myoglobins*. Source: PALI [17].

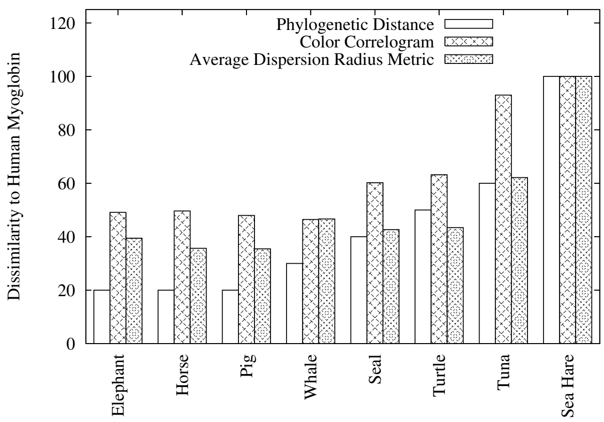

Figure 6. Dissimilarity rank between Human *Myoblobin* and the *Myoglobins* from other animal species

7. Conclusions and future works

In this work, we presented an original methodology to analyze protein similarity. We use image-matching techniques to, given an image database and a contact map image as the query, retrieve the contact map images of proteins with similar chemical interaction patterns. To evaluate our methodology, we implemented the Color Correlogram proposed by [6] and proposed an algorithm based on the Average Dispersion Radius Metric.

Our experiments show that the proposed methodology is appropriate for the similarity analysis of the contact maps image database. As each protein structural motif has a unique and specific pattern of contact, the algorithms used can distinguish between contact maps from different protein structures.

We analyzed the accuracy of the proposed system and it presented good results as with the Average Dispersion Radius Metric algorithm the whole set of same-class proteins was on the top of the rank.

Another applicability of this work is the similarity analysis of proteins with the same structural motif. We did some experiments with a varied set of *Myoglobins* from different animal species and we could verify that our rank is very close to the hierarchy presented in PALI[17], a web site with the phylogeny and alignment of homologous protein structures.

As the next steps we intend to, firstly, test the proposed methodology with a wider image database with contact maps of thousands of proteins. We intend to use the PDB Select [2] which is a representative subset of the PDB [2] and contains structures with low homologs sequences. This database provides a varied set of proteins with different cutoffs of similarity what makes possible to analyze the algorithms with proteins of different similarity degrees.

Secondly, we can add other types of chemical interactions to the image database used. The STING [13] contains information about cystein bridges, aromatic stacking, dissulfide bonds, among others. We believe that including more information in the images will improve the precision of the classification.

Finally, we intend to use other recently developed algorithms for content-based image retrieval and image registration and analyze their precision in protein structure classification. We plan to use a mixture of techniques and even propose a new algorithm that better deals with contact map matching.

References

[1] J. M. Berg, J. L. Tymoczko, and L. Stryer. *Biochemistry*. W. H. Freeman and Company, 2002.

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

[3] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics and Image Processing*, 34:344–371, 1984.

[4] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, Inc., 1999.

[5] L. G. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, 1992.

[6] J. Huang, S. R. Kumar, M. Mitra, W. J. Zhu, and R. Zabih. Image indexing using color correlograms. In *Proc. Computer Vision and Pattern Recognition Conf.*, pages 762–768, 1997.

[7] K. N. Kutulakos. Approximate N-view stereo. In *Proc. European Conf. on Computer Vision*, pages 67–83, 2000.

[8] J. B. A. Maintz and M. A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.

[9] A. Del Bimbo. *Visual information retrieval*. Morgan Kaufmann, 1999.

[10] A. Mojsilović, J. Gomes, and B. Rogowitz. Semantic-friendly indexing and quering of images based on the extraction of the objective semantic cues. *Int. J. Computer Vision*, 56(1-2):79–107, 2004.

[11] A. G. Murzin, L. Conte, A. Andreeva, D. Howorth, B. G. Ailey, S. E. Brenner, T. J. P. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures, 1995. http://scop.mrc-lmb.cam.ac.uk/scop/.

[12] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Molecular Biology*, 48:443–453, 1970.

[13] G. Neshich et al. Sting millennium: a web based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Research*, 31(13):3386–3392, 2003.

[14] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath – a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[15] C. A. Orengo and W. R. Taylor. Ssap: Sequential structure alignment program for protein structure comparison. *Methods in Enzymology*, 266:617–635, 1996.

[16] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1994.

[17] Phylogeny and alignment of homologous protein structures. http://pauling.mbu.iisc.ernet.in/pali/.

[18] T. ups and downs of protein topology; rapid comparison of protein structure. A. c. r. martin. *Protein Engineering*, 13(12):829–837, 2000.

[19] D. R. Westhead, D. C. Hatton, D. R. Gilbert, and J. M. Thornton. A www site devoted to protein structural topology. *embnet.news*, 5, 1998.