BIOINFORMATICS APPLICATIONS NOTE

Vol. 21 no. 13 2005, pages 3056–3057
doi:10.1093/bioinformatics/bti458

Structural bioinformatics

A web-based three-dimensional protein retrieval system by matching visual similarity

Jeng-Sheng Yeh[1,*], Ding-Yun Chen[1], Bing-Yu Chen[2] and Ming Ouhyoung[1,3]

[1]Department of Computer Science and Information Engineering, [2]Department of Information Management and [3]Graduate Institute of Network and Multimedia, National Taiwan University, Taipei 106, Taiwan

Received on July 10, 2004; revised on January 24, 2005; accepted on April 18, 2005
Advance Access publication April 19, 2005

ABSTRACT

Summary: A web-based three-dimensional (3D) protein retrieval system is available for protein structure data including all PDB and FSSP dataset. In this system, we use a visual-based matching method to compare the protein structure from multiple viewpoints. It takes less than three seconds for each query with 90% accuracy on an average.

Availability: The web-based query interface and downloadable files can be accessed via http://3d.csie.ntu.edu.tw/ProteinRetrieval/

Contact: jsyeh@cmlab.csie.ntu.edu.tw

Supplementary information: Further details of the proposed method are available at http://graphics.csie.ntu.edu.tw/~jsyeh/3Dprotein/

INTRODUCTION

There are more than 25 000 protein structure files in Protein Data Bank (PDB) (Berman *et al*., 2000, http://www.pdb.org/) now, with an additional one hundred added per week. Hence an increasing necessity for protein structural retrieval. Therefore we propose a visual-based method to find the similarity of protein structures automatically, and which can also provide some clues for protein classification.

Several algorithms and servers have been proposed to analyse those protein structures in Protein Data Bank (PDB) in order to help in the prediction of protein functions, as the shape of protein may determine its function. The following tools are mainly based on alignment of primary structure (1D sequence data), secondary structure (helix/sheet) and/or 3D atom coordinates. For instance, EMBL SSM (Krissinel and Henrick, 2003) uses a graph-matching algorithm to map secondary structure elements as a first step to iteratively align atoms. To compare the 3D protein structures, the Dali/FSSP (Holm and Sander, 1998) database has been developed based on exhaustive 3D structure comparison of protein structures currently in PDB. Several image processing-based methods were also proposed for protein structure comparisons (Sandak *et al*., 1995; Chi *et al*., 2004). Shape histogram (Ankerst *et al*., 1999) is used to compare the 3D structure of the surface of proteins. Here, however, we would like to provide an alternative tool based on views instead of atom positions only.

In this paper, we present a visual-based protein retrieval system, which is available on Internet with web-enabled interface. The concept of the visual-based matching method is based on human perception, therefore, the result of retrieval can be used and manipulated more intuitively and quickly. Biologists can receive the ranked results of a given query. The design of user interface is described as follows.

Using our system, the user can specify a PDB ID as an input to query similar protein structures. That is, the proteins that look similar to the query protein will be displayed in terms of visual similarity ranking. The users can also pick one of the results for further query by clicking again. If users want to query by an unpublished protein structure, they can upload the protein structure file in PDB format. The server will calculate the necessary 3D features and make a query.

For output display, users can choose their preference for display. One of the configurations is to display all figures of protein in similarity ranking. Another configuration displays the results with metadata information from PDB files including protein name, EC number and SwissProt ID. This system output can link to other online databases, such as OCA (Prilusky, 2004, http://bip.weizmann.ac.il/oca-bin/ocamain/) and PDBsum (Laskowski, 2001).

METHODS

The proposed method is based on LightField descriptors (Chen *et al*., 2003) to match 3D protein structures in visual-based similarity. The core idea of the multi-view projection method is to compare 3D object with multiple 2D projection views. The retrieval process is divided into off-line feature extraction and on-line protein retrieval.

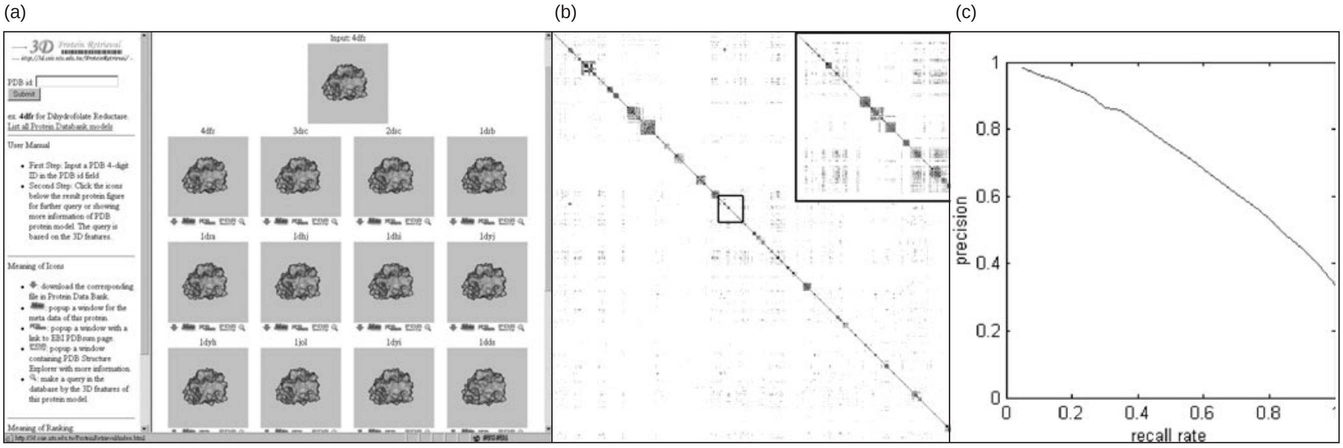
In off-line feature extraction, the projection views are first pre-rendered from the solvent-accessible surface of protein, which is computed by Connolly's msp package (Connolly, 1983). Then the 2D shape Zernike moment descriptors and Fourier descriptors are extracted as features for each projection view. In our system, 100 projection views are rendered around the centre of 3D structure for the visual-based matching.

In on-line protein retrieval, the dissimilarity value of two proteins is calculated by the summation of the distance between descriptors in each corresponding view. In addition, in order to accelerate the matching speed in such a large database, we use iterative algorithms and early rejection of non-relevant models. To iteratively reject non-relevant protein structures, lower frequency parts of Zernike moment descriptors and Fourier descriptors are matched in the initial stage, and higher frequency parts of those descriptors are applied in each stage to refine the top ranked results of retrieval. After iteratively rejecting models stage-by-stage, the entire database (more than 25 000 proteins) can be queried in <3 s in a Pentium 4 2.4 GHz PC. Figure 1a shows a typical example of protein retrieval in the proposed web-based system.

DISCUSSION

In our experiments, the 4997 proteins, which are listed in the first 362 classes (representative sets from 12asA–1bsvA) in the FSSP

[*]To whom correspondence should be addressed.



3D Protein Retrieval

http://3d.csie.ntu.edu.tw/ProteinRetrieval/

PDB id:

Submit

ex. **4dfr** for Dihydrofolate Reductase.

List all Protein Databank models

User Manual

- First Step: Input a PDB 4-digit ID in the PDB id field
- Second Step: Click the icons below the result protein figure for further query or showing more information of PDB protein model. The query is based on the 3D features.

Meaning of Icons

- download the corresponding file in Protein Data Bank.
- popup a window for the meta data of this protein.
- popup a window with a link to EBI PDBsum page.
- popup a window containing PDB Structure Explorer with more information.
- make a query in the database by the 3D features of this protein model.

Meaning of Ranking

http://3d.csie.ntu.edu.tw/ProteinRetrieval/index.html

Fig. 1. (**a**) The query result of our server after submitting the shape of a query protein (4dfr: dihydrofolate reductase). (**b**) The resulting similarity matrix (4997 × 4997) while the intensity of (x, y) shows the dissimilarity value between protein x and protein y, i.e. a darker pixel (x, y) means that protein x and protein y are much similar. The box in the upper-right corner is the enlarged sample of the small box in centre. (**c**) The precision-recall plot: 'given different recall rates (x-axis, 0–100%), plots the precision values (y-axis, 0–100%) of the correct classification.' For comparison purpose, we choose 4997 proteins to retrieve similar shapes to see if proteins with same FSSP class name will be retrieved. Please visit the supplementary web page for further details.

database, updated in October 2001 (Holm and Sander, 1998), are analysed and classified. Every class with only one molecule is skipped. Figure 1b is the similarity matrix, which shows that the proteins with the same FSSP class name will be clustered together. The box in the upper-right corner is the enlarged part of the small box in the centre. The similarity value is the inverse of dissimilarity value, which is the sum of the distances in all the corresponding views. Figure 1c, calculated by psbPlot (Shilane *et al.*, 2004), is the precision-recall plot of these 4997 proteins. We create a query for each protein from the 4997 proteins and plot the recall rate of other proteins having the same FSSP class name as in the 4997 proteins. It shows that our visual-based matching method may provide some useful clues to help biochemists retrieve and analyse protein 3D structure.

Compared with the shape histogram method (Ankerst *et al.*, 1999), the accuracy of nearest neighbour classification derived by using our method is 92.8% (4997 proteins in 362 classes, and actually 25 591 proteins are also tested with similar result.), which is very similar to the 91.6% in Ankerst's method on the previous version of FSSP dataset (3422 proteins in 281 classes).

In http://3d.csie.ntu.edu.tw/ProteinRetrieval/ a full set of 25 120 proteins is available. In http://3d.csie.ntu.edu.tw/ProteinRetrieval/. Note that DNA files in PDB are not included. As for statistics of our web server, there are 1177 accesses from the first prototype (2051 proteins inside) of June 16, 2003 to current release (June 16, 2004). Now the system is extended to 25 120 proteins and synchronized to RCSB PDB weekly.

ACKNOWLEDGEMENT

This project is supported in parts by CIET-NTU (MOE), NSC93-2622-E-002-033, NSC93-2752-E-002-007-PAE, NSC93-2213-E-002-083 and NSC93-2213-E-002-084.

REFERENCES

Ankerst,M. *et al.* (1999) Nearest neighbor classification in 3D protein databases. *Proc. Int. Conf Intell. Syst. Mol. Biol.*, **1999**, 34–43.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Chen,D.-Y. *et al.* (2003) On visual similarity based 3d model retrieval. *Computer Graphics Forum*, **22**, 223–233.

Chi,P.-H., Scott,G. and Shyu,C.-R. (2004) A fast protein structure retrieval system using image-based distance matrices and multidimensional index. In *Proceedings of BIBE 2004*, IEEE Computer Society Press, Taichung, Taiwan, pp. 522–532.

Connolly,M.L. (1983) Solvent-accessible surfaces of proteins and nucleic acid. *Science*, **221**, 709–713.

Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.

Krissinel,E. and Henrick,K. (2003) Protein structure comparison in 3d based on secondary structure matching (ssm) followed by ca alignment, scored by a new structural similarity function. In Andreas,J.K. and Penelope,J.K. (eds), *Proceedings of the 5th International Conference on Molecular Structure Biology*, Austrian Chemical Society, (GoeCH), Biochemistry Subgroup, Vienna, p. 88.

Laskowski,R.A. (2001) PDBsum: summaries and analyses of pdb structures. *Nucleic Acids Res.*, **29**, 221–222.

Prilusky,J. (2004) OCA, a browser-database for structure/function.

Sandak,B. *et al.* (1995) An automated computer vision and robotics-based technique for 3-D flexible biomolecular docking and matching. *Comput. Appl. Biosci.*, **11**, 87–99.

Shilane,P. *et al.* (2004) The Princeton Shape Benchmark. In *Proceedings of International Conference on Shape Modeling and Applications*, (*SMI'04*), Palazzo Ducale, Genova, IEEE Computer Society Press.