Exploring protein architecture using 3d shape-based signatures

Paquet, Eric; Viktor, H.L.

Publisher's version / Version de l'éditeur:

Proceedings of the 29th Annual International Conference of the IEEE Engineering in medicine and Biology Society (ECMB) in Conjunction with the Biennial Conference of the French Society of Biological and Medical Engineering

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Exploring Protein Architecture using 3D Shape-Based Signatures *

Paquet, E., Viktor, H.L.
August 2007

* published in the Proceedings of the 29th Annual International Conference of the IEEE Engineering in medicine and Biology Society (ECMB) in Conjunction with the Biennial Conference of the French Society of Biological and Medical Engineering (SFGBM). August 2007. NRC 49828.

Canada

Exploring Protein Architecture using 3D Shape-based Signatures

Eric Paquet and Herna L Viktor, *Member, IEEE*

Abstract— **Consider the scenario where, for a prescription drug designed to treat a terminal illness, a particular protein has been successfully identified as a crucial, beneficial component in the drug compound. However, this protein has contra-indications and causes severe adverse effects in a certain subset of the population. If another protein from the same family, with similar structure and functionality, but without these adverse effects, can be found, the subsequent modification of the harmful drug has obvious benefits.**

This paper describes a new indexing and similarity search system to retrieve such protein structure family members, based on their 3D shape. Our approach is translation, scale and rotation invariant, which eliminates the need for prior structure alignment. Our experimental evaluation against seven (7) diverse protein families indicate that our system accurately and precisely locate all members of a family. We further illustrate this by showing that our system precisely retrieves the Homo Sapiens Hemoglobin family members, against a database containing 26,000 protein structures.

I. Introduction

The functional analysis of protein structures is an important research issue in molecular biology, bio-informatics and pharmaceutics. A protein's function is often dependent on the shape and physical properties of the active sites of the molecular surface [1-3]. Current research suggests that, if two proteins have similar active sites, the function of the two proteins may be closely related and they may thus belong to the same family or super-family [4]. This observation is of importance in many domains, including the discovery of new structures and/or mutations, the study of protein interactions, and the implications for drug design.

The identification of family membership thus becomes crucially important. To this end, this paper presents an indexing and retrieval system which utilize the 3D structure of a protein, in order to find the other members of the family a protein structure belongs to. The algorithm is translation, scale and rotation invariant, which eliminates the need for prior structure alignment. The main benefit of using 3D structural indexing is that the protein functionality is related to its 3D shape. 3D shape indexing is a natural way to index the functionality with all the foreseen applications in bioinformatics, genomic as well as for the pharmaceutical industry.

Manuscript received April 2, 2007. Eric Paquet is with the National Research Council of Canada, Ottawa, ON, K1A 0R6, Canada (phone: 613-991-5035; fax: 613-952-0215; e-mail: eric.paquet@ nrc-cnrc.gc.ca).

Herna L Viktor is with the University of Ottawa, ON, K1N 6N5, Canada (email: hlviktor@site.uottawa.ca).

Our results against more than 26,000 protein structures as contained in the Protein Data Bank shows that our system is able to accurately and efficiently retrieve protein families using a "query by prototype" approach, where a family member acts as the "seed". In this way, domain experts are aided in the task of labelling new structures effectively, finding the families of existing proteins, identifying mutations and unexpected evolutions.

This paper is organized as follows. The next section provides an overview of similarity search and its application in the protein structure domain. In Section 3, we present our 3D indexing and retrieval system. This is followed, in Section 4, with an experimental evaluation. Section 5 concludes the paper.

II. Background

It is estimated that the amount of newly discovered protein structures will growth linearly [1]. There are currently more than 45,000 known structures in the Protein Data Bank, and an additional 100 are added every week [5]. Ideally, new structures should be labelled by a domain expert. However, given the rate in which molecular biology technologies develop, this ideal unfortunately becomes unrealistic.

Recently, a number of research teams have focused their attention on the use similarity search for protein structure retrieval, mainly using structure alignment [5]. For example, the approach followed by [4], uses a local approach to calculate the similarities between proteins. However, a drawback of their method is that they accumulate error and that it does not scale well. Similarly, [2, 8] and [9] uses a local approach to find the similarity between aligned structures: thus loosing semantic information about the interrelationship of the substructures on the protein. In their research, reference [10] uses a shape-based approach in the form of a sphere, grid or pie in order to compare structures. Reference [3] also employs a spherical approach, and uses a weighted distance measure to determine similarity.

The next section introduces our algorithm for the 3D indexing and similarity search of protein structures.

III. Shape-based 3D Indexing and Searching

The 3D shape on a protein structure provides us with a high level of detail regarding the functionality of protein, which cannot be inferred when considering the sequence alone [1]. This is especially relevant when aiming to identify family membership based on the similarity of their

structures [4]. The family membership should be identified without having to super-impose or precisely align structures to see how much they overlap. Otherwise, the process becomes time consuming and tedious.

Our objective is thus to define an index (or signature) that describes a protein from a three-dimensional shape point of view and that is translation, scale and rotation invariant. The later invariants are essential because the protein can have an arbitrary location and pose in space.

That is, for the 3D structures, the signatures are shape-based, and the proteins with a shape-based distance closest to each other are considered similar. Figure 1 depicts our algorithm, which can be described as follows. Firstly, the protein structure is triangulated into a mesh. Next, the centre of mass of the protein is calculated and the coordinates of its vertices are normalised relatively to the position of its centre of mass. A translation invariant representation is then achieved. Translation invariance is important since a priori we do not know the location of the protein.

Then, the tensor of inertia of the object is calculated, using the following formula

$$I = \left[I_{qr}\right] = \left[\frac{1}{n}\sum_{i=1}^{n}\left[S_i\left(q_i - q_{CM}\right)\left(r_i - r_{CM}\right)\right]\right] \quad (1)$$

where S_i is the area of the i^{th} triangle; q_i is the coordinates x, y or z of the i^{th} triangle and r_i is the coordinate x, y or z of the i^{th} triangle; and q_{CM} and r_{CM} are the coordinates x, y or z of the barycentre. This tensor results in a 3 x 3 matrix.

In order to take into account the tessellation in the computation of these quantities, we do not utilise the vertices per se but the centres of mass of the corresponding triangles; the so-called tri-centres. In all subsequent calculations, the coordinates of each tri-centre are weighted with the area of their corresponding triangle. The later is being normalised by the total area of the protein, i.e. with the sum of the area of all triangles. In this way, the calculation can be made robust against tessellation, which means that the index is not dependent on the method by which the protein was virtualised: a "sine qua non" condition for real world applications. Under certain assumptions, the area of the triangles is related to the local curvature: the smaller the area, the higher the curvature. Such a hypothesis is valid if the number of acquired points is related to the complexity of the local structure and constitutes a sine qua non condition for any realistic shape acquisition.

To be able to achieve rotation invariance, the Eigen vectors of the tensor of inertia are calculated. The Jacobi method, which has been proven successful for real symmetric matrices, is used. Once normalised, the unit vectors define a unique reference frame, which is independent on the pose and the scale of the corresponding protein: the so-called Eigen frame. The unit vectors are identified by their corresponding Eigen values. It is very common to encounter axes which are orientated along the shortest and the longest dimensions of the protein. For instance, let us suppose we have an ellipsoidal protein: in that particular case, one axis would correspond to the revolution symmetry of the protein and the other two axes would correspond to the minor and major axis of the protein.

Algorithm **Calculate3DIndexDescriptor**

Input: A *proteinFile* from the *ProteinDataBankDatabases*
Output: The *3DIndexDescriptor* of *proteinFile*
1. read(*proteinFile*);
2. *triangulatedProtein*=triangulate(*proteinFile*);
3. **for each** *triangle* in *triangulatedProtein* **do**
   triCentres.**add**(computeTriCentre(*triangle*));
   weights.**add**(computeWeight(*triangle*));
4. *barycentre*=computeBarycentre(*triCentres*);
5. *tensorInertia*=computeTensorInertia(*barycentre*, *triCentres*, *weights*);
6. *principalComponents*=jacobi(*tensorInertia*);
7. sortByPrincipalValues(*principalComponents*);
8. **for each** *triCentre* **in** *triCentres* **do**
   cords.**add**(computeCord(*barycentre*, *triCentre, principalComponents*));
9. **for each** *cord* **in** *cords* **do**
   angles.**add**(computeAngles(*cord*, *principalComponents*));
   radii.**add**(computeRadii(*cord*));
10. *histogramsAngles*=computeHistogramsAngles(*angles*);
11. *histogramRadii*=computeHistogramsRadii(*radii*); (14)
12. normalisedAnglesHistograms(*histogramsAngles*);
13. normalisedRadiiHistogram(*histogramRadii*);
14. write(*histogramsAngles*);
15. write(*histogramRadii*);
16. Return *3DIndexDescriptor*
End **Calculate3DIndexDescriptor**

Fig. 1. The Calculate3DIndexDescriptor algorithm

The descriptor is based on the concept of a cord. A cord is a vector that originates from the centre of mass of the protein and that terminates on a given tri-centre. The coordinates of the cords are calculated in the Eigen reference frame in cosine coordinates. The cosine coordinates consist of two cosine directions and a spherical radius. The cosine directions are defined in relation with the two unit vectors associated with the smallest Eigen values i.e. the direction along witch the protein presents the maximum spatial extension. In other words, the cosine directions are the angles between the cords and the unit vectors. The radius of the cords are normalised relatively to the median distance in between the tri-centres and the centre of mass in order to be scale invariant. It should be noticed that the normalisation is not performed relatively to the maximum distance in between the tri-centres and the centre of mass in order to achieve robustness against outliers or extraordinary tri-centres. From that point of view, the median is more efficient that the average. The cords are also weighted in terms of the area of the corresponding triangles;

the later being normalised in terms of the total area of the protein.

The statistical distribution of the cords is described in terms of three histograms: one histogram for the radial distribution and two for the angular distribution of the cords. That is, the first histogram described the distribution of the cosine directions associated to the unit vector associated with the smallest Eigen value. The second histogram described the distribution of the cosine directions associated with the unit vector associated with the second smallest Eigen value. The third histogram described the distribution of the normalised spherical radius as defined in the previous paragraph.

The ensemble of the three histograms constitutes the shape index of the corresponding protein structure, which is placed in database PB3D to be used when querying the Protein Data Bank, in order to locate family members.

IV. Experimental evaluation

This section describes our experimental evaluation of our system. We implemented the system using Java and ran the experiments on workstations with two 3.4 GHz CPUs and 2.8 GB of RAM. We used a total of 26,000 protein structures in our experiments. These protein structures were taken from the Protein Data Bank and we used the family classification information as contained in SCOP database to verify our findings [1].

A. 3D Retrieval of Protein Structures

In our experiments, our objective was to determine the effectiveness of our system when aiming to obtain all protein structures within a particular family. The nearest-neighbourhood approach is used to identify such families, using a "query by prototype" approach where the query protein structure is used as a seed. Here, we used the Euclidian distance as measure of similarity.

Table 1 shows the protein families used in our first set of experiments. The table indicates the family, the number of family members and the protein structure used as seed query, together with the results obtained by our system, indicated by the number of nearest neighbours within the family (FNN). The table shows that our system was able to obtain accurate and precise results, when exploring a database of 26,000 protein structures. For example, for the Pertussis toxin s2/s3, Flourescent protein, GluRS and Sex hormone-binding globulin, all of the *n* family members were retrieved first, i.e. an accuracy of 100% was obtained and the family formed the *n* nearest neighbours. Similarly, for the Ligand-gated protein channel and Pyridoxine 5'-Phosphate synthase, a total of *n-1* family member were the seed's nearest neighbours, with accuracies of 88.8% and 85.7%, respectively.

Next, we focused our attention on the Homo Sapiens Hemoglobin protein structure family. This family contains 162 members of which 95 are present in our database.

TABLE I
PROTEIN FAMILIES AND THEIR RETRIEVAL ACCURACIES

Family	Members (#)	Seed Query	FNN
Pertussis toxin S2/S3	3	`1bcp`	3 (100%)
Fluorescent protein	2	`1ggx`	2 (100%)
GluRS	6	`1n78`	6 (100%)
Ligand-gated Protein Channel	9	`1qfg`	8 (88.8%)
Pyridoxine 5'-Phosphate Synthase	7	`1bcp`	6 (85.7%)
Sex hormone-binding Globulin	8	`1f5f`	8 (100%)

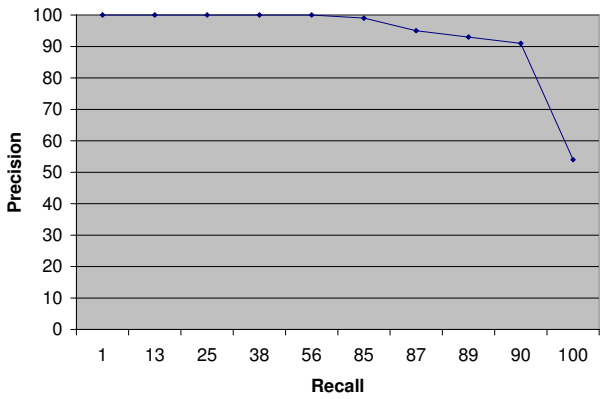

Fig. 2. Precision-recall curve (in percentage) for the Homo Sapiens Hemoglobin family, for the `1rly` structure as seed query

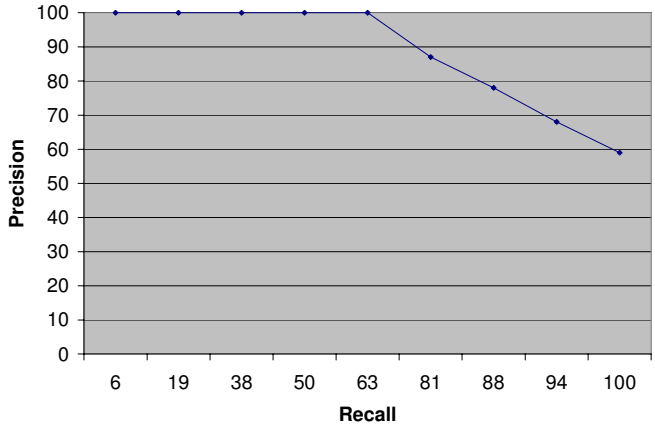

Fig. 3. Precision-recall curve (in percentage) for the Homo Sapiens

Hemoglobin family, for the `1lfg` structure as seed query.

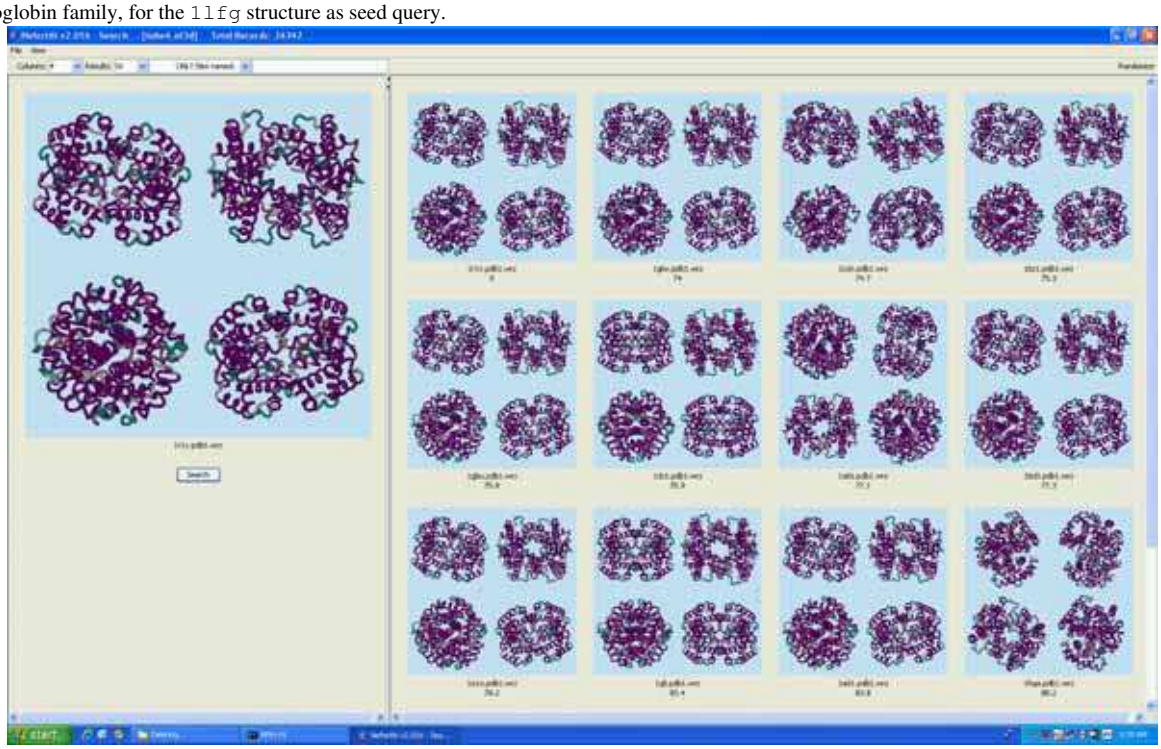

Fig. 4. Retrieval of the first 12 family members of the Homo Sapiens Hemoglobin family within 77 protein structures, using the `1rly` protein structure as seed query.

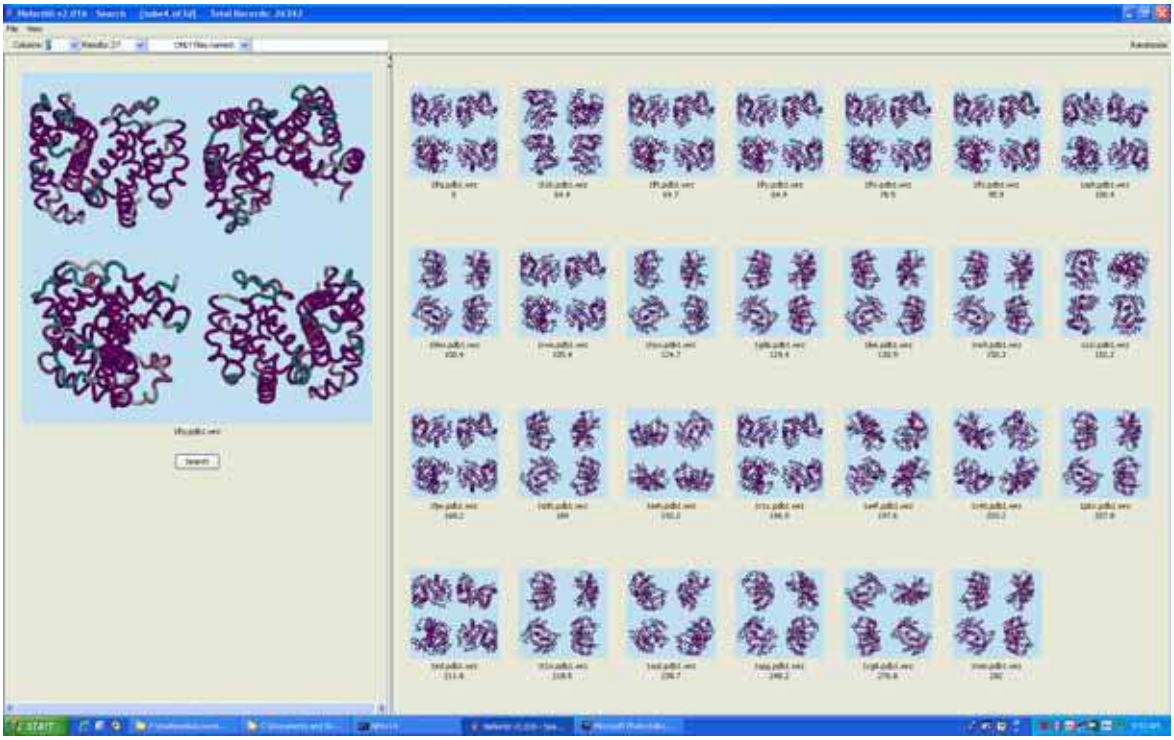

Fig. 5. Retrieval of all 16 members of the Homo Sapiens Hemoglobin family within 27 protein structures, using the `1flg` structure as seed query, with precision 59% and recall 100%.

When employing our 3D indexing and similarity search system with the lrly protein as the query structure, the first 55 of the similar structures retrieved belong to the Homo Sapiens Hemoglobin protein structure family. This result is shown in Figure 4. Our query results are highly precise, with 86 of the first 100 structures retrieved are members of this family, i.e. an accuracy of 86% was obtained. Further analysis of our results indicates that, those protein structures which are not of the Homo Sapiens family, are also Hemoglobin protein structures. Namely, the structure located at position 56 from the query structure is a Hemoglobin structure of a cow, the next incorrectly retrieved structure is the Hemoglobin of a rookcod (position 69), a chicken (position 71) and so on.

This result convinces that we were able to retrieve similar protein structures within families, and that our system succeeds in finding the relevant groupings within the data.

In general, Hemoglobin protein structures may be grouped, by visual inspection, into two distinct formats, as can be seen from Figures 4 and 5. Our second query thus used the 1lfg protein structure and subsequently located all 16 members of this subset of the Homo Sapiens Hemoglobin, again against more than 26,000 protein structures.

The precision-recall curves for these two queries are shown in Figures 2 and 3. Here, the precision refers to the number of structures retrieved that is relevant, divided by the total number of structures that are retrieved. The recall denotes the number of relevant protein structures retrieved, divided by the total number of structures that are relevant.

For example, for the lrly protein, our database of 26,000 tuples contains 79 members of this "sub-family". This first 44 structures retrieved by our system all belonged to the Homo Sapiens Hemoglobin protein structure family, with a precision of 100% (44/44) and a recall of 56% (44/79). Furthermore, all family members were retrieved within the first 147 structures, giving us a precision of 54% (79/147) and recall of 100% (79/79). That is, we were able to retrieve all family members by considering less than twice the number of structures, from a database containing 26,000 members.

We obtain similar results for the 16 member "sub-family" when using the 1lfg protein structures as query, as shown in Figures 3 and 5. In this case, the first ten (10) protein structures retrieved all belonged to this subfamily, with a precision of 100% and a recall of 63% (10/10). We are able to obtain the 16 relevant structures within the first 27 query results, giving up as recall of 59% (16/27) and precision of 100% (16/16).

In summary, our system not only achieves a high precision, but also finds the most pertinent results with a minimum of outliers. This implies that the most pertinent results can be obtained with a minimum number of queries. Namely, for the results presented in this paper, we were able to retrieve structure belonging to the same family within a database containing 26,000 structures, using only one (1) or two (2) queries at most. Importantly, our system is very fast, providing us with query results in less than one second. Both visual inspection and the experimental evaluation of our results, as shown in this paper, show that we are able to retrieve family members accurately, and effectively.

I. Conclusion

Molecular biologist and other role players must be able to obtain fast, relevant information about the family membership of existing and new protein structures. Through the use of 3D-shape based similarity search, domain experts can verify whether a structure is a new member of a family, if it possibly represent an unexpected evolution, and/or indicate a new breakthrough in domains such as drug design and protein functionality discovery.

The paper presented a system for indexing protein structure based on the 3D structure. The 3D shape-based index is translation, scale and rotation invariant, thus enabling us to compare structures which are not aligned. Our experimental results show that we are able to accurately find families of proteins from a very large database. Also, these families are found very fast, making our 3D shape-based search approach applicable for wide-spread use.

References

[1] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C., SCOP: A Structural Classification of Proteins Database of the Investigation of Sequences and Structures, Journal of Molecular Biology, Volume 247, 1995, 536-540.

[2] Ohkawa, T., Hirayama, S. and Nakamura, H. A Method of Comparing Protein Structures Based on Matrix Representation of Secondary Structure Pairwise Topology, In Proceedings of the International Conference on Information Intelligence and Systems (Bethesda, MD, USA), 1999, 10-15.

[3] Ankers, M., Kastenmuller, G., Kriegel, H-P and Siedi, T. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB99) (Heidelberg, Germany), AAAI Press, 1999, 34-43.

[4] Park, S.-H., Park, S.-J. and Park, S.H., A Protein Structure Retrieval System Using 3D Edge Histogram, Key Engineering Materials, Vols. 277-279, 2005, 324-330.

[5] Yeh, J.-S., Chen, D.-Y. and Ouhyoung, M., A Web-based Protein Retrieval System by Matching Visual Similarity, Bioinformatics, Vol. 21, no 13, 2005, 3056-3057.

[6] Ohkawa, T., Nonomura, Y. and Inoue, K., Logical Cluster Construction in a Grid Environment for Similar Protein Retrieval, In Proceeding of the 2004 International Symposium on Applications and the Internet Workshops (SAINTW'04) (Tokyo, Japan), 2004, 5-16.

[7] Chi, P.H., Scott, G. and Shyu, C.-R., A Fast Protein Structure System Using Image-Based Distance Matrices and Multidimensional Index, In Proceeding of the Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04) (Taichung, Taiwan), 2004, 522-532.

[8] Akbar, S., Kung, J. and Wagner, R., Exploiting Geometrical Properties of Protein Similarity Search, In Proceeding of the 17th International Conference on Database and Expert Systems Applications (DEXA'06) (Krakow, Poland), 2006, 228-234.