Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

# A contact map matching approach to protein structure similarity analysis

**Raquel C. de Melo[1,2], Carlos Eduardo R. Lopes[1], Fernando A. Fernandes Jr.[1], Carlos Henrique da Silveira[1,2], Marcelo M. Santoro[2], Rodrigo L. Carceroni[1], Wagner Meira Jr.[1] and Arnaldo de A. Araújo[1]**

[1]Departamento de Ciência da Computação, UFMG, Belo Horizonte, MG, Brasil
[2]Departamento de Bioquímica e Imunologia, UFMG, Belo Horizonte, MG, Brasil
Corresponding author: R.C. de Melo
E-mail: raquelcm@dcc.ufmg.br

Genet. Mol. Res. 5 (2): 284-308 (2006)
Received October 1, 2005
Accepted February 23, 2006
Published May 15, 2006

**ABSTRACT.** We modeled the problem of identifying how close two proteins are structurally by measuring the dissimilarity of their contact maps. These contact maps are colored images, in which the chromatic information encodes the chemical nature of the contacts. We studied two conceptually distinct image-processing algorithms to measure the dissimilarity between these contact maps; one was a content-based image retrieval method, and the other was based on image registration. In experiments with contact maps constructed from the protein data bank, our approach was able to identify, with greater than 80% precision, instances of monomers of apolipoproteins, globins, plastocyanins, retinol binding proteins and thioredoxins, among the monomers of Protein Data Bank Select. The image registration approach was only slightly more accurate than the content-based image retrieval approach.

**Key words:** Protein structure, Image-matching, Image registration, Contact maps, Content-based image retrieval

 

## INTRODUCTION

Bioinformatics is an emerging field undergoing rapid growth. This growth has mainly been fueled by advances in DNA sequencing and mapping techniques. The Genome Project has resulted in an exponentially growing database of genetic sequences, while the Structural Genomics Initiative is doing the same for the Protein Data Bank (PDB; Berman et al., 2000). One of the most active research areas in bioinformatics is the study of the relation between protein structure and function.

Proteins are the most versatile macromolecules in living systems, serving crucial functions in all biological processes. They function as catalysts and transporters, store other molecules, such as oxygen, provide mechanical support and immune protection, generate movement, transmit nerve impulses, and control growth and differentiation. Proteins are composed of sequences of amino acids, which is called primary structure. Different regions of the sequence form organized secondary structures, such as α-helices or β-strands. The tertiary structure, which is the three-dimensional structure of the protein, is formed by packing these 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 domains with specific three-dimensional structure, from a linear chain. It is known that the amino acid chain completely determines the structure of a protein. However, many proteins can have the same structure and the same function, but with very little sequence identity or similarity (see Gan et al., 2002). The study of a protein structure is very important, because the structure determines the protein function. In Figures 1-5, we present five proteins with different topologies.

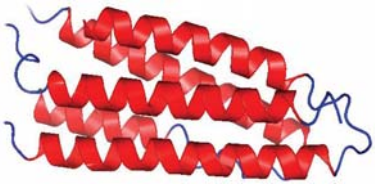

**Figure 1.** Human apolipoprotein from the protein data bank 1b68.

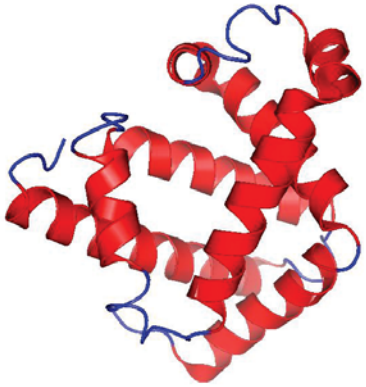

**Figure 2.** Whale globin from the protein data bank 101m.

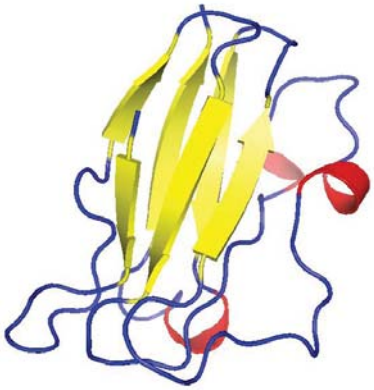

**Figure 3.** Spinach plastocyanin from the protein data bank 1ag6.

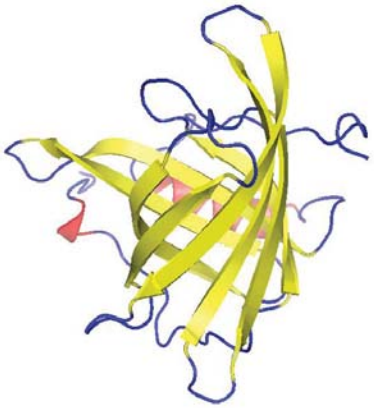

**Figure 4.** Pig retinol-binding protein from the protein data bank 1aqb.

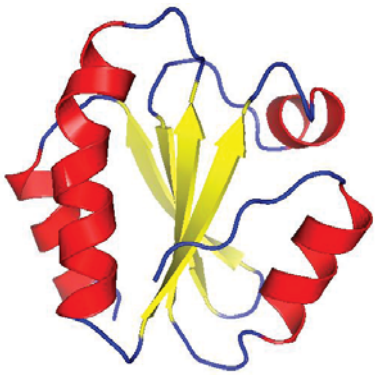

**Figure 5.** Human thioredoxin from the protein data bank 1aiu.

By January 2005, the PDB had approximately 29,000 proteins in its archives; this number is increasing continuously. In 2004, approximately 5,000 proteins were added. Even though

the PDB is now a huge data set, considerable information still needs to be added, especially concerning protein structure.

The inter-residue chemical interactions are very important for the folding of a protein and for keeping its shape once it is folded. Each protein family presents a specific pattern of contacts that we believe can produce a structural signature for that family.

Here, we have used the traditional contact maps and an image-matching approach to analyze the similarity of protein structures. We selected proteins with different topologies, and we used PDB Select (Hobohm et al., 1992; Hobohm and Sander, 1994) to evaluate the performance of this approach. The PDB Select database is a subset of the structures in the PDB that does not contain (highly) homolog sequences. For each selected protein, we built the respective contact map, which is an image in which the colors represent types of chemical interactions between two amino acids. These contact maps are a two-dimensional representation of the protein structure. Content-Based Image Retrieval (CBIR) and Image Registration (IR) techniques are used to measure the dissimilarity between contact maps, making it 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 of the substrate, and they are involved in catalysis as well. Thus, it is as important to study the similarity of proteins based on their internal chemical interactions, as it is to find a pattern of chemical interactions for each protein family.

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

- The hydrophobic interactions consist of the attraction between hydrophobic side chains of residues of 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 or repulsion between amino acids with different or equal charges. Most of them are located on 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 $\alpha$-helices and the $\beta$-strands, structures that maintain the folding of the protein (Branden and Tooze, 1999).

We used the spatial location of these three types of interactions as the features from which protein structure is identified. Structural comparison is a central task in biomedical research. Identifying structural similarities can provide significant insights into the relation between structure and function in proteins. Reliable and efficient structural matching plays a key role in rational drug design and in assessing the structure prediction methods. Other applications of protein structure analysis include validation of protein models, identification of native folding motifs among incorrect alternatives, identification of possible folds for a sequence of unknown structure, and finding sequences compatible with given structures.

## RELATED STUDIES

Protein structure has been a topic of great interest during recent years. Some researchers

have explored the positioning of the secondary structures to classify protein structures; others work on atomic detail and try to develop templates for each protein family. There is also some research on contact maps, in an attempt to align protein pairs structurally.

TOPS (Westhead et al., 1998) is a web site for protein structure classification. It presents an atlas of drawings representing the structure of proteins. These are two-dimensional schemes that display a fold as a sequence of secondary structures, along with their relative orientation and spatial position. TOPS performs protein classification, based on pattern searching, using a string-graph algorithm. Another method used for the comparison of protein structures is TOPSCAN (Martin, 2000). This system also uses secondary structures and their relative direction, proximity, accessibility, and length. This method uses the Needleman and Wunsch dynamic programming algorithm called Needleman. The CATH (Orengo et al., 1997) 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 (Orengo and Taylor, 1996). Parameters for clustering domains into fold families were determined by empirical trials throughout the database. SCOP (Murzin et al., 1995) is a web site of protein hierarchy. It was created by manual inspection and abetted by automated methods, aiming to provide a description of the structural and evolutionary relationships between proteins with known structures.

In a study on atomic detail (Chew and Kedem, 2002), coordinates of the $\alpha$-carbons were used to generate a signature for each protein, and a protein consensus was compiled for each family of proteins. Through this consensus, it is possible to analyze the similarity of protein families and also to classify protein structures into families. Gan et al. (2002) also examined atomic detail to analyze the variations in the three-dimensional structures of two proteins through their root mean square values, and they compared their findings with the sequence similarity of these proteins.

All of these studies used information on the chemical interactions between residues in the proteins. 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 similarity of interactions between the molecules. Furthermore, using dissimilarity measurements that are well-defined mathematically and continuously valued, we can measure how much proteins from a single family differ from each other.

Lancia (2001) and Carr et al. (2002) used the traditional contact maps to align pairs of proteins structurally. Their maps do not differentiate among types of contacts, while ours do. They attempted to overlap contact maps, mapping contacts of one map to contacts of the other; this overlap gives a score that indicates how many contacts were matched. This score makes it possible to use their methodology for protein classification. However, these algorithms are very expensive computationally. Among the 528 alignments that they examined, only 42 were optimally solved in less than an hour. With our methodology, the dissimilarity between a pair of proteins can be computed in a minute, or even less.

## MATERIAL AND METHODS

### The contact maps

Contact maps are useful tools to study protein structure. In our system, each contact

map consists of a colored image that represents different kinds of chemical interactions between all the amino acids of a protein. Our database is composed of contact map images with *n* x *n* pixels, where *n* is the number of residues 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 the limitation that amino acids can have more than one kind of interaction, which cannot be expressed in these two-dimensional images. So, we decided to prioritize the electrostatic interactions, followed by 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.

Also, since the value of *n* varies from a protein to another, all the images must be normalized to the same dimensions before applying the image-matching algorithms.

In Figures 6-10, we present plots of protein chemical interactions as contact map images in our database. These figures came from the proteins presented in Figures 1, 2, 3, 4, and 5, respectively.

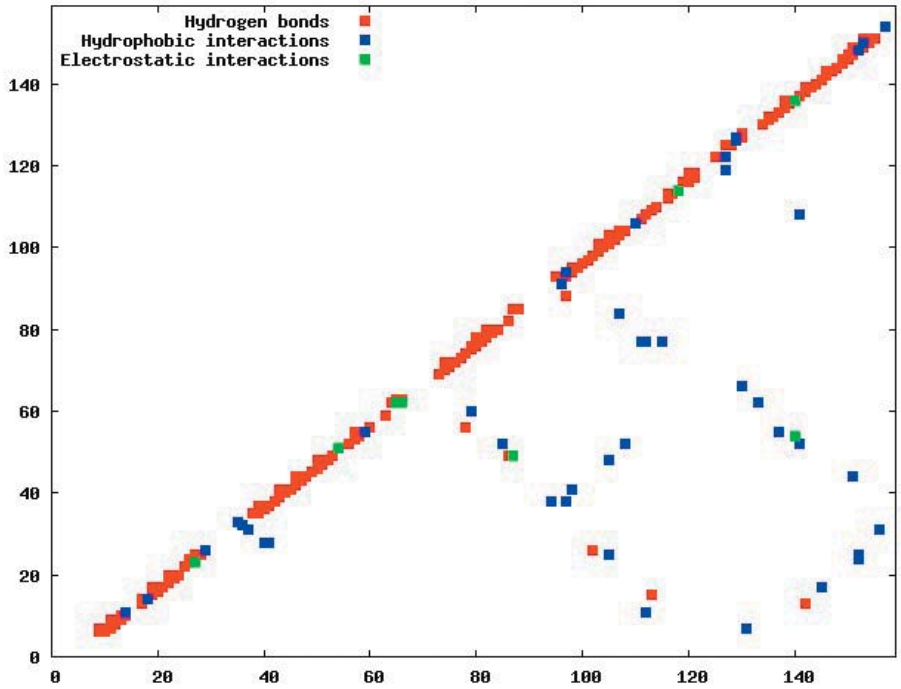


**Figure 6.** Contact map of apolipoprotein from the protein data bank 1b68 (Figure 1). The hydrophopic interactions are presented in blue, the electrostatic interactions in green and the hydrogen bonds in red.

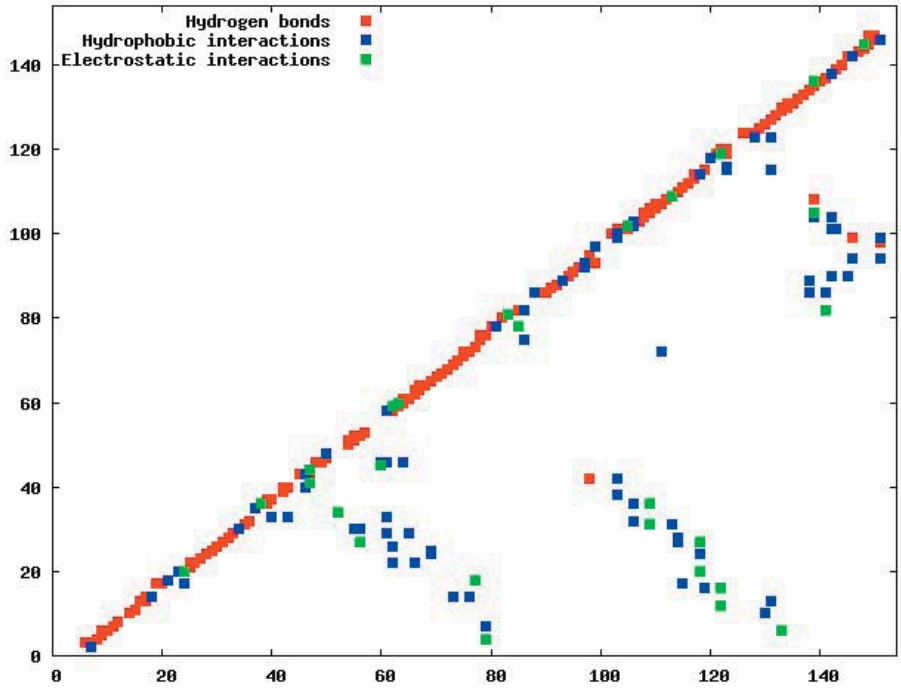


**Figure 7.** Contact map of globin from the protein data bank 101m (Figure 2).

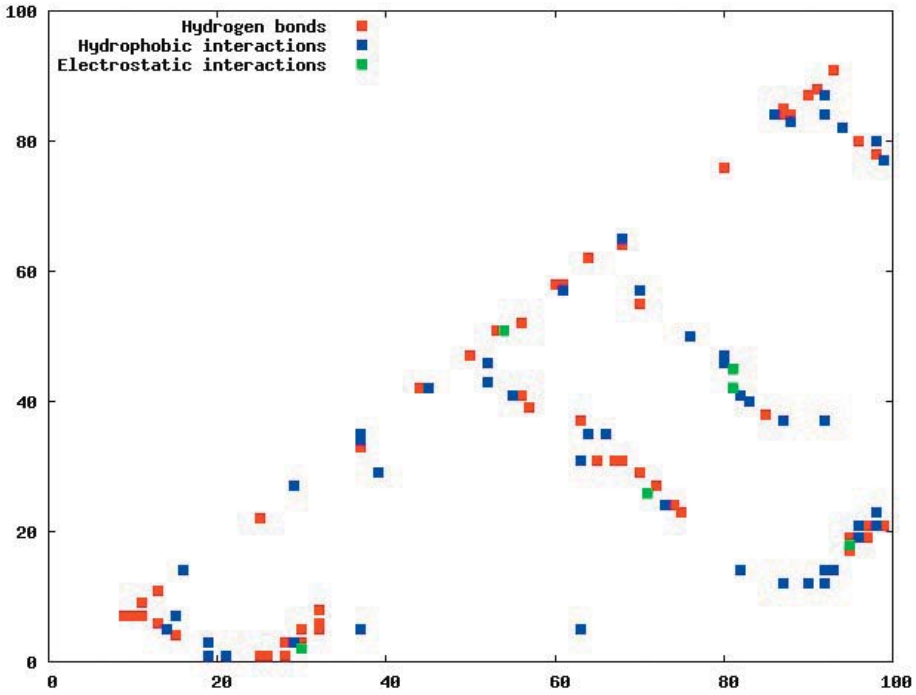


**Figure 8.** Contact map of plastocyanin from the protein data bank 1ag6 (Figure 3).

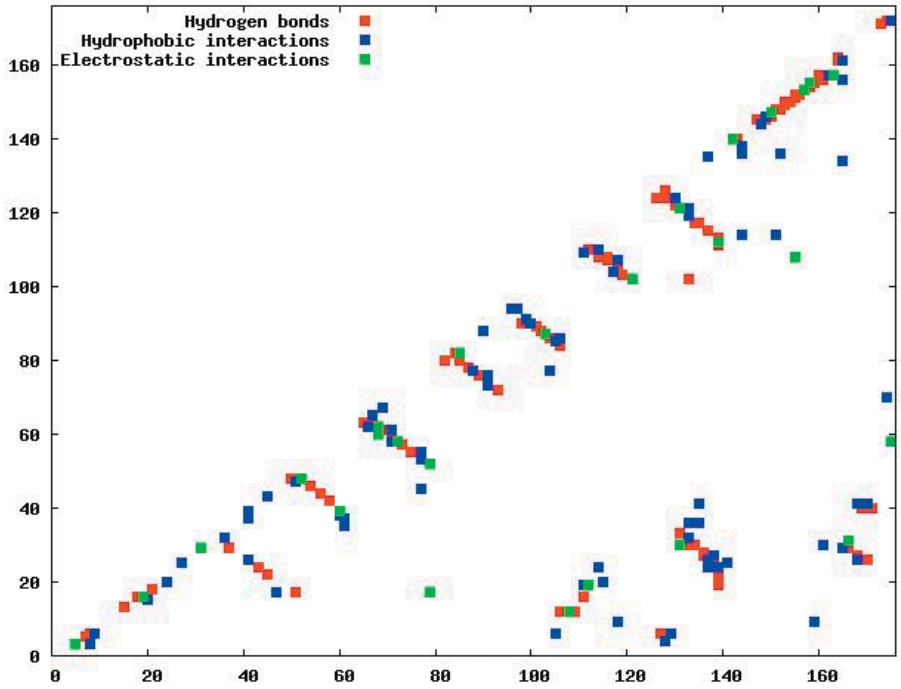


**Figure 9.** Contact map of pig retinol-binding protein from the protein data bank 1aqb (Figure 4).

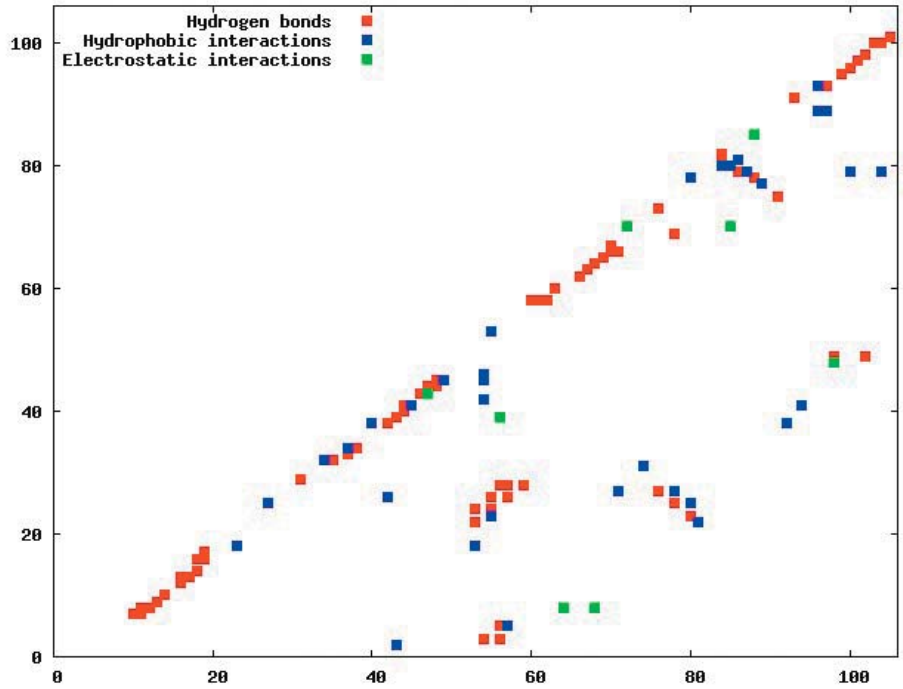


**Figure 10.** Contact map of thioredoxin from the protein data bank 1aiu (Figure 5).

With the increasing growth 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 searches are performed by text query on protein identifications, names and other features, but all of them include analyses of structural similarity. Using contact maps, as visual representations of proteins, allows us to use an image-matching methodology in order to analyze structural similarity between them. Hence, we expect to, given a specific protein, select the protein maps that have similar interaction patterns and thus similar protein structures.

## The database

We can distinguish among different kinds of chemical interactions through the different colors in the images in our database. This is useful because each type of interaction has a different rule in the folding and in the activity of the proteins. The 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 for hydrophobic contacts, [2.0, 6.0] Å for the electrostatic interactions and [2.0, 3.2] Å for the hydrogen bonds. There are other kinds of contacts that we intend to analyze in the future.

All biological data came from the PDB. The contact maps were generated using a module of STING (Neshich et al., 2003). To test the performance of our approach, we selected all the monomers with five different protein topologies from the PDB. We selected the globins, which are the oxygen carriers in the muscles and the blood and which are very well-studied proteins. This family of proteins is composed only of α-helices. There are 224 globins in the PDB. We also used the 13 apolipoproteins, which are lipoproteins composed of bundled α-helices. The plastocyanins are electron transporters, composed mostly of β-sheets. We found 15 of them in the PDB. Another family used is the retinol-binding proteins (RBP), which are also composed of β-sheets, but are barrel shaped. There are actually 18 of them in the PDB. Finally, we used the thioredoxins, which are electron transporters composed of both α-helices and β-sheets. We found eight of them in the PDB. The identifications of all of this test set are indicated in Table 1.

Our objective was to retrieve all proteins of similar structures, within a mixture with other proteins of different topologies. We tested this with 187 proteins of different topologies. These proteins are all monomers of the PDB Select list (Table 2).

## Image-matching methodology

We examined how different image-matching approaches may be used to compute measures of structural similarity between two arbitrary proteins from their contact maps. In particular, we consider two conceptually distinct ways of treating this problem: as a CBIR problem or as a problem of IR.

CBIR is a scientific discipline largely based on the notion that it is ideally possible to perform some form of semantics-preserving compression (Pentland et al., 1994) of each image in a database into a signature vector, which 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 (Mojsilovic et al., 2004), i.e.,

**Table 1.** Protein data bank identification for the protein families used in our experiments.

| Protein family | PDB identification |
|---|---|
| Apolipoproteins | 1aep, 1b68, 1bz4, 1ea8, 1gs9, 1h7i, 1le2, 1le4, 1lpe, 1nfn, 1nfo, 1or2, 1or3 |
| Globins | 101m, 102m, 103m, 104m, 105m, 106m, 107m, 108m, 109m, 110m, 111m, 112m, 1a6g, 1a6k, 1a6m, 1a6n, 1abs, 1ajg, 1ajh, 1ash, 1azi, 1b0b, 1b2v, 1bje, 1bvc, 1bvd, 1bz6, 1bzp, 1bzr, 1ch1, 1ch2, 1ch3, 1ch5, 1ch7, 1ch9, 1cik, 1cio,1co8, 1co9, 1cp0, 1cp5, 1cpw, 1dlw, 1dly, 1dm1, 1do1, 1do3, 1do4, 1do7, 1dti, 1dtm, 1duk, 1duo, 1dwr, 1dws, 1dwt, 1dxc, 1dxd, 1ebc, 1ebt, 1eca, 1ecd, 1ecn, 1eco, 1emy, 1f63, 1f65, 1f6h, 1fcs, 1flp, 1gdi, 1gdj, 1gdk, 1gdl,1gjn, 1h1x, 1hbg, 1hjt, 1hlb, 1hlm, 1hrm, 1hsy, 1iop, 1irc, 1j52, 1jdo, 1jl6, 1jl7, 1jp6, 1jp8, 1jp9, 1jpb, 1jw8, 1kfr, 1kr7, 1lh1, 1lh2, 1lh3, 1lh5, 1lh6, 1lh7, 1lhs, 1lht, 1ltw, 1lue, 1mba, 1mbc, 1mbd, 1mbi, 1mbn, 1mbo,1mbs, 1mcy, 1mgn, 1mlf, 1mlg, 1mlh, 1mlj, 1mlk, 1mll, 1mlm, 1mln, 1mlo, 1mlq, 1mlr, 1mls, 1mlu, 1mnh, 1moa, 1mob, 1moc, 1mod, 1moh, 1mti, 1mtj, 1mtk, 1mym, 1myt, 1myz, 1mz0, 1n9f, 1n9h, 1n9i, 1n9x, 1naz, 1npf, 1npg, 1nz2,1nz3, 1nz4, 1nz5, 1o16, 1obm, 1ofj, 1ofk, 1q1f, 1rse, 1rtx, 1spe, 1swm, 1tes, 1tu9, 1utg, 1uvy, 1v07, 1v5h, 1vxa, 1vxb, 1vxc, 1vxd, 1vxe, 1vxf, 1vxg, 1vxh, 1wla, 1xch, 1yma, 1ymb, 1ymc, 1yog, 1yoh, 1yoi, 2cmm, 2fal, 2fam,2gdm, 2hbg, 2lh1, 2lh2, 2lh3, 2lh5, 2lh6, 2lh7, 2lhb, 2mbw, 2mga, 2mgb, 2mgc, 2mgd, 2mge, 2mgf, 2mgg, 2mgh, 2mgi, 2mgj, 2mgk, 2mgl, 2mgm, 2mm1, 2mya, 2myb, 2myc, 2myd, 2mye, 2spl, 2spm, 2spn, 2spo, 3mba, 4mba, 4mbn, 5mba, 5mbn |
| Plastocyanins | 1ag6, 1byp, 1iuz, 1kdi, 1oow, 1plc, 1pnc, 1pnd, 2pcy, 2plt, 3pcy, 4pcy, 5pcy, 6pcy, 7pcy |
| Retinol-binding proteins | 1aqb, 1brp, 1brq, 1erb, 1fel, 1fem, 1fen, 1hbp, 1hbq, 1iiu, 1jyd, 1jyj, 1kt3, 1kt4, 1kt5, 1kt6, 1kt7, 1rbp |
| Thioredoxins | 1aiu, 1faa, 1gh2, 1h75, 1tho, 1thx, 1wou, 2tir |

**Table 2.** Protein data bank identification for all monomers of PDB Select.

| PDB identification | Classification | Number of residues |
|---|---|---|
| 1c53 | Electron transport | 79 |
| 2ila | Cytokine | 145 |
| 1efm | Elongation factor | 158 |
| 1tia | Hydrolase (carboxylic esterase) | 271 |
| 1dpi | Nucleotidyltransferase | 546 |
| 1aat | Aminotransferase | 411 |
| 1ian | Serine/threonine-protein kinase | 328 |
| 1pho | Outer membrane protein | 330 |
| 1xrc | Methyltransferase | 377 |
| 1cne | Oxidoreductase (nitrogenous acceptor) | 260 |
| 1nom | Nucleotidyltransferase | 242 |
| 4hb1 | Designed helical bundle | 44 |
| 1fdi | Oxidoreductase | 715 |
| 1bpm | Hydrolase (α-aminoacylpeptide) | 481 |
| 1rgs | Kinase | 264 |
| 1lfb | Transcription regulation | 77 |
| 1fsz | Cell-division protein | 334 |

Continued on next page

**Table 2.** Continued.

| PDB identification | Classification | Number of residues |
|---|---|---|
| 1bbs | Aspartic proteinase | 331 |
| 1ysc | Hydrolase (carboxypeptidase) | 421 |
| 1bgw | DNA-binding protein | 679 |
| 2cah | Oxidoreductase ($H_2O_2$ acceptor) | 475 |
| 1pex | Metalloprotease | 192 |
| 1glv | Glutathione biosynthesis ligase | 299 |
| 1a0i | Ligase | 332 |
| 1wkd | tRNA-modifying enzyme | 372 |
| 1cby | Toxin | 227 |
| 1aod | Hydrolase | 274 |
| 1914 | ALU domain | 171 |
| 1gwz | Hydrolase | 280 |
| 1dtp | Toxin | 190 |
| 1hup | C-type lectin | 141 |
| 1cyw | Electron transport | 159 |
| 1vdc | Oxidoreductase | 322 |
| 1juk | Lyase | 247 |
| 1pdy | Lyase (carbon-oxygen) | 433 |
| 1ax8 | Cytokine | 130 |
| 1ah5 | Lyase | 299 |
| 8ohm | Helicase | 435 |
| 1a41 | Isomerase | 221 |
| 1c25 | Hydrolase | 161 |
| 1bob | Acetyltransferase | 306 |
| 1aln | Hydrolase | 294 |
| 1auq | Willebrand | 208 |
| 1gal | Oxidoreductase (flavoprotein) | 581 |
| 1cen | Cellulose degradation | 334 |
| 8prn | Membrane protein | 289 |
| 1pfo | Toxin | 471 |
| 1gcb | DNA-binding protein | 452 |
| 1grj | Transcription regulation | 151 |
| 1kte | Electron transport | 105 |
| 1am2 | Intein | 181 |
| 1tul | Telokin-like protein | 102 |
| 1ash | Oxygen storage | 107 |
| 1br9 | Proteinase inhibitor | 182 |
| 1rmd | DNA-binding protein | 116 |
| 1ryt | Electron transport | 190 |
| 1a3k | Galectin | 137 |
| 1pht | Phosphotransferase | 83 |
| 1aol | Viral glycoprotein | 228 |
| 1a1x | Proto-oncogene | 106 |
| 1tig | Ribosome-binding factor | 88 |
| 1btn | Signal transduction protein | 106 |
| 1amx | Bacterial adhesin | 150 |
| 1hoe | Glycosidase inhibitor | 74 |

Continued on next page

**Table 2.** Continued.

| PDB identification | Classification | Number of residues |
|---|---|---|
| 1uox | Oxidoreductase | 295 |
| 4mt2 | Metallothionein | 61 |
| 1sra | Calcium-binding protein | 151 |
| 1vid | Transferase (methyltransferase) | 214 |
| 1poc | Hydrolase | 134 |
| 1by2 | Extracellular module | 112 |
| 1xer | Electron transport | 103 |
| 3tdt | Acyltransferase | 274 |
| 2dtr | Repressor | 214 |
| 1ptq | Phosphotransferase | 50 |
| 1bea | Serine protease inhibitor | 116 |
| 1rss | Ribosomal protein | 140 |
| 1at0 | Developmental signaling molecule | 145 |
| 1alu | Cytokine | 157 |
| 1mai | Signal transduction protein | 119 |
| 1cif | Electron transport (heme protein) | 108 |
| 1dxy | Oxidoreductase | 130 |
| 1fen | Transport protein | 176 |
| 1rec | Calcium-binding protein | 185 |
| 1cpo | Oxidoreductase | 299 |
| 4bcl | Electron transport | 350 |
| 1sfp | Spermadhesin | 111 |
| 3tss | Toxin | 190 |
| 1fit | Chromosomal translocation | 126 |
| 1vls | Chemotaxis | 146 |
| 2abk | Endonuclease | 211 |
| 3mag | mRNA processing | 292 |
| 1ak0 | Endonuclease | 264 |
| 1mml | Reverse transcriptase | 251 |
| 1ayl | Kinase (transphosphorylating) | 532 |
| 1bdo | Transferase | 80 |
| 1hyp | Hydrophobic seed protein | 75 |
| 1tml | β-amylase | 286 |
| 2sak | Plasminogen activator | 121 |
| 1al3 | Transcription regulation | 237 |
| 2acy | Acylphosphatase | 98 |
| 1pgs | Endoglycosidase | 311 |
| 1nar | Plant seed protein | 289 |
| 1iab | Zinc endopeptidase | 200 |
| 1bkb | Translation | 136 |
| 1cv8 | Cysteine protease | 173 |
| 1chd | Carboxyl methylesterase | 198 |
| 1amf | Binding protein | 231 |
| 3cla | Transferase (acyltransferase) | 213 |
| 1ajj | Receptor | 37 |
| 1nkr | Inhibitory receptor | 195 |
| 1vie | Oxidoreductase | 60 |

Continued on next page

**Table 2.** Continued.

| PDB identification | Classification | Number of residues |
|---|---|---|
| 1pdo | Phosphotransferase | 129 |
| 1ako | Nuclease | 268 |
| 1mof | Coat protein | 53 |
| 1dhn | Pterine binding | 121 |
| 1cnv | Seed protein | 283 |
| 1vcc | DNA binding | 77 |
| 1gvp | DNA-binding protein | 87 |
| 1ads | Oxidoreductase | 315 |
| 1jer | Electron transport | 110 |
| 1a3c | Transcription regulation | 166 |
| 1edg | Cellulose degradation | 380 |
| 16pk | Kinase | 415 |
| 1b6a | Angiogenesis inhibitor | 355 |
| 1a8e | Iron transport | 329 |
| 1aru | Peroxidase (donor:$H_2O_2$ oxidoreductase) | 336 |
| 3cyr | Electron transport | 107 |
| 1nif | Oxidoreductase (nitric oxide(a)) | 333 |
| 1mrj | Ribosome-inactivating protein | 247 |
| 2ilk | Cytokine | 155 |
| 1ppn | Hydrolase (sulfhydryl proteinase) | 212 |
| 1nox | Flavoenzyme | 200 |
| 2a0b | Sensory transduction | 118 |
| 1a8d | Neurotoxin | 452 |
| 1moq | Glutamine amidotransferase | 366 |
| 1a62 | Transcription termination | 125 |
| 1orc | Gene-regulating protein | 64 |
| 1kpf | Protein kinase inhibitor | 111 |
| 1whi | Ribosomal protein | 122 |
| 1rie | Electron transport | 127 |
| 1mla | Acyltransferase | 305 |
| 1opd | Phosphotransferase | 85 |
| 1ezm | Hydrolase | 298 |
| 1cyo | Electron transport | 88 |
| 1brt | Haloperoxidase | 277 |
| 2sns | Hydrolase (phosphoric diester) | 141 |
| 8abp | Binding proteins | 305 |
| 3seb | Toxin | 238 |
| 1g3p | Minor coat protein | 192 |
| 1bgf | Transcription factor | 124 |
| 1aba | Electron transport | 87 |
| 1yge | Dioxygenase | 839 |
| 3vub | CCDB | 101 |
| 1eca | Oxygen transport | 136 |
| 2ctc | Hydrolase (C-terminal peptidase) | 307 |
| 1nxb | Neurotoxin (post-synaptic) | 62 |
| 1ppt | Pancreatic hormone | 36 |
| 1rhs | Transferase | 293 |

Continued on next page

**Table 2.** Continued.

| PDB identification | Classification | Number of residues |
|---|---|---|
| 1utg | Steroid binding | 70 |
| 1plc | Electron transport | 99 |
| 1bk0 | B-lactam antibiotic | 329 |
| 1dcs | Oxidoreductase | 279 |
| 1c52 | Electron transport protein | 131 |
| 1oaa | Oxidoreductase | 259 |
| 2pth | Hydrolase | 193 |
| 2sn3 | Toxin | 65 |
| 1amm | Crystallin | 174 |
| 1bx7 | Anti-coagulant | 51 |
| 1mun | DNA repair | 225 |
| 1ifc | Lipid-binding protein | 131 |
| 1b6g | Hydrolase | 310 |
| 1ctj | Electron transport | 89 |
| 2igd | IGG-binding protein | 61 |
| 1nkd | Transcription regulation | 59 |
| 3sil | Glycosidase | 379 |
| 2erl | Pheromone | 40 |
| 1a6m | Oxygen transport | 151 |
| 1cex | Serine esterase | 197 |
| 1ixh | Phosphate transport | 321 |
| 1byi | Ligase | 224 |
| 1aho | Neurotoxin | 64 |
| 1nls | Agglutinin | 237 |
| 2fdn | Electron transport | 55 |
| 3lzt | Hydrolase | 129 |
| 1rb9 | Iron-sulfur protein | 52 |
| 3pyp | Photoreceptor | 125 |
| 1gci | Serine protease | 269 |

from attributes that can be measured directly in 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 is 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 means of measuring how dissimilar any two proteins are based on the cost of registering the images formed by their contact maps. The IR paradigm (Brown, 1992) is often used to match multiple images of a single object that suffers non-rigid deformations (Maintz and Viergever, 1998). A cost is attributed to each deformation that the

object may suffer and the image-to-image dissimilarity is computed by finding the lowest-cost deformation that maps one image onto the other.

A motivation to apply this idea to contact maps is that distinct proteins evolved from common ancestral molecules, and consequently proteins with the same structure (thus similar contact maps) can be more than 80% different in their amino acid composition. 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 protein structure, the structural dissimilarity between any two proteins can be computed by finding the minimum-cost sequence of such transformations between their contact maps.

This proposed methodology was tested with two techniques from different paradigms; there are trade-offs in the choice between them. Like feature-based methods, CBIR techniques tend to be more efficient with very large data sets, but on the other hand, like direct methods, IR techniques tend to be more accurate, at least in terms of matching pairs of images that are indeed closely related.

*The content-based image retrieval approach*

In order to fully specify a CBIR algorithm, it is necessary to define how the signature vector of each possible image is generated and how the dissimilarity between two arbitrary vectors is computed (Del Bimbo, 1999). Figure 11 presents a schematic representation of our proposed CBIR system.

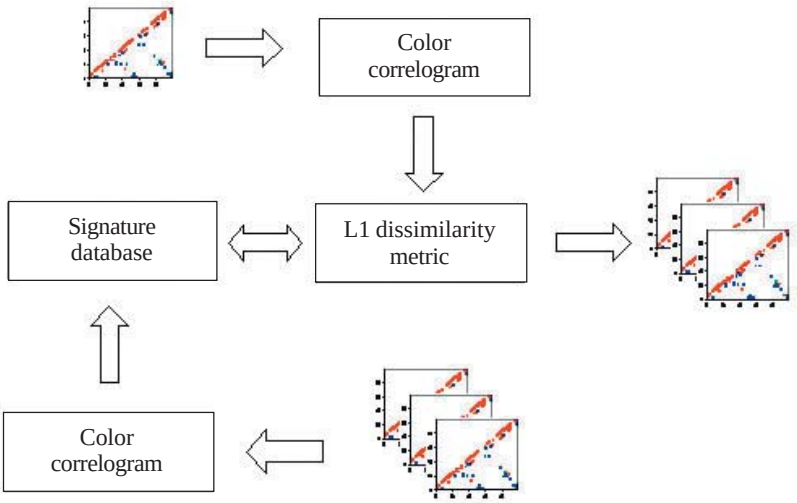


**Figure 11.** Content-based image retrieval (CBIR) system for protein similarity analysis. IR= image registration.

We used the Color Correlogram (Huang et al., 1997) as the image signature, the $d_1$ distance measure for dissimilarity 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* x *n* image with a color space quantized into *m* colors $c_1,\ldots, c_m$. Also let 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:

$$\gamma_{c_i,c_j}^{(k)}(I) \triangleq \underset{p_1 \in I_{c_i},\, p_2 \in I}{Prob} [p_2 \in I_{c_j} \,||\, p_1 - p_2| = k] \qquad \text{(Equation 1)}$$

where the notation $p_1 \in I_i$ 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:

$$\gamma_{c_i,c_j}^{(k)}(I) = \frac{\Gamma_{c_i,c_j}^{(k)}(I)}{h_{c_i} \cdot 8k} \qquad \text{(Equation 2)}$$

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

$$|I - I'|_{\gamma,d_1} \triangleq \sum_{i,j \in [m],\, k \in [d]} \frac{|\gamma_{c_i,c_j}^{(k)}(I) - \gamma_{c_i,c_j}^{(k)}(I')|}{1 + \gamma_{c_i,c_j}^{(k)}(I) + \gamma_{c_i,c_j}^{(k)}(I')} \qquad \text{(Equation 3)}$$

The $d_1$ measure is known to be relatively insensitive to the contents of individual vector elements. Instead, it corresponds to a weighted 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 the sum of the correlograms, i.e., the larger this sum 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 correlogram of images $I$ and $I'$ is

$$\Gamma_{c_i,c_j}^{(k)} \triangleq |\{p_1 \in I_{c_i}, p_2 \in I_{c_j} \,||\, p_1 - p_2| = k\}| \qquad \text{(Equation 4)}$$

where 1 in the denominator avoids division by zero. Importantly, once the color correlograms of two images have been built, the calculation time increases linearly, based on the signature vector size to be computed, which means that queries on large databases are answered efficiently.

*The image registration approach*

With IR, it is not necessary to compute a signature for each image, but, as with the Color Correlogram, this method computes a dissimilarity measure between two maps.

This methodology is loosely inspired on the Approximate Stereo work of Kutulakos (2000), which introduced an algorithm to match multiple images in a way that is invariant within 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 protein structural similarity 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 an *r*-shuffle.

Here, instead, we define the concept of Average Dispersion Radius 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* x *n* images is defined as:

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

where

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

Thus, to compute the dissimilarity index between a query map, and another which we call base map, for each contact in the query map, the closest contact (of the same type) to the corresponding position in the base map is searched, and the distance between these two positions is calculated. The distances are accumulated for all the searches. Then, the maps have their roles inverted, such that the base map becomes the query map, and vice-versa. This process is repeated, and the distances obtained are accumulated with the previous values. In this way, the measure is not dependent on which one is taken as query or base map. Finally, the accumulated distance is divided by the number of searches. The dissimilarity measure is then defined as the mean distance in the searches.

## RESULTS AND DISCUSSION

Given a database of contact map images and an algorithm, we need to use a specific image as a query to search for similar proteins. That is, when we want to search for globins in the database we have to use one specific globin as query. Thus, to verify the accuracy of the methodology and of the proposed algorithms in retrieving globins, we are expected to query the database using all the globin contact maps of the database.

### Evaluation of retrieval performance

We selected five different protein families to test this proposed methodology. Our objective was to determine if the system is able to retrieve similar protein structures using each of the proteins of each family as queries. For that, we used the well-known statistical concepts of the confusion matrix and receiver operating characteristic (ROC) curves. A confusion matrix (Provost and Kohavi, 1998) contains information about actual and predicted classifications done

by a classifier and makes it possible to evaluate the performance of classification systems. This matrix gives the true-negative, true-positive, false-negative, and false-positive rates.

ROC curves are another way to examine the performance of classifiers (Swets, 1988). An ROC graph is a plot with the false-positive rate on the X-axis and the true-positive rate on the Y-axis. The false-positive rate is the number of negative instances predicted as positives divided by the number of negative instances. The true-positive rate is the number of positive instances predicted as positives divided by the number of positive instances.

In the ROC space, the point (0,1) is the perfect classifier: it classifies all positive cases and negative cases correctly. It is (0,1) because the false-positive rate is 0 (none), and the true-positive rate is 1 (all). The point (0,0) represents a classifier that predicts all cases to be negative, while the point (1,1) corresponds to a classifier that predicts every case to be positive. Point (1,0) is the classifier that is incorrect for all classifications.

In many cases, a classifier has a parameter that can be adjusted to increase true-positives at the cost of increasing false-positives or decreasing false-positives at the cost of decreasing true-positives. Each parameter setting provides a (false-positive, true-positive) pair and a series of such pairs can be used to plot an ROC curve. In our algorithms, the parameter used is a threshold that we use to decide if a protein is or is not of a given family.

An ROC curve is independent of class distribution or error costs, and it encapsulates all information contained in the confusion matrix, since false-negatives is the complement of true-positives and true-negatives is the complement of false-positives (Swets, 1988). These curves provide a visual tool for examining the tradeoff between the ability of a classifier to correctly identify positive cases and the number of negative cases that are incorrectly classified. Another interesting feature of these curves is that the area beneath them can be used as measure of accuracy in many applications (Swets, 1988). Another way of comparing ROC points is by using a formula that equates accuracy with the Euclidean distance from the perfect classifier, point (0,1) on the graph.

It is necessary to evaluate the performance of our classifiers with all the proteins of the families as queries. By doing that, we obtained 13 curves for apolipoproteins, 224 for globins, 15 for plastocyanins, 18 for RBPs, and 8 for thioredoxins. Each curve was a result of retrieving the proteins of a specific family (among the above) from the database of 187 proteins from PDB Select. We produced average curves with the standard errors for each protein family (Figures 12-16).

The values of the areas beneath each average curve and the false-positive rate and true-positive rate for the better cases, that is, the points of the curves that present the smaller Euclidean distances to point (0, 1) were calculated (Table 3). These rates are the ones obtained with the threshold that presented the best trade-off between the false-positive and true-positive rates.

We can see that it is possible to analyze the structural similarity between proteins and also to classify them using only the information about chemical interactions between their residues (Table 3). We used five protein families of quite different topologies mixed with all the monomers of PDB Select, which is a representative subset of the PDB, and we were always able to identify the families with an average precision above 80%.

The IR approach gave slightly better results than CBIR for almost all the protein families. The rightness probabilities (areas beneath the curves) were about 0.63% higher on average, and the standard errors were also smaller.

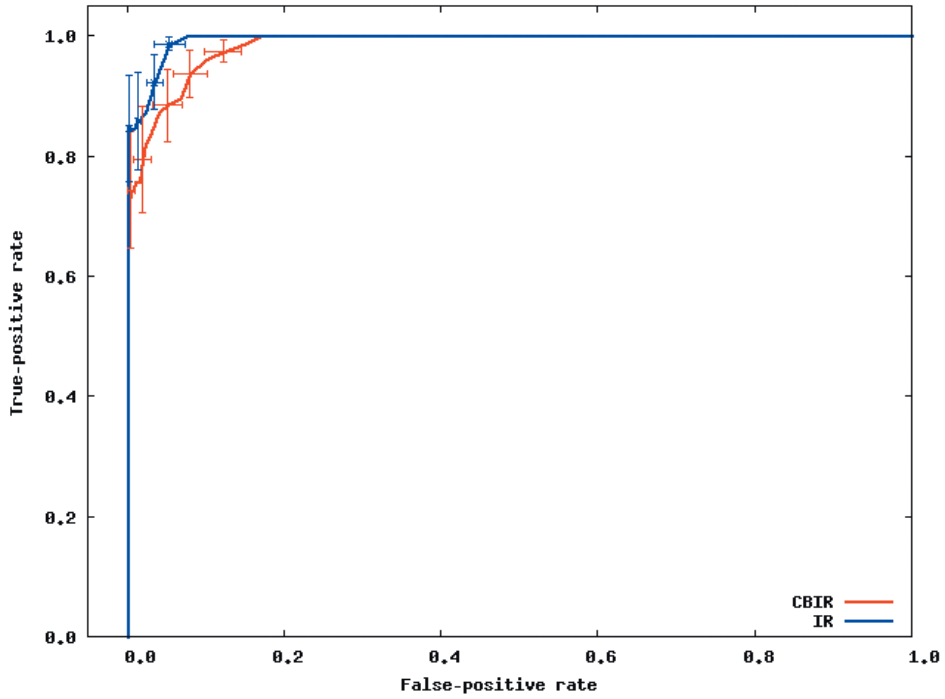


**Figure 12.** Receiver operating characteristic curve for content-based image retrieval (CBIR) and image registration (IR) for the retrieval of 13 apolipoproteins among the 187 different proteins.

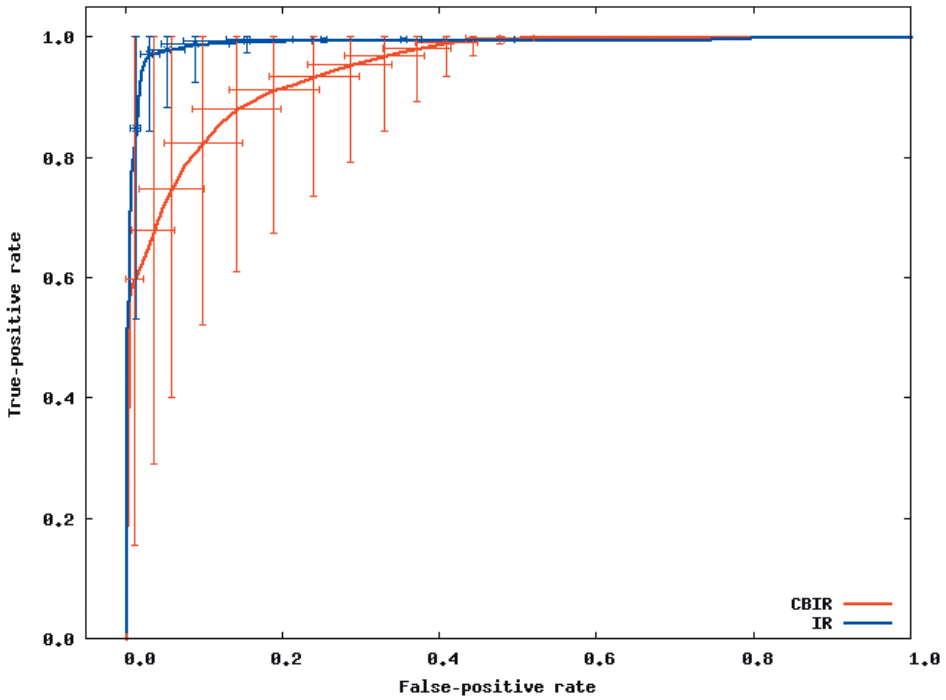


**Figure 13.** Receiver operating characteristic curve for content-based image retrieval (CBIR) and image registration (IR) approaches for the retrieval of 224 globins among the 187 different proteins.

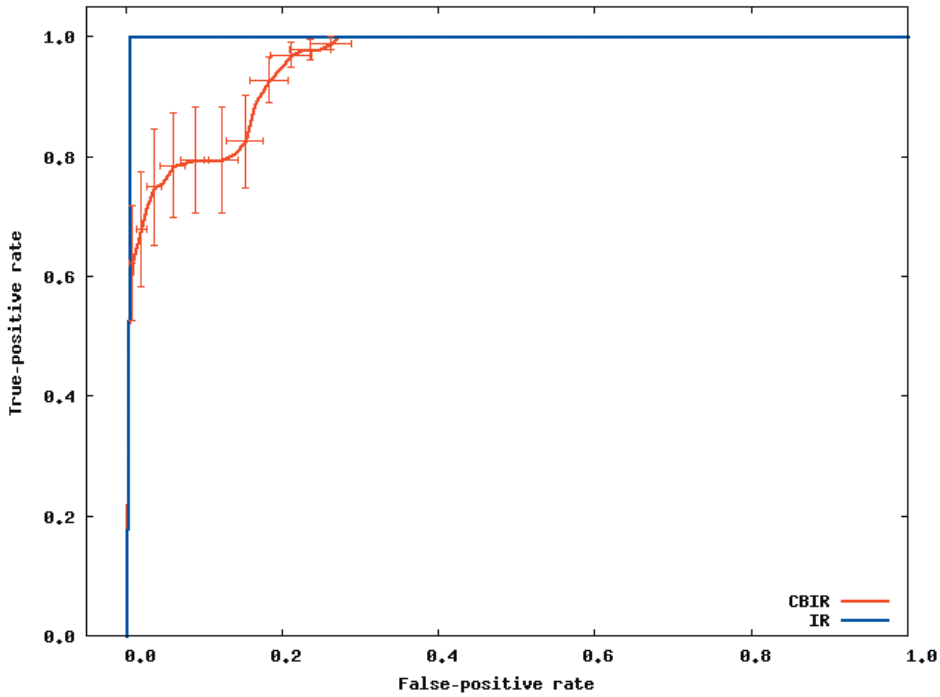


**Figure 14.** Receiver operating characteristic curve for content-based image retrieval (CBIR) and image registration (IR) approaches for the retrieval of 15 plastocyanins among the 187 different proteins.

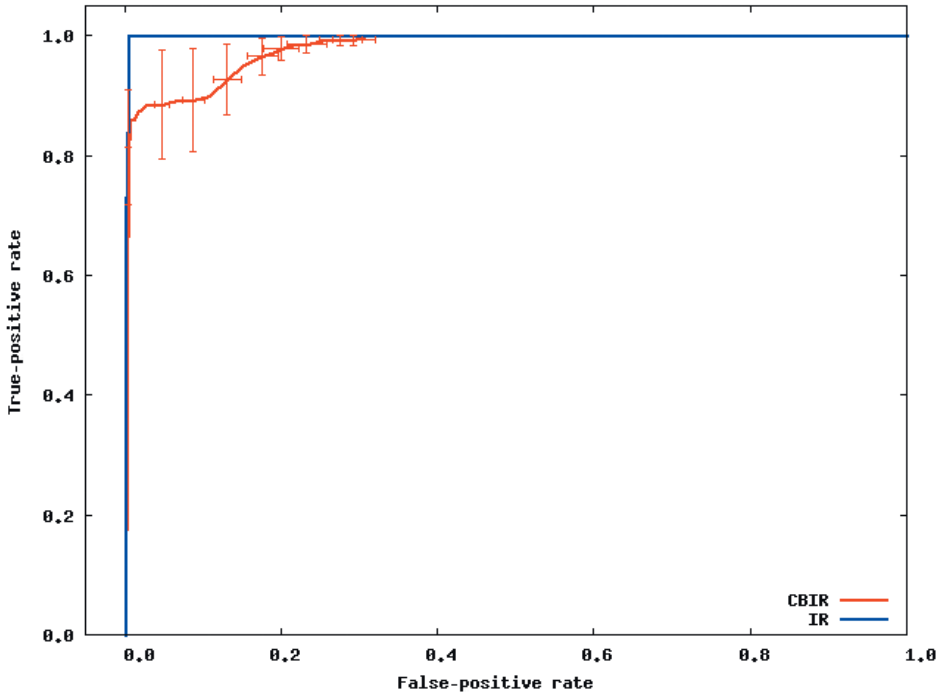


**Figure 15.** Receiver operating characteristic curve for content-based image retrieval (CBIR) and image registration (IR) approaches for the retrieval of 18 retinol-binding proteins among the 187 different proteins.

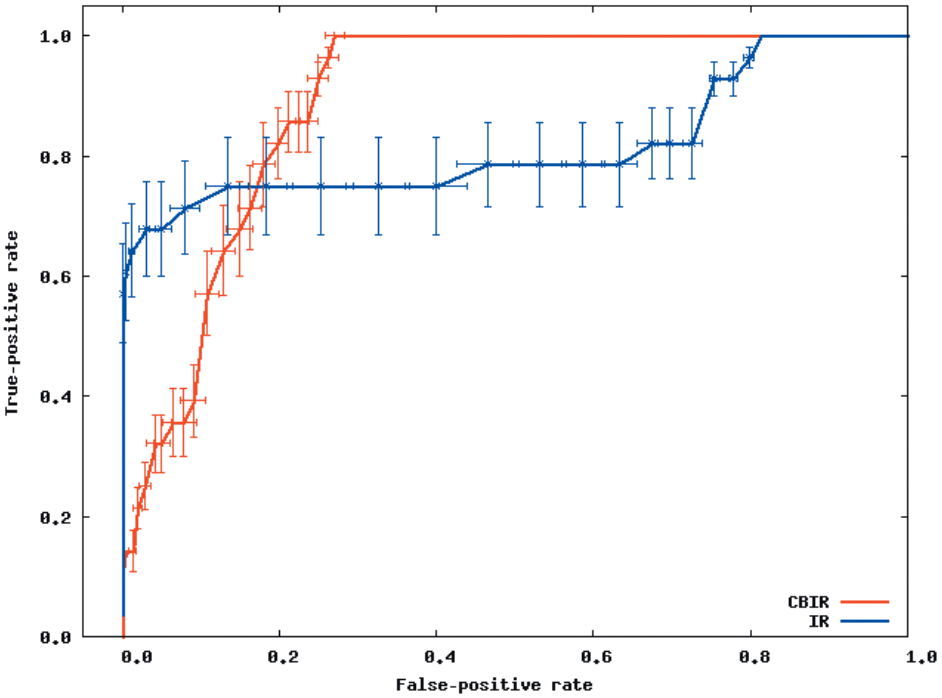


**Figure 16.** Receiver operating characteristic curve for content-based image retrieval (CBIR) and image registration (IR) approaches for the retrieval of 8 thioredoxins among the 187 different proteins.

**Table 3.** Area under receiver operating characteristic curves, false- and true-positive rates for the experiments with apolipoproteins, globins, plastocyanins, retinol-binding proteins (RBPs), and thioredoxins.

| | Apolipoproteins | | | Globins | | | Plastocyanins | | | RBPs | | | Thioredoxins | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Area (%) | FPR (%) | TPR (%) | Area (%) | FPR (%) | TPR (%) | Area (%) | FPR (%) | TPR (%) | Area (%) | FPR (%) | TPR (%) | Area (%) | FPR (%) | TPR (%) |
| IR | 99.36 | 5.41 | 98.72 | 99.00 | 3.15 | 97.08 | 99.68 | 0.53 | 100.00 | 99.84 | 0.54 | 100.00 | 81.69 | 13.36 | 75.00 |
| CBIR | 98.44 | 8.06 | 93.59 | 94.86 | 14.22 | 88.08 | 95.63 | 18.35 | 92.82 | 97.91 | 2.76 | 88.56 | 89.12 | 21.22 | 85.71 |

IR = image registration; CBIR = content-based image retrieval; FPR = false-positive rate; TPR = true-positive rate.

The protein families that gave the best results in classification also had the smallest structural deviations. This was the case for plastocyanins and RBPs. The thioredoxins gave the worse results, and we can see that there are important structural deviations in these proteins. These deviations can add, move or delete some of the inter-residue contacts. Some examples of alignments are presented in Figures 17-21.

## Evaluation of the dissimilarity measure through structural alignments

To analyze the reliability of the dissimilarity indexes of our proposed algorithms, we

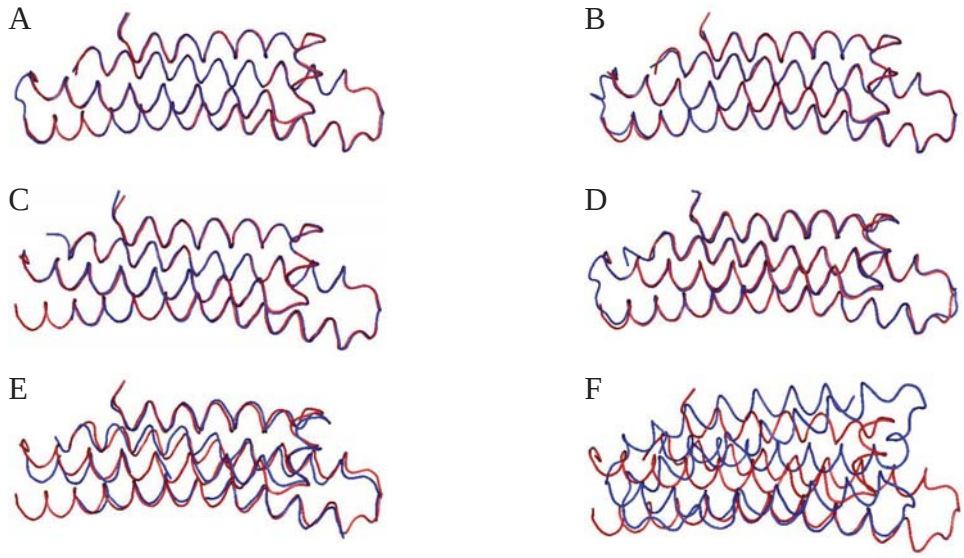


**Figure 17.** Human apolipoprotein (in red, PDB identified as 1b68) and some of the other apolipoproteins used in our tests (in blue). The blue proteins are ordered by the dissimilarity indexes in a way that the first one (A) is the most similar to 1b68. **A.** 1ea8 - 1st place in the classification rank. **B.** 1le4 - 3rd place in the classification rank. **C.** 1nfn - 5th place in the classification rank. **D.** 1gs9 - 7th place in the classification rank. **E.** 1or2 - 9th place in the classification rank. **F.** 1aep - 12th place in the classification rank.

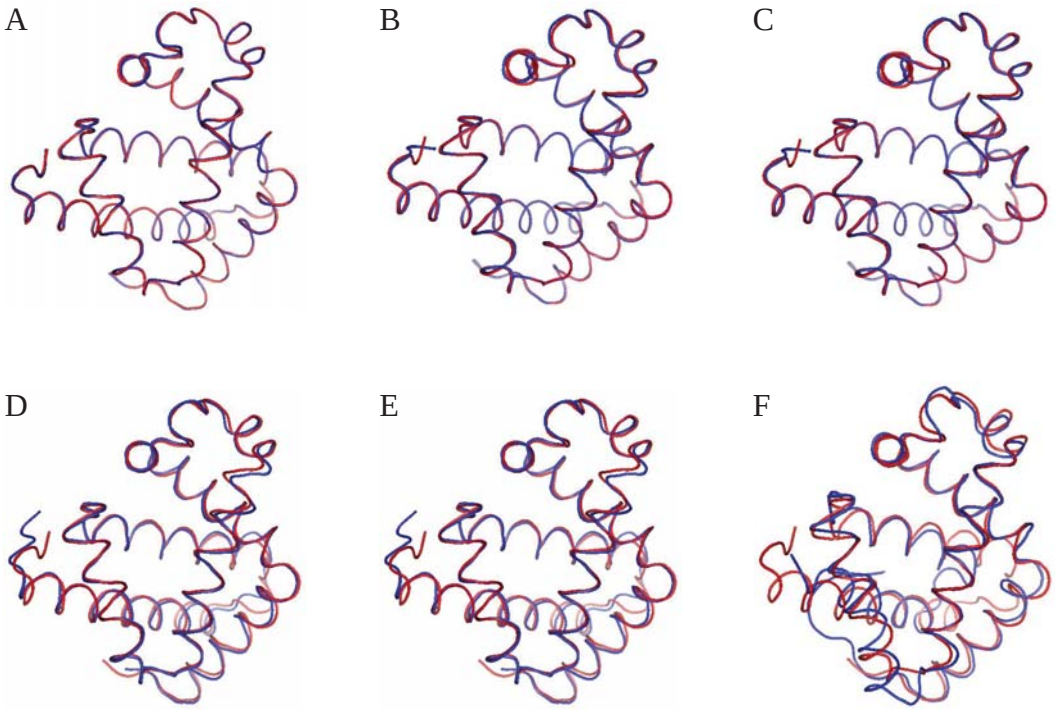


**Figure 18.** Whale globin (in red, PDB identified as 101m) and some of the other globins used in our tests (in blue). **A.** 101m - 1st place in the classification rank. **B.** 2spo - 33rd place in the classification rank. **C.** 1dxc - 65th place in the classification rank. **D.** 1hsy - 129th place in the classification rank. **E.** 1ymc - 161th place in the classification rank. **F.** 2lhb - 221st place in the classification rank.

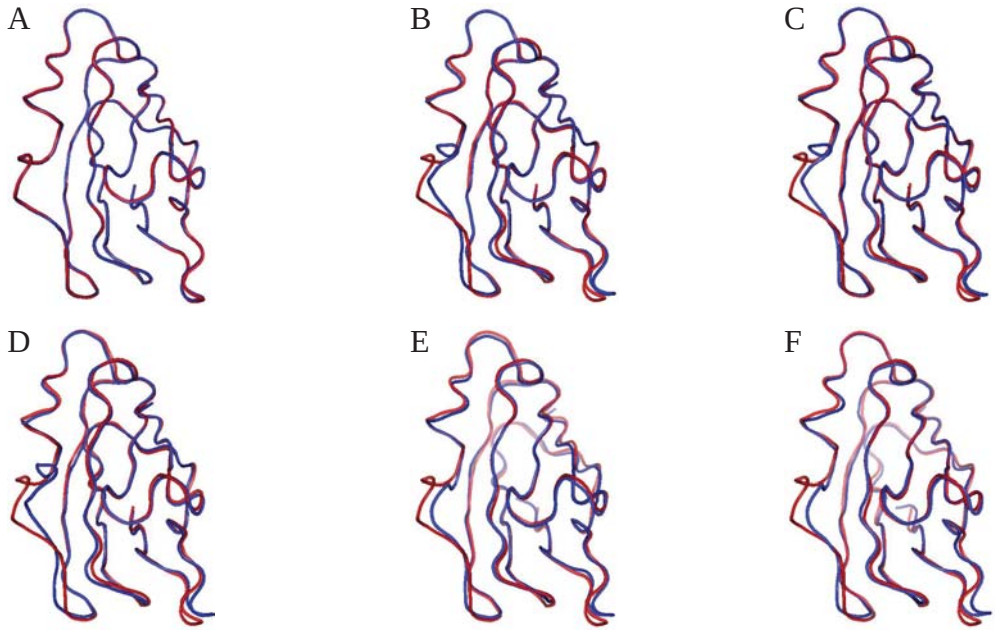


**Figure 19.** Spinach plastocyanin (in red, PDB identified as 1ag6) and some of the other plastocyanins used in our tests (in blue). **A.** 1oow - 1st place in the classification rank. **B.** 2pcy - 3rd place in the classification rank. **C.** 4pcy - 5th place in the classification rank. **D.** 1pnd - 9th place in the classification rank. **E.** 2plt - 11th place in the classification rank. **F.** 1iug - 14th place in the classification rank.

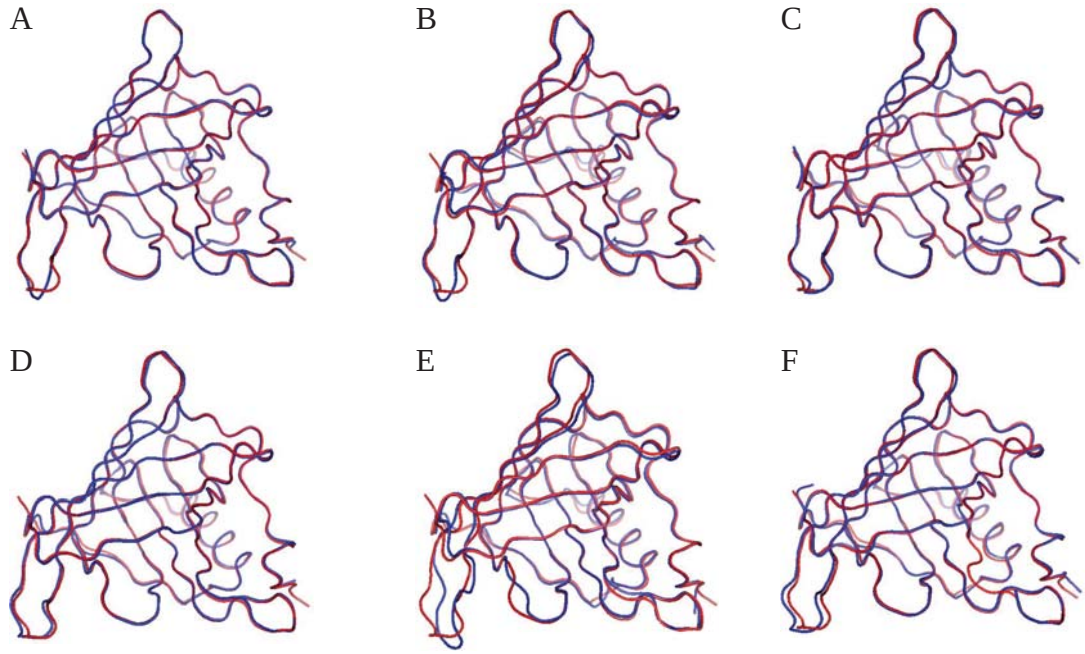


**Figure 20.** Pig retinol-binding protein (in red, PDB identified as 1aqb) and some of the other retinol-binding proteins used in our tests (in blue). **A.** 1iiu - 1st place in the classification rank. **B.** 1brp - 3rd place in the classification rank. **C.** 1kt5 - 7th place in the classification rank. **D.** 1erb - 11th place in the classification rank. **E.** 1jyd - 13th place in the classification rank. **F.** 1fem - 17th place in the classification rank.

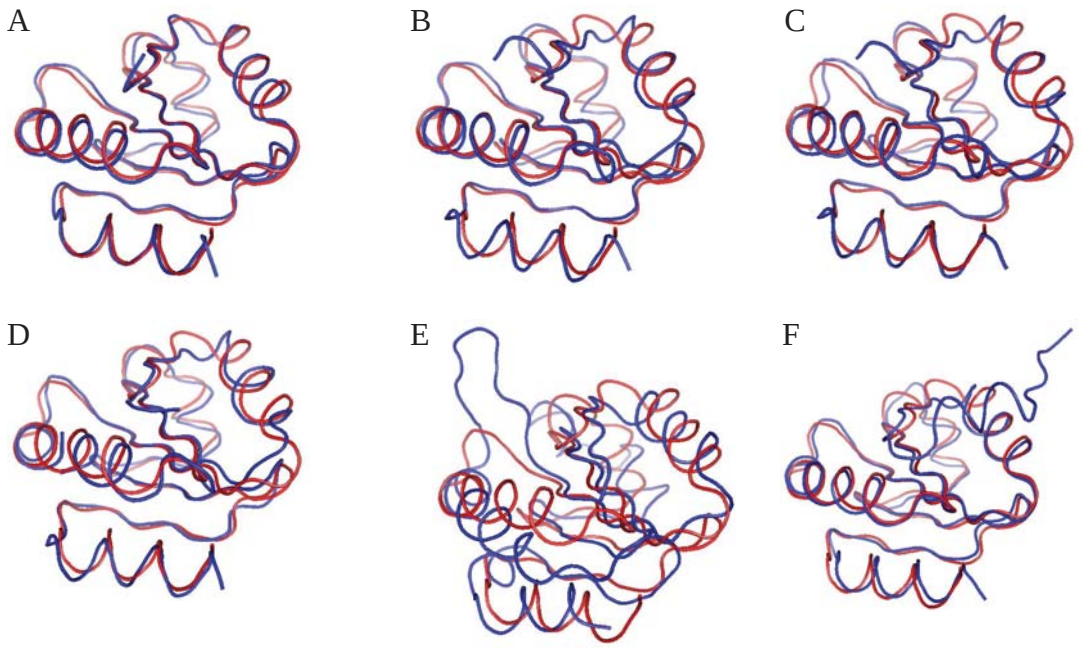


**Figure 21.** Human thioredoxin (in red, PDB identified as 1aiu) and some of the other retinol-binding proteins used in our tests (in blue). **A.** 1gh2 - 1st place in the classification rank. **B.** 2tir - 2nd place in the classification rank. **C.** 1tho - 3rd place in the classification rank. **D.** 1thx - 4th place in the classification rank. **E.** 1wou - 5th place in the classification rank. **F.** 1faa - 6th place in the classification rank.

compared them to the deviations in the three-dimensional structures of the proteins. In Figures 17-21, we can see the structural alignments of the proteins used as queries (shown in red) and some of the other proteins of the same family (in blue).

Through these images, we can see that the proteins that were best classified are really much closer to the query than the others structurally. Thus, we believe that the dissimilarity indexes of the developed algorithms are appropriate to describe the similarity of protein structures, even when they are very similar.

## CONCLUSIONS AND FUTURE RESEARCH

We developed a methodology to analyze protein structure similarity. We used image-matching techniques to retrieve the contact map images of proteins with similar chemical interaction patterns, given an image database and a contact map image as the query. We implemented the Color Correlogram proposed by Huang et al., 1997 and proposed an algorithm-based IR to evaluate our methodology.

Based on experimental analyses, this methodology is appropriate for the similarity analysis of a contact map image database. As each protein structure has a unique and specific pattern of contact, the algorithms can distinguish between contact maps with different protein structures. We analyzed the accuracy of this system and found that it produced good results with apolipoproteins, globins, plastocyanins, RBPs, and thioredoxins as queries. We also identified the minimum set of interactions that exists in each protein family and its role in folding.

We now intend to analyze other types of chemical interactions in the image database, such as cysteine bridges, aromatic stacking, and disulfide bonds. We believe that including more information in the images will improve the precision of the classification. We also intend to evaluate the importance of each type of interaction in the analysis.

## ACKNOWLEDGMENTS

The authors are grateful to Brazilian CNPq and CAPES agencies for the support of this study.

## REFERENCES

Berman HM, Westbrook J, Feng Z, Gilliland G et al. (2000). The protein data bank. *Nucleic Acids Res.* 28: 235-242.

Branden C and Tooze J (1999). Introduction to protein structure. Garland Publishing, Inc., New York, USA.

Brown LG (1992). A survey of image registration techniques. *ACM Comput. Surv.* 24: 325-376.

Carr B, Hart WE, Krasnogor N, Burke EK et al. (2002). Alignment of protein structures with a memetic evolutionary algorithm. Proceedings of the Genetic and Evolutionary Computation Conference, Morgan Kaufman, San Francisco, CA, USA.

Chew LP and Kedem K (2002). Finding the consensus shape for a protein family. Proceedings of the 18th ACM Symposium on Computational Geometry, Barcelona, Spain, June 5-7, pp. 64-73.

Del Bimbo A (1999). Visual information retrieval. Morgan Kaufmann, San Francisco, CA, USA.

Gan HH, Perlow RA, Roy S, Ko J et al. (2002). Analysis of protein sequence/structure similarity relationships. *Biophys. J.* 83: 2781-2791.

Hobohm U and Sander C (1994). Enlarged representative set of protein structures. *Protein Sci.* 3: 522-524.

Hobohm U, Scharf M, Schneider R and Sander C (1992). Selection of a representative set of structures from the Brookhaven Protein Data bank. *Protein Sci.* 1: 409-417.

Huang J, Kumar SR, Mitra M, Zhu WJ et al. (1997). Image indexing using color correlograms. Computer Vision and Pattern Recognition (CVPR 97), San Juan, Puerto Rico, June 17-19, pp. 762-768.

Kutulakos KN (2000). Approximate n-view stereo. Proceedings of the European Conference on Computer Vision, Dublin, Ireland, June 26-July 1, pp. 67-83.

Lancia G, Carr R, Walenz B and Istrail S (2001). 101 optimal PDB structure alignments: a branch-and-cut algorithm for the maximum contact map overlap problem. Proceedings of the Annual Conference on Research in Computational Molecular Biology, Montreal, Canada, April 22-25, pp. 193-202.

Maintz JBA and Viergever MA (1998). A survey of medical image registration. *Med. Image Anal.* 2: 1-36.

Martin ACR (2000). The ups and downs of protein topology; rapid comparison of protein structure. *Protein Eng. Des. Sel.* 13: 829-837.

Mojsilovic A, Gomes J and Rogowitz B (2004). Semantic-friendly indexing and quering of images based on the extraction of the objective semantic cues. *Int. J. Comput. Vision* 56: 79-107.

Murzin AG, Brenner SE, Hubbard T and Chothia C (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247: 536-540.

Neshich G, Togawa RC, Mancini AL, Kuser PR et al. (2003). STING millennium: A web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Res.* 31: 3386-3392.

Orengo CA and Taylor WR (1996). SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* 266: 617-635.

Orengo CA, Michie AD, Jones S, Jones DT et al. (1997). CATH - a hierarchic classification of protein domain structures. *Structure* 5: 1093-1108.

Pentland A, Picard R and Sclaroff S (1994). Photobook: content-based manipulation of image databases. Proceedings of the SPIE, Vol. 2185, Storage and Retrieval for Image and Video Databases, San Jose, CA, USA, February 6-10, pp. 34-37.

Provost F and Kohavi R (1998). On applied research in machine learning. *Mach. Learn.* 30: 127-132.

Swets JA (1988). Measuring the accuracy of diagnostic systems. *Science* 240: 1285-1293.

Westhead DR, Hatton C, Gilbert DR and Thornton JM (1998). A www site devoted to protein structural topology. *Trends Biochem. Sci.* 23: 35-36.