Paper

PROTEIN: A Visual Interface for Classification of Partial Reliefs of Protein Molecular Surfaces

Keiko NISHIYAMA, Takayuki ITOH (Member)
Graduate School of Humanitics and Sciences, Ochanomizu University

⟨Summary⟩ 3D structure of proteins deeply relates to their functionality. It is well-known that functions of proteins strongly appear in the bumpy parts of the molecular surfaces, and therefore geometric analysis of protein molecular surfaces is important.

We propose a technique to extract geometric features of protein molecular surfaces, and a visual interface to effectively visualize the extracted results. Assuming that molecule surfaces are approximated as triangular meshes, the technique extracts groups of triangles forming partial reliefs, calculates their feature values, and finally clusters them according to the feature values. In addition, our technique extracts larger similarly shaped parts consisting of two or more reliefs, by simplifying the triangular meshes and applying the graph route problem. The technique provides a visual interface for visualizing the above results applying a hierarchical data visualization technique "HeiankyoView".

Key words: Molecular surfaces, Partial relief, Clustering, Visualization

1. INTRODUCTION

Proteins are raw materials that are the constituent of enzyme, internal organs, hormone, and other important materials for human body. The proteins also have roles of important actions for various in vivo reacts. Analysis of the protein is important in a lot of fields, including medicine, pharmacology, and biology. Conventional studies on protein analysis were mainly based on the decoding of amino acid sequence, so called the primary structure of the protein. On the other hand, recent many reports argue that functions of the protein greatly depend on the shapes of molecular surfaces. Especially, shapes of partial reliefs of the surfaces deeply relate to functionality of proteins. Therefore, we can expect to understand the interaction and function of proteins, and their correlations to other already-known proteins, by analyzing the partial reliefs of proteins. Thanks to recent enrichment of molecular surface database, such as eF-site[1)], retrieval and comparison of geometry of molecular surfaces are hot topics for protein analysis. Since there have been many techniques on 3D geometry retrieval and categorization in the field of CG and CAD, we expect that the techniques can be applied to the retrieval and comparison of geometry of molecular surfaces. We also expect that this application can contribute to various academic and industrial protein-related fields.

This paper proposes PROTEIN (Partial Relief Observation TEchnique and INterface), a technique and visual interface focusing on retrieval and classification of partial reliefs of proteins, as well as visual interface to explore the partial reliefs of proteins. The technique consists of the following steps:

1. Extract partial reliefs from molecule surfaces.
2. Calculate feature values of partial reliefs.
3. Cluster the partial reliefs according to the feature values.
4. Simplify the triangular meshes, and discover larger similar parts applying the graph route problem to the simplified meshes.
5. Visualize the extracted results.

The proposed technique assumes that geometry of protein molecular surfaces is modeled as triangular meshes. Several existing molecular surface analysis techniques also assume to deal with triangular meshes; however, these techniques retrieve features of geometry per triangle or vertex, and therefore their computation time may be very high. On the other hand, our approach is based on classification of partial reliefs of geometry, because we think it is computationally reasonable, and it should be an interest of protein researchers since functionality of proteins highly relates to their partial reliefs.

Functional analysis of proteins is a very complicated problem, because it is highly related not only to geometry but also chemical properties. We think visual analysis tools are useful for this complicated problem, because the tools can incorporate knowledge and experiments of researchers. Our technique applies a hierarchical data visualization technique "HeiankyoView" for the overview of clustering results. It interlocks with a 3D molecular surface viewer so that users can easily look the geometry of partial reliefs of particular clusters.

2. RELATED WORK

2.1 Comparison of Protein Structure

There have been many techniques on comparison of the protein structure; however, many of traditional techniques do not refer geometry of molecular surfaces. A typical approach is based on the comparison of position of atoms between proteins. However, this approach may cause very high computation times, and essentially difficult because atoms are always moving and they do not have any stable positions. Another approach compares amino arrays and fold structures. In this approach, some techniques compare primary structures of proteins[2), and some of others compares secondary structures and distances between atoms[3). However, there have been many cases that functions of two proteins are not similar though they are serologically related and their fold structures are therefore determined as similar.

A lot of recent protein comparison techniques have been based on their molecule surfaces[4). A typical technique constructs "vector pairs"[5), which are the pairs of adjacent normal vectors with their physical properties, and then extracts the collection of similar vector pairs as the parts of similarly shaped surfaces. Another technique applies Creek retrieval method for normal vectors with physical properties[6), and it is implemented on eF-site. However, these techniques may also cause very high computation times.

2.2 Comparison of 3D Geometry

Many of 3D geometry comparison techniques are based on geometric features, and others are based on topological features. Typical techniques are summarized as follows:

- Octree- or voxel-based comparison,
- Frequency-domain comparison,
- 2D-projection-based comparison, and
- Scatter-point-based comparison.

The technique proposed in this paper uses points scattered onto the triangular surfaces, and calculates feature values using the points. Existing feature value calculation techniques include D2[7), based on the histogram of distances between corresponding points, and PS[8), based on the histogram of distances, variances, and other values around the medial axis of the geometry.

2.3 Graph matching

Graph matching is a famous problem to discover common parts from graphs, and has been applied to various node-link data, such as chemical compounds and human networks. One problem of graph matching is that it may take very large computation time. Many acceleration techniques have been therefore presented, including $A*$algorithm-based best-first search[9), subgraph decomposition[10) , and subgraph-based graph simplification[11).

2.4 Hierarchical data visualization

Our technique uses "HeiankyoView"[12)to visualize the clustering result. HeiankyoView represents leaf nodes of hierarchical data as square icons, and the branch nodes as rectangular borders, as shown in **Fig. 1**. It targets all-in-one display of leaf nodes of whole hierarchical data, rather than representation of connectivity between parent and children nodes.

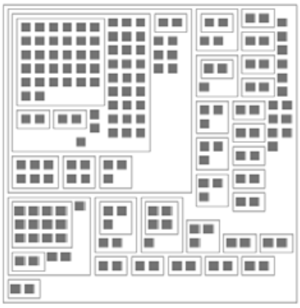

Fig. 1 Example of hierarchical data visualized by "HeiankyoView".

3. PROPOSED TECHNIQUE

3.1 Data structure of molecular surface

The proposed technique assumes that molecular surfaces are modeled as triangular meshes. Currently we use the molecular surfaces retrieved from eF-site[1]. The molecular surfaces are constructed from solid protein structural information registered in PDB (http://www.pdb.org/), based on the definition of Connolly surface. eF-site publishes molecular surfaces as triangular meshes, consisting of vertices, edges, and triangles, as XML documents. Vertices have geometric values including coordinates, normal vectors, maximum and minimum curvatures, and chemical values including hydrophobe, temperature, and potential values. Molecular surfaces retrieved from eF-site may contain tens or hundreds of thousands of vertices and triangles, and the sizes of XML documents may become several megabytes. That is why we mentioned in Section 2.1 that vertex or triangle oriented comparison techniques may cause very high computation times.

In this paper we formalize the data structure of triangular mesh of molecular surfaces as follows:

- Vertex $V = \{v_1, ..., v_{nV}\}$, where nV is the number of vertices.
- Edge $E = \{e_1, ..., e_{nE}\}$, where nE is the number of edges.
- Triangle $T = \{t_1, ..., t_{nT}\}$, where nT is the number of triangles.
- Relief $R = \{r_1, ..., r_{nR}\}$, where nR is the number of reliefs, and a relief r_i consists of a set of vertices and triangles.
- Cluster $C = \{c_1, ..., c_{nC}\}$, where nC is the number of clusters, and a cluster c_i consists of a set of reliefs.

3.2 Partial relief extraction

Our technique extracts partial reliefs from the molecular surfaces, by the following two steps:

Step 1: Shape determination

The technique first assigns attributes to vertices. Let the position of vertex v_i as (x_i, y_i, z_i), and its normal vector as $(n_{x_i}, n_{y_i}, n_{z_i})$. The tangent plane at v_i is represented as equation (1), where as $t = 0$:

$$t = n_{x_i}(x - x_i) + n_{y_i}(y - y_i) + n_{z_i}(z - z_i) \qquad (1)$$

The technique calculates values t by equation (1), for vertices connected to v_i via edges of the triangular mesh. If all of t are positive, the technique determines that v_i belongs to a convex. If all of t are negative, it determines that v_i belongs to a concave. If there are both positive and negative values, it determines that v_i does not belong either convex or concave. By repeating this process, the technique assigns either "convex", "concave", or "other", for all vertices of the triangular mesh.

Step 2: Labeling

The technique then recursively traverses the adjacent vertices which have same marks, either "convex" or "concave", and forms groups of the traversed adjacent vertices. It then assigns sequential numbers (called "labels" in this paper) to the groups. It also assigns the specific label to vertices belonging to the group, and to triangles whose three vertices have the same label. It recognizes the group of vertices and triangles as a partial relief. **Fig. 2** shows an example of the partial reliefs extracted by the above procedure. The example applies a small protein named "1gcn". The example draws convex parts in red, and concave parts in blue.

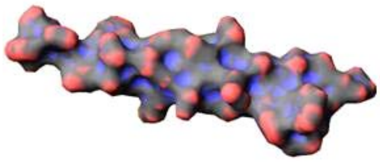

Fig. 2 Labeling result. Convex parts are in red, and concave parts are in blue.

3.3 Feature value calculation

This section describes a technique to calculate feature values of partial reliefs, using medial axis, similar to PS method[8]. Following is the procedure of the technique, and **Fig. 3** illustrates the procedure.

1. Specify the "top vertex" which locates at the top of a convex, or the bottom of a concave.
2. Generate the medial axis from the top or bottom vertex, and divide the medial axis into k pieces of segments.
3. Generate points randomly on the triangles.
4. Calculate distances from each of the points to the medial axis.
5. Calculate the average and variance of the distances for each pieces of the medial axis.
6. Form a histogram from these values, and use them as a feature vector.

Here, a base surface of a partial relief is approximated as the plane which is averagely closest to the vertices on the border of a partial relief. A top vertex is then defined as the vertex that is the most distant vertex from the base surface.

By the way, similarly shaped pairs of convex and concave are often chemically sensitive, and therefore discovery of such similarly shaped pairs is important for the analysis of sensitive parts of proteins. Our technique therefore reverses the geometry of concave symmetrically around the base surface, so that it can easily discover the similarly shaped pairs of convex and concave.

Functionality of partial reliefs is different if their geometry is similar but their sizes are different: our technique therefore does not normalize the sizes of partial reliefs before calculating the feature values.

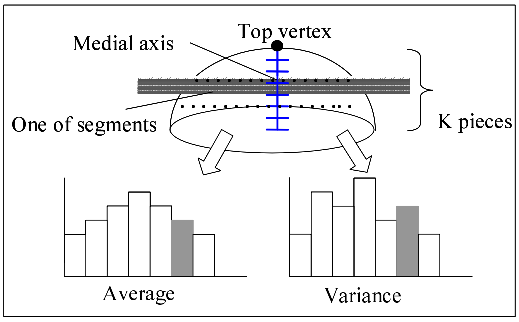

Fig. 3 Calculation of feature values of partial reliefs.

3.4 Clustering

Next, our technique clusters the partial reliefs according to the feature values calculated in the above procedure. Clustering and dimension reduction techniques are hot topics, not only for 3D geometry but also for various kinds of media (i.e. still image and audio). Application of such clustering and dimension reduction techniques will be our future works.

3.5 Extraction of larger similar parts

The technique can extract larger geometric features consists of two or more reliefs. The technique generates "partial relief graph" that connects top vertices, where top vertices are vertices positioned at tops of reliefs. It then extracts common parts from the partial relief graphs, as larger similar parts.

The technique first applies mesh simplification to obtain the triangular mesh which never contains non-top vertices, where top vertices are vertices locating at the top or bottom of partial reliefs. The technique is based on well-know mesh simplification techniques[13),14)], except our technique never deletes the top vertices. We call the simplified mesh as partial relief graph.

Here, the technique may generate different topology from very similar geometry. For example, a quadrilateral shown in **Fig. 4** (left) may be divided as Fig. 4 (a) or (b), if the quadrilateral is almost rectangular. In this case our technique generates both edges, as shown in Fig. 4(c). This process is done after removing every non-top vertices.

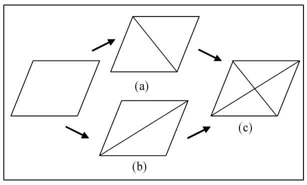

Fig. 4 Edge addition.

Fig. 5 shows an example of the partial relief graph generated from a molecular surface of the protein named as 1gcn. In this figure, yellow numeric characters denote the sequential number of clusters c_1 to c_{nC}.

By applying the graph route problem to the partial relief graphs, we can discover candidates of the larger similar parts among proteins. Note that this idea

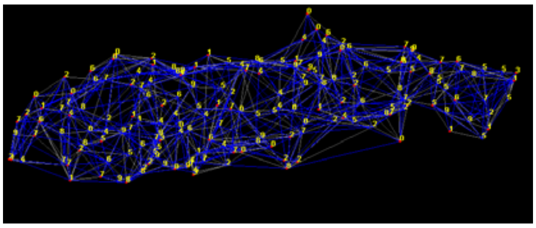

Fig. 5 Partial relief graph generated from 1gcn.

just extracts partial features which adjacency of similar reliefs is similar; it does not guarantee that the extracted parts are always geometrically very similar. It would be better to combine other techniques for determination of shape similarity; our idea quickly discovers candidates of larger similar parts, and then other techniques can determine if the extracted parts are geometrically similar or not.

3.6 Visual interface

The technique applies HeiankyoView, introduced in Section 2.4, as a visual interface to explore the clustering result. Our technique constructs hierarchical data by classifying the partial reliefs as follows.

1. Classify them according to clustering result.
2. Classify them according to their parent proteins.
3. Classify them into convex and concave.

The technique then visualizes the hierarchical data, where colored icons denote partial reliefs, and rectangular borders denote clusters. We expect the visualization by HeiankyoView is useful for the following purposes:

- Understanding of distribution of the clustering result.
- Understanding of relationship between geometric features and chemical properties of reliefs.
- Discovery of protein pairs that share a lot of similarly shaped pairs of convex and concave.

Moreover, we provide a 3D viewer for the visualization of molecule surfaces. It can selectively display the partial reliefs that belong to the specific cluster. When users discover interesting clusters by using HeiankyoView, they can specify the cluster on the 3D viewer, and it then displays only the partial reliefs belonging to the specific cluster. The combination of these two viewers enables all-in-one display of clustering results and specific cluster observation on the 3D viewer.

4. RESULTS

We developed extraction, feature calculation, clustering, and hierarchical data visualization modules of the technique on Java SDK 1.5. We also developed 3D molecular surface viewer on GNU gcc 3.4 and OpenGL/GLUT. However, we would like to redevelop the 3D viewer on Java with Java3D or JOGL, for the improvement of interoperability between 3D viewer and our other modules, or other many famous bioinformatics software which are developed in Java.

In this experiment, we downloaded molecular surfaces of two proteins, named "1yee" and "1yec", from eF-site. It is well-known that the two proteins are roughly similarly shaped. **Fig. 6** shows the entire surfaces of the proteins, where we can observe that the proteins look similar. **Tab. 1** shows the number of triangles of the two proteins used in our experiment. This experiment required 0.41 seconds for partial relief extraction, and 1.34 seconds for feature calculation and classification, on IBM ThinkPad T60 (CPU 1.8GHz, RAM 1GB).

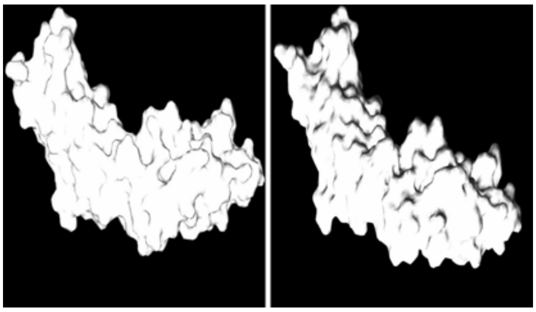

Fig. 6 (left) Molecular surface of 1yee. (right)Molecular surface of 1yec.

Table 1 Composition triangle number of each protein

Protein name	1yee	1yec
Number of polygons	26066	26418

Fig. 7 and **Fig. 8** show the result of clustering of partial reliefs extracted from 1yee and 1yec. We applied three levels of clustering processes for this visualization. We first generated clusters according to the feature values of the partial reliefs, as described in Section 3.4, where we call them "top-level clusters".

We then divided the reliefs in each of the top-level clusters according to the protein that contains the reliefs, and formed the "second-level clusters". Finally, we divided the reliefs in each of the second-level clusters into two clusters "convex" and "concave", and formed the "third-level clusters". In the figures, clusters denoting reliefs of 1yee are displayed in the left side of the top-level clusters, and clusters denoting reliefs of 1yec are displayed in the right side. Also, clusters denoting convex reliefs are displayed in the upper side of the second-level clusters, and clusters denoting concave reliefs are displayed in the lower side.

Looking at Fig. 7 and Fig. 8, we can observe that the numbers of reliefs in most of second-level clusters are almost same in the most of top-level clusters; it may indicate the two proteins, 1yee and 1yec, are geometrically similar. Moreover, we can observe that the number of concave parts is much more than that of convex parts. From this result, we can suppose that the two proteins tend to become passive for the reaction.

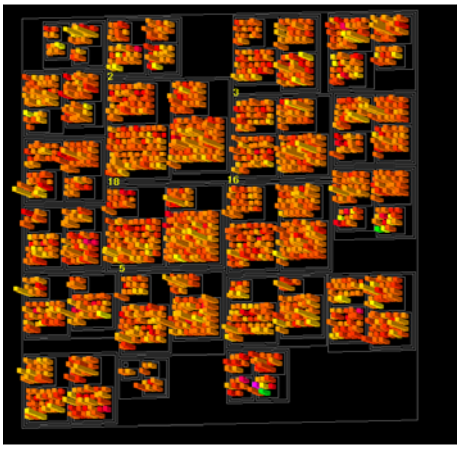

Fig. 7 Heights denote temperature value and colors denote potential value

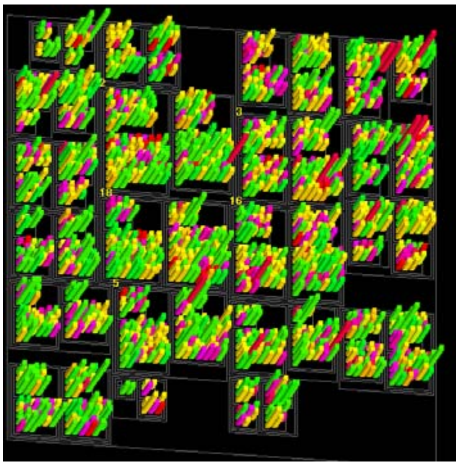

Fig. 8 Heights denote shape and colors denote Protein ID.

By using HeiankyoView, we can also observe the relationship between clustering results and chemical properties. In **Fig. 7**, heights of icons denote temperature at top vertices of the reliefs, and colors denote potential value at the vertices. We can observe that many of tall icons are yellow. From the result, we can observe the correlativity between potential and temperature. Also, though most of icons are orange, pink icons which denote these potentials are high, and yellow-green icons which denote these potentials are low, are mixed only in several clusters. All such clusters were concave reliefs extracted from 1yec. Such interesting features can be easily discovered by using HeiankyoView.

In **Fig. 8**, colors of icons denote hydrophobicity, and heights of icons denote temperature. If hydrophobe is high, the color approaches to red; if it is low, the color approaches to green. We can observe that most of tall icons are red or green. It denotes that hydrophobicity is apart from average, if temperature is high. Such interesting correlativity can be also easily discovered by using HeiankyoView.

Fig. 9 shows the partial reliefs of the two proteins belonging to the same cluster. We can observe that the partial reliefs of the two proteins similarly distribute on the two proteins. **Fig. 10** shows the zoom-up view of the partial reliefs belonging to the specific cluster of 1yee. We can observe that the partial reliefs are similarly shaped, even if they are whether convex or concave.

Fig. 11 shows an example of extracted similar parts of partial relief graphs of 1yee and 1yec, enclosed by purple polygons, which consist of four partial reliefs. **Fig. 12** shows the corresponding parts of molecular surfaces of 1yee and 1yec. This figure encloses convex parts by rectangles in double dotted lines, and concave parts by rectangles in solid lines.

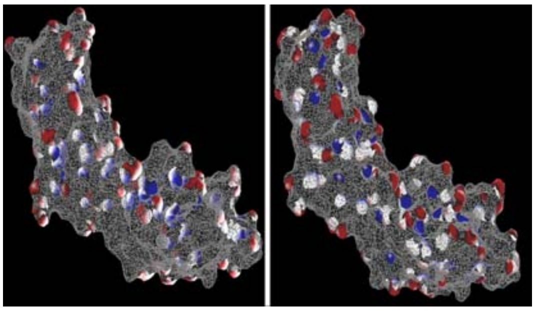

Fig. 9 Partial reliefs belong to the specific cluster.

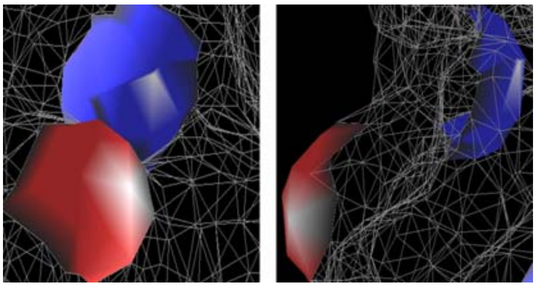

Fig. 10 Zoom-up view of Fig. 9. (left) Convex part is in red, and concave part is in blue. (right) Same parts from a different viewpoint.

Fig. 11 denotes the identification numbers of clusters: where numbers of four reliefs in the extracted parts are 0, 2, 2, and 0.

Concave parts of proteins may easily become the passive of the reaction with compounds than the convexes. Since the extracted parts contain three concave reliefs, we could discover candidate parts that may be passive of the similar reactions, from the two proteins 1yee and 1yec.

Note that the extracted two parts look geometrically a little bit different. As discussed ahead, the technique just extracts parts where adjacency of similar reliefs is similar, and therefore we think the extracted parts are just candidates of similar parts. We would like to integrate with other techniques for determination of shape similarity, so that we can guarantee the similarity of extracted parts.

5. CONCLUSION

This paper proposed a technique and a visual interface for partial reliefs of protein molecular surfaces. It presented extraction, feature vector calculation, and clustering of partial reliefs of protein molecular surfaces, and extraction of candidates of larger similar parts applying mesh simplification and graph matching. It also presented a visual interface applying HeiankyoView and a 3D viewer, to interactively explore the partial reliefs.

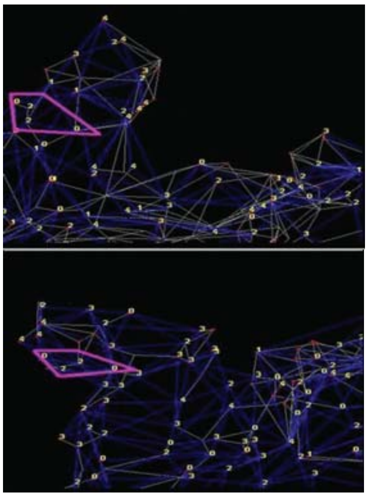

Fig. 11 Extracted similar parts. (upper) Partial relief graph of 1yee. (lower) Partial relief graph of 1yec.

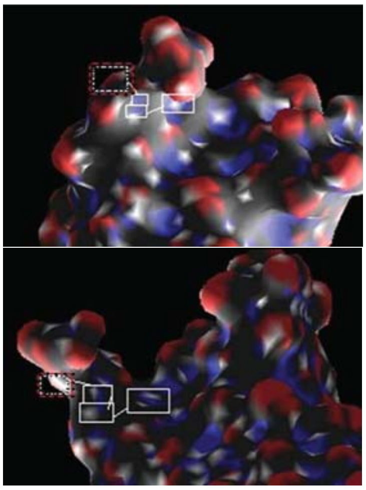

Fig. 12 Extracted similar parts. (upper) Molecular surface of 1yee. (lower) Molecular surface of 1yec.

As future works, we would like to discuss the following issues:

- Consideration of chemical values in addition to geometric values for feature value calculation.
- Balancing the computational complexity and accuracy by controlling level of detail control of triangular meshes.

- Application of sophisticated clustering and dimension reduction techniques.
- Consideration of the length of edges of partial relief graphs.
- Subjective evaluation of the user interface.
- Quantitative evaluation of clustering results.
- Construction of partial relief database with large number of protein molecular surfaces retrieved from eF-site.

Acknowledgment We appreciate to Prof. Kengo Kinoshita with the University of Tokyo, for overall valuable suggestion for the calculation of structural similarity of the protein molecular surfaces. We appreciate to Prof. Mariko Hagita with Ochanomizu University, for fruitful discussion of geometric processing to extract partial reliefs.

References

1) eF-site, http://pi.protein.osaka-u.ac.jp/eF-site/

2) Russell, R. B., Sasieni, P. D., Sternberg, M. J. E., Supersites within superfolds: binding site similarity in the absence of homology, Journal of Molecular Biology, 282(4), 903-918, 1998.

3) Singh, A. P., Brutlag, D. H., Hierarchical Protein Structure Superposition Using Both Secondary Structure and Atomic Representations, Proceedings of Intelligent Systems for Molecular Biology, 284-293, 1997.

4) Via. A., Ferre. F., Brannetti. B., Helmer-Citterich. M., Protein surface similarities: a survey of methods to describe and compare protein surfaces, Cell Mol Life Sci., 57(13-14), 1970-1977, 2000.

5) Shimizu. Y., Nripendra. L. S., A method of parallel processing of protein surface motifs extraction, Journal of Information Processing Society: Mathematical principle modeling and application (TOM), 47-SIG1(TOM14), 120-129, 2006.

6) Kinoshita. K, Nakamura. H., Identification of protein biochemical functions by similarity search using the molecular surface database eF-site, The Biophysical Society of Japan, ISSN:05824052, 42(1), 20-23, 2002.

7) Osada. R., Funkhouser. T., Chazelle. B., Dobkin. D., Shape Distributions, ACM Transactions on Graphics, 21(4), 807-832, 2002.

8) Otagiri. T., Ibato. M., Takei. T., Ohbuchi. R., Shape-Similarity Search of 3D Models by Using Moment Envelopes, The journal of the Institute of Image Information and Television Engineers, 56(10), 1589-1597, 2002.

9) Asanobu. K., Mikio. T., Fast Matching of Hierarchical Attributed Relational Graphs for an Application to Similarity-Based Image retrieval, MIRU'96, 2, 331-336, 1996.

10) Bruno. T. M., Horst. B., A New Algorithm for Error-Tolerant Subgraph Isomorphism Detection, IEEE Trans. Pattern Analysis and Machine Intelligence, 20(5), 493-504, 1998.

11) Akira. M., Chien. N. P., Kouzou. O., Hiroshi. M., Takashi. W., Analysis of Hepatitis Dataset by Using Cl-GBI (Medical Active Mining), Information Processing Society of Japan, 2004(125), 43-48, 2004.

12) Itoh. T., Takakura. H., Sawada. A., Koyamada. K., Hierarchical Visualization of Network Intrusion Detection Data in the IP Address Space, IEEE Computer Graphics and Applications, 26(2), 40-47, 2006.

13) Michael. G., Paul S. H., Surface Simplification Using Quadric Error Metrics, ACM SIGGRAPH 97 Conference Proceedings, ISBN:0-89791-896-7, 209-216, 1997.

14) Hugues. H., Progressive meshes, Computer Graphics (SIGGRAPH 96), ISBN:0-89791-746-4, 99-108, 1996.

Keiko Nishiyama

She received B.S. and M.S. degrees from Ochanomizu University in 2006 and 2008 respectively. She has been studying on bioinformatics-related analysis and visualization. She is a member of IPSJ.

Takayuki Itoh (Member)

He received B.S., M.S., and Ph.D degrees from Waseda University in 1990, 1992 and 1997 respectively. He has been an associate professor in Ochanomizu University since 2005. His research interest includes scientific and information visualization, geometry, image, and music processing, computer graphics, and distributed computing systems. He is a member of IEEE CS, ACM, IPSJ, IIEEJ, and so on.

(Received December 31, 2007)