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

Keiko NISHIYAMA *Takayuki ITOH*
Graduate School of Humanitics and Sciences, Ochanomizu University

ABSTRACT

3D structure of proteins deeply takes part in the expression of the protein. Molecular surfaces of the protein generally have very complex and bumpy shapes. It is well-known that functions of proteins strongly appear in the bumpy parts of the molecular surfaces.

We propose a visual interface to effectively visualize the partial reliefs of molecular surfaces of proteins. This technique assumes that molecule surfaces are approximated as triangular meshes. It first extracts groups of triangles forming partial reliefs, and calculates their feature values as histograms. It finally clusters the partial reliefs according to the histogram. The presented technique then visualizes the clustering results applying a hierarchical data visualization technique "HeiankyoView", as a visual interface to explore the clustered partial reliefs.

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 (primary structure of the protein). However, recently there are many reports 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. We 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 comparison and classification of proteins. 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 an interface to observe partial reliefs of molecular surface geometry of proteins. Our approach focuses not only on retrieval and classification of partial reliefs of proteins, but also on 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. Visualize the clustering result.

The proposed technique assumes that geometry of molecular surfaces of proteins 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.

Since function of proteins highly related not only to geometry but also chemical properties, it is a very complicated problem if we analyze them only by quantitative or numerical evaluation. We think subjective analysis tools are useful for the study of proteins because the tools can incorporate researchers' knowledge and experiments, and therefore visualization can be a useful tool for this study. 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. 3D Protein Structure Comparison

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 frame structures of proteins [2], and some of others compares secondary structure 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", which are the pairs of adjacent normal vectors with their physical properties, and then extracts the similar collection of vector pairs as the parts of similarly shaped surfaces [5]. 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. 3D Geometry Comparison

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, distribution, and other values around the interior axis of the geometry.

2.3. Hierarchical data visualization by "HeiankyoView"

Our technique uses "HeiankyoView" [9] to visualize the clustering result. HeiankyoView represents leaf nodes of hierarchical data as square icons, and the branch nodes as rectangular borders. The technique focuses on all-in-one display of leaf nodes of whole hierarchical data, rather than representation of connectivity between parent and children nodes.

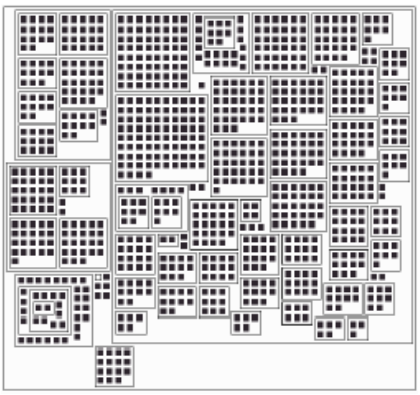

Figure 1. Example of hierarchical data visualized by "HeiankyoView".

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 calculated based on the definition called Connolly surface, from solid protein structural information registered in PDB (http://www.pdb.org/). eF-site publishes molecular surfaces as triangular meshes, consisting of vertices, edges, and triangles, and provides as XML documents. Vertices have geometric values including coordinates, normal vectors, maximum and minimum curvatures. They also have chemical values including hydrophobe, temperature, and potential values.

Molecular surfaces retrieved from eF-site usually contain tens or hundreds of thousands of vertices and triangles, and the sizes of XML documents 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.

3.2. Partial relief extraction

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

Step 1 : Shape judgment in the each top

The technique first assigns relief attributes to vertices. Let the position of vertex A as (x_A, y_A, z_A) , and the normal vector as (n_{xA}, n_{yA}, n_{zA}). The tangent plane of A is represented as equation (1), where $t=0$:

$$t = n_{xA}(x - x_A) + n_{yA}(y - y_A) + n_{zA}(z - z_A) \ldots (1)$$

The technique calculates values t by equation (1), for vertices connected to A via edges of the triangular mesh. If all of t are positive, the technique determines that A belongs to a convex. If all of t are negative, it determines that A belongs to a concave. If there are both positive and negative values, it determines that A 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 number to vertices belonging to the group, and to triangles whose three vertices have the same number. It recognizes the groups of vertices and triangles as partial reliefs.

We experimentally observed that the partial reliefs formed by the above method look a little bit smaller than we expected, and we therefore developed an additional procedure to adjust the sizes of the partial reliefs. The procedure is based on a heuristic that the sizes of partial relief are usually adjusted by adding one layer of triangles. It adds triangles those one or two vertices have the specific number of group to the group. It also adds vertices connected to the added triangles to the group.

Figure 2 shows the 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.

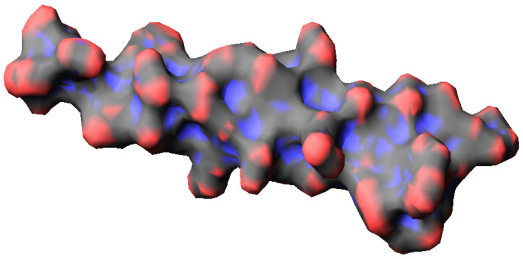

Figure 2. Labeling result. Convex parts are in red, and concave parts are in blue.

3.3. Feature value calculation

This section describes a technique to calculate feature values of partial reliefs, using medial axis, similar to PS method described in [8]. Following is the procedure of the technique, and Figure 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 to the base surface of the partial relief, and divide the medial axis into k pieces of segments.
3. Generate points on the triangles of the partial relief. Our current implementation randomly generates same number of points on each of triangles.
4. Make segments vertically intersecting to the medial axis from the points, and calculates the length of the segments.

5. Calculate the average and variance of the length of the segment, and aggregates them for each pieces of the medial axis. Finally, the technique forms a histogram from these values, and uses it 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. 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 geometry before calculating the feature values.

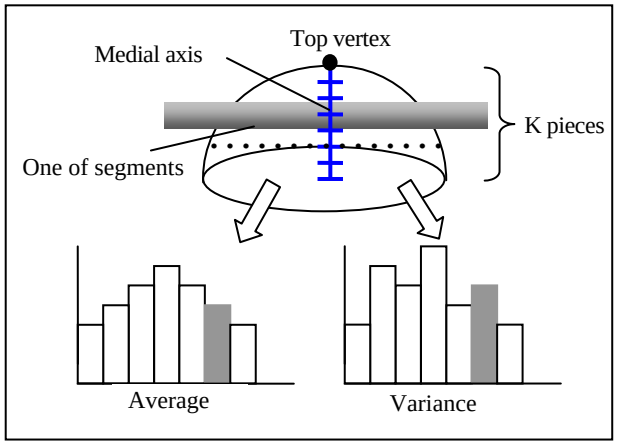

Figure 3. Calculation of feature values of partial reliefs.

3.4. Clustering

Next, our technique clusters the partial reliefs according to the feature values calculated in the above procedure. Current our implementation applies K-means method as a clustering algorithm. 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. 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 from the clustering result as follows.

- Classify the partial reliefs according to clustering result.
- Classify them according to their parent proteins, if necessary.
- Classify them into convex and concave, if necessary.

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 partial reliefs.
- Comparison of the clustering results among proteins.
- 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 display only the partial reliefs that belong to the specific cluster. Since users discover interesting clusters and know the label of the cluster on HeiankyoView, they can specify the label 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 blocks 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 re-develop the 3D viewer on Java with Java3D or JOGL. One reason is that many famous bioinformatics software have been developed on Java. Another reason is that HeiankyoView has been also developed on Java, and we would like to improve interoperability between HeiankyoView and the 3D viewer.

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. Figure 4 shows the entire surfaces of the proteins. We can observe that the proteins look similar.

Table 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).

Table 1: Composition triangle number of each protein

Protein name	1yee	1yec
Number of polygons	26066	26418

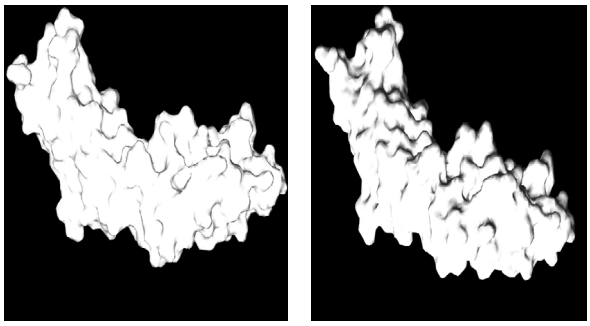

Figure 4: (left) Molecular surface shape of 1yee. (right)Molecular surface shape of 1yec.

Figures 5, 6, and 7 show the result of clustering of partial reliefs extracted from the two proteins. 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 which the reliefs are contained, 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, left clusters denote reliefs of 1yee, and right clusters denote reliefs of 1yec. Also, upper clusters denote convex reliefs, and lower clusters denote concave reliefs.

In Figure 5, partial reliefs extracted from 1yee are represented as magenta icons, and from 1yec are represented as yellow-green icons. Also, heights of icons denote the shape of partial reliefs, where convex reliefs correspond to short icons, and concave reliefs correspond to tall icons. We observed that the numbers of reliefs in most of second-level clusters were almost same in the most of top-level clusters It may indicated the two proteins, 1yee and 1yec, were geometrically similar. Moreover, we could observe that the number of concave parts was much more than that of convex parts from the heights of icons. From this result, we could suppose that the two proteins tend to become passive voice for the reaction.

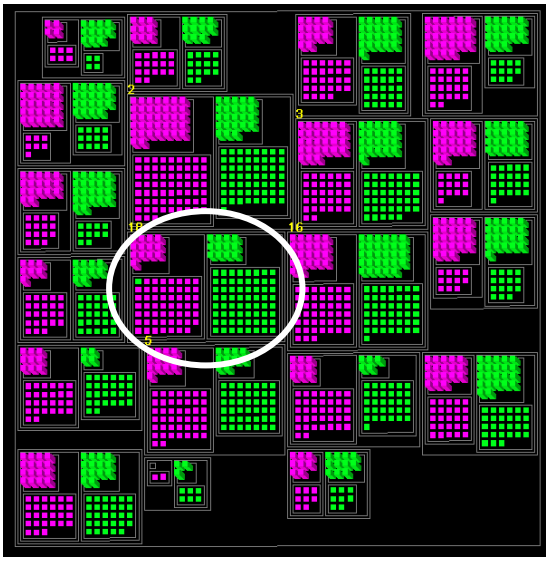

Figure 5: Visualization of clustering result by HeiankyoView. Heights denote shape and colors denote Protein ID.

Figure 6: Heights denote temperature value and colors denote potential value.

Hierarchical data visualization by HeiankyoView has variety of capability to represent multivariate values as well as hierarchical structures, where we can apply to visualization of relationship between clustering results and chemical features. By using HeiankyoView, we can also observe the relationship between clustering results and chemical properties.

In Figure 6, 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, as indicated by yellow-green circles in Figure 6. From the result, we can observe the correlativity between potential and temperature.

Let us make an attention to distribution of colors in Figure 6. Most of icons are orange, but in several clusters, pink icons which denote these potentials are high, and yellow-green icons which denote these potentials are low, are mixed, as indicated in blue dotted circles. All such clusters were concave reliefs extracted from 1yec. Such interesting features can be easily discovered by using HeiankyoView.

In Figure 7, colors of icons denote hydrophobicity, and heights of icons denote temperature. 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.

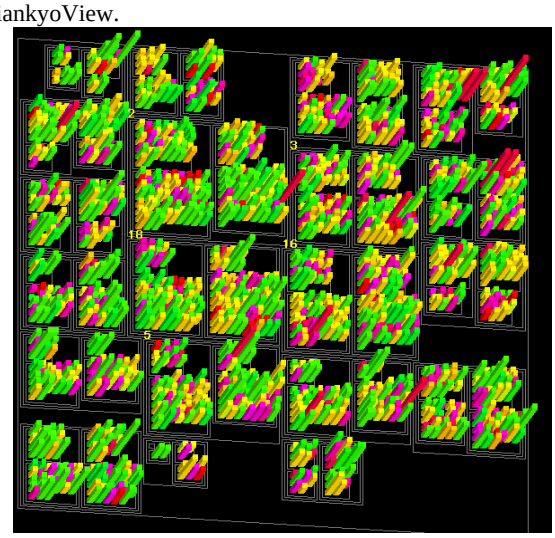

Figure 7: Heights denote shape and colors denote Protein ID.

Figure 8 shows the partial reliefs of the two proteins belonging to the same cluster (marked as a while circle in Figure 5). We observed that the partial reliefs of the two proteins similarly distributes on the two proteins. Figure 9 shows the partial reliefs belong to the specific cluster of 1yee. We observed that the partial reliefs are similarly shaped, even if they have different attributes such as “convex” or “concave”.

6. CONCLUSTION

We proposed a technique of a visual interface for partial reliefs of protein molecular surfaces. The technique visualizes clusters of partial reliefs by HeiankyoView and 3D viewer.

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

- Consideration of chemical values in addition to geometric values for feature value calculation.
- Balance the computational complexity and accuracy by controlling level of detail control of triangular meshes.
- Application of sophisticated clustering and dimension reduction techniques.
- Subjective evaluation of the user interface.

- Experiments of visualization of additional information (i.e. chemical values) on HeiankyoView.
- Quantitative evaluation of clustering results.
- Construction of partial relief database with large number of protein molecular surfaces retrieved from eF-site.

Also, we would like to extend the proposed technique to discover larger similar geometric parts of surfaces. Following is the procedure of our planning extension:

1. Cluster partial reliefs, and assign labels to them.
2. Simplify the triangular mesh so that it remains only top vertices. Consequently the mesh forms "partial relief graph".
3. Apply the simple graph route problem, so that we can discover the larger similar parts among proteins.

Proteins have revitalization parts that directly relate to the functions of the proteins. The revitalization parts locate in the very limited area on the molecule surfaces. It is important to discover the revitalization parts, and analyze their physical and geometric properties, to study function of proteins. Also, it is well-known that revitalization parts which have the same function are similarly structured and shaped. These revitalization parts can be discovered in some different proteins, and it means that such proteins may have common functionality even they have roughly different geometry. We think that the extension of the proposed technique will help to discover and analyze such revitalization parts of proteins.

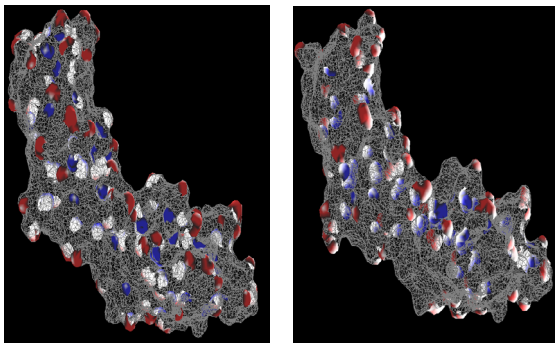

Figure 8: Partial reliefs belong to the specific cluster (shown as a white circle in Figure 5).

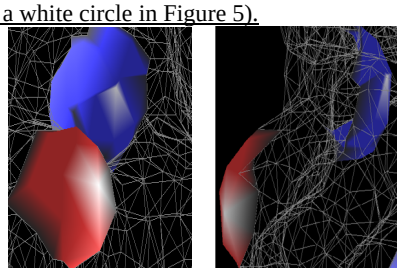

Figure 9: Zoom-up view of Figure 6. (left) Convex part is in red, and concave part is in blue. (right) Same parts from a different viewpoint.

ACKNOWLEDGEMENT

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.

REFEERNCES

[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, Vol. 282, No. 4, pp. 903-918, October 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, pp. 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., Vol. 57, No. 13-14, pp. 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), Vol. 47, No. SIG1(TOM14), pp. 120-129, 2006.2.

[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, Vol. 42, No. 1, pp. 20-23, 2002.

[7] Osada. R., Funkhouser. T., Chazelle. B., Dobkin. D., Shape Distributions, ACM Transactions on Graphics, Vol. 21, No. 4, pp. 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, Vol. 56, No. 10, pp. 1589-1597, 2002.

[9] 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, Vol. 26, No. 2, pp. 40-47, 2006.