A method for retrieving proteins with a local structure similar to known interaction sites using profiling

Yusuke Nonomura[1], Koichi Yoshino[1], and Takenao Ohkawa[2]
[1]Graduate School of Information Science and Technology, Osaka University, JAPAN
2-1, Yamadaoka, Suita, Osaka 565-0871 Japan
[2]Graduate School of Science and Technology, Kobe University, JAPAN
1-1, Rokkodai, Nada, Kobe 657-8501 Japan
yoshino.koichi@ist.osaka-u.ac.jp, ohkawa@cs.kobe-u.ac.jp

Abstract

It is known that a protein's function is determined by its interaction with other molecules at a local site. Since proteins with similar structures often have similarity in their functions, finding structurally similar proteins at a local site plays an important role in protein functional analysis. This paper proposes a method of retrieving a protein with a local structure that is similar to a known interaction site structure in terms of its binding ablility to a specific compound. In this method, a profile, which is defined as a set of common atoms in some interaction sites that bind to the same compound, is introduced for generalizing features of the interaction sites. Experimental results demonstrate that the proposed method succeeds in detecting interaction sites that cannot be found by using non-profiled queries.

Keywords: protein, retrieval system, interaction site, protein structure database, profile

1 Introduction

A protein's function depends on its interaction with other molecules at a local site. There are many proteins whose structures have been determined but whose functions remain unknown. Identifing compounds that have a chance of binding to such proteins plays an important role in clarifying a given protein's function.

In this paper we propose a method for retrieving a protein with a local structure that is similar to a known interaction site in terms of its ability to bind to a specific compound.

Figure 1 presents outline of the method. The input data are structural data on the interaction site of a protein-compound complex. Such structural data of the interaction site are expressed as coordinate data of atoms that compose the interaction site [1]. A protein with a local structure similar to the input structure is retrieved from the database on protein structures for which the interaction site is not specified, and the data is given as a coordinate data of each atom. The retrieved protein is expected to have some possibility of binding to the same compound as the input one.

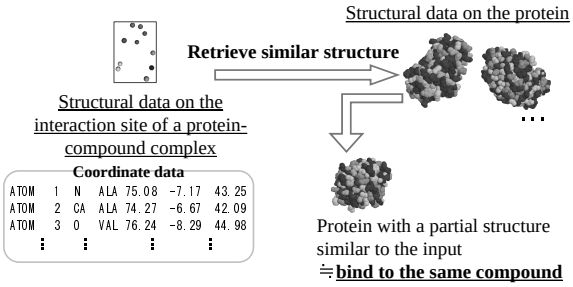

Figure 1: Outline of the retrieval method

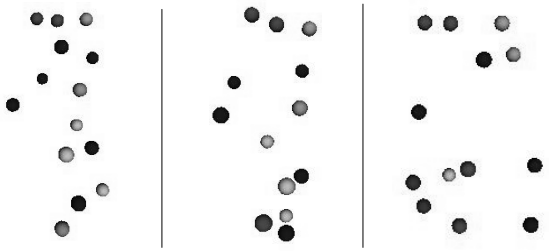

Figure 2: Example of an interaction site that binds to the compound GDP

In general, structures at the interaction site binding to the same compound are not necessarily identical [2, 3]. For example, Figure 2 illustrates groups of atoms constituting interaction sites that bind to the compound "Guanosine-5'-Diphosphate" (GDP), where shading of the spheres in the figure represents the type of atom. Therefore, raw data on the interaction site is not suitable as an input for querying similar structures.

In the proposed method, a query is generated from structural data of more than one interaction site to extract features that are essential for interaction.

2 Profile of an Interaction Site

The structures of the interaction sites binding to a specific compound exhibit common features in their structures. Such features can be regarded as a query of the retrieval, and we call such features a profile of an interaction site. In other words, the profile is defined as a set of common atoms at some

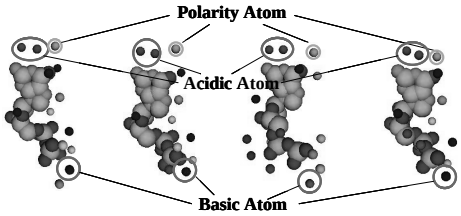

Figure 3: Classification according to physical properties of atoms

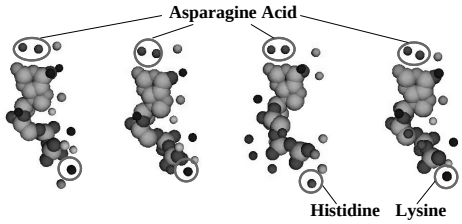

Figure 4: Classification according to the type of amino acid residue to which the atoms belong

interaction sites binding to the same compound. The input of the retrieval method's input is assumed to be more than one interaction site, and the profile obtained from them is regarded as a retrieval query.

The profile can be defined in more detail by employing some actual interaction sites.

Classification of atoms

Figures 3, 4 and 5 show four actual interaction sites, all of them binding to compound GDP. In the center of each figure, the object that consists of combined spheres indicates the compound GDP, while small spheres around the compound represent atoms at the protein's interaction site. The atoms that compose the interaction site are characterized by certain physical properties, namely acidity, basicity, and polarity. These physical properties are determined from the amino acid residue to which the atoms belong, and the constituent elements. As Figure 3 shows, an atom with certain physical properties often appears at almost the same position in different interaction sites. Therefore, to generate a profile from more than one interaction site, it is necessary to classify each atom in terms of its physical properties.

An atom that has the same physical properties and also belongs to the same type of amino acid residue exists at almost the same position in several different interaction sites. As Figure 4 shows, physical properties of the atoms located lower down in the figure are the same, but the types of amino acid residue to which these atoms belong are different. However, the atoms located at the top of the figure have the same physical properties and also belong to the same type of amino acid residue.

Therefore, in order to extract common atoms for the profile, it is necessary to classify each atom by not only its physical properties but also the type of amino acid residue to which it belongs.

Positions of atoms

In Figure 5, atoms commonly found in some interaction sites are circled. The relative positions of some common atoms are unchanged in coordinates, although relative positions of other atoms do change, as the figure indicates. Therefore, it is necessary to specify areas in the profile where each atom can exist.

Moreover, the atoms that appear at limited interaction sites are assumed not to play an important role in interaction, and should be ignored.

Formal definition

The profile is composed of a set of common atoms at some interaction sites that bind to the same compound. The positions of the atoms in the profile are relatively expressed by the distance matrix between each atom. The profile P is formally defined as follows.

$P = \langle \{a_i\}, D \rangle \quad (0 \leq i \leq n)$
$a_i = \langle x, y, z, r, p \rangle$
$p \in (H \cup K)$
n: the no. of atoms in the profile
a_i: the atom i in the profile
p: the property of the atom
H: the physical property of the atom
K: the type of amino acid to which the atom belongs
r: the radius of the sphere where atom i can exit
D: the distance matrix between each atom
x, y, z: the center coordinates of the atom

3 Retrieval Method using Profiling

Figure 6 illustrates an the retrieval method using the profile. The structures are compared at the atomic level between the profile and the structural data of proteins in the database. The protein with a local structure similar to the structure of the profile is retrieved by superimposing the profile on the proteins in the database. First, any atom that is

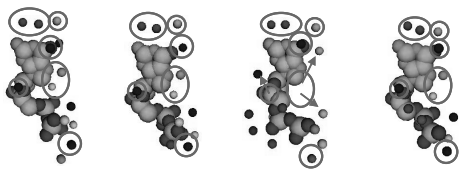

Figure 5: Positions of atoms

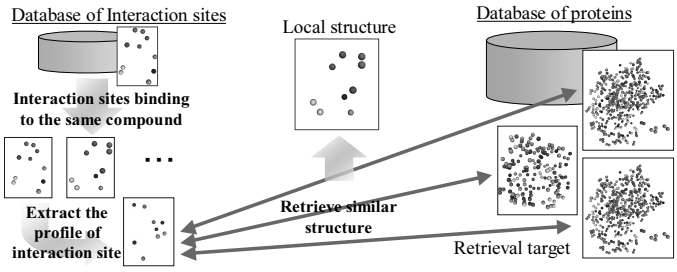

Figure 6: Outline of the retrieval method using a profile

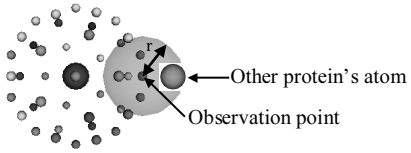

Figure 7: A sphere that centers on the observation point

of the same type as the atom in the profile is extracted from a target protein in the database. Next, the atom that has the same condition as the rest of the atoms in the profile is extracted from the target protein. Here, the condition regarding the atom refers to the type of atom and whether the atom exists in a feasible area. Finally, the group of atoms satisfying the above condition is extracted from the target protein and considered as having a similar structure to the profile.

Improvement of retrieval efficiency

A given protein in the database is composed of several thousand atoms. However, the portion of atoms buried inside the protein cannot constitute an interaction site. Therefore, specifying these internal atoms and excluding them helps to improve the efficiency of matching and the accuracy of retrieval by reducing unnecessary output.

Internal atoms of the protein are identified based on the fact that they are surrounded by other atoms. Let α be an atom to be identified. A sphere of radius $R[\mathring{A}]$ whose center is at α is assumed, and 43 points are placed evenly over the sphere. These points are called observation points. The observation points correspond to the vertices of a polyhedral structure called the Geodesic Dome [4].

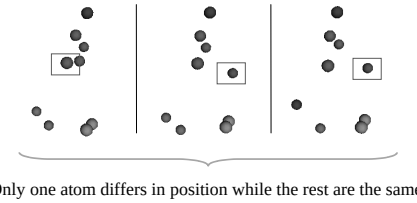

Figure 8: Example of structures that are almost the same

Next, we assume spheres of radius $r[\mathring{A}]$ exist whose centers lie at observation points, and check whether other protein atoms exist in each sphere, as Figure 7 shows. If any atom is observed in all spheres whose centers are at the observation points, α is regarded as an internal atom.

We conducted an experiment excluding internal atoms from consideration in the search process. Results indicate that 13%-23% of atoms other than those belonging to interaction sites were deleted from the retrieval target, which implies that this method adequately excludes only the internal atoms.

Integration of similar outputs

Some local structures that consist of almost the same atoms are often obtained as retrieval results. Figure 8 shows an example of this. These structures should be integrated into one structure.

To integrate the structures, the degree of similarity between the structures of the output is introduced using the Jaccard coefficient method. The degree of similarity J_{ij} between a local structure i and j is defined as follows:

$$J_{ij} = \frac{N_{ij}}{N_{ij} + N_i + N_j},$$

where N_{ij} is the number of atoms included in both the local structures i and j, N_i, and N_j represent the number of atoms included only in the local structures i or j. Then, the degree of dissimilarity D_{ij} is defined as follows:

$$D_{ij} = 1 - J_{ij}.$$

The local structures that have a high degree of similarity are integrated by clustering using D_{ij}. The longest-distance method is used for calculating the degree of dissimilarity between clusters.

The clustering may be performed until the degree of dissimilarity between each cluster does not become smaller than threshold T. This T is usually set to about 0.667, which means that local structures in a cluster share more than half the atoms, assuming that the number of atoms belonging to the local structure in a cluster is equal.

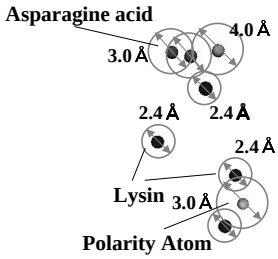

Figure 9: Profile of the interaction sites that bind to compound GDP

4 Experiment

To evaluate the accuracy of retrieval using the profile as a query, we performed a retrieval experiment by applying actual interaction site structural data. For comparison, we also conducted a retrieval experiment using structural data on a single interaction site as a query (non-profiled query).

An overview of the experiment is as follows:

- The profile is manually defined referring to ten interaction sites that bind to compound GDP. Figure 9 shows an example of the defined profile, in which virtual atoms and their possible areas and properties are indicated.
- In the retrieval using the non-profiled query, five interaction sites are used that bind to compound GDP.
- The retrieval target is 75 proteins that include fifteen proteins (A) that bind to compound GDP. These proteins are different from the ten proteins used to make the profile.
- The number of proteins that have been detected (B), the number of proteins whose interaction site has been detected correctly (C), and retrieval time, are counted.
- Recall and precision are calculated by the following definition.
  recall (%) = 100 ×(C)/(A)
  precision (%) = 100 × (C)/(B)
- In the retrieval using the non-profiled query, atoms are classified by their physical properties, and the permissible error margin is set to a constant value (1.5 Å).

Table 1 shows the results. By using the profile, more interaction sites are detected correctly compared to using the non-profiled query. Some interaction sites that are not found through using the non-profiled query can be detected, which demonstrates the validity of the proposed method. Figure 10 shows some examples of the structures detected by employing the profile.

5 Conclusion

In this paper, we proposed a method for retrieving a protein with a local structure similar to the known interaction site structure. In this method, the profile of the interaction site is defined and applied as a retrieval query instead of using a single interaction site. Using the proposed method, some interaction sites that are not detected by using the non-profiled query are correctly detected.

Our future work is to develop a method of automatically extracting the profile from some interaction sites.

Table 1: Results of the retrieval experiment

Query	(B)	(C)	recall (%)	precision (%)	time (m)
Profile	14	8	57.1	53.3	65
Non-profiled	25	1	4.0	6.7	361

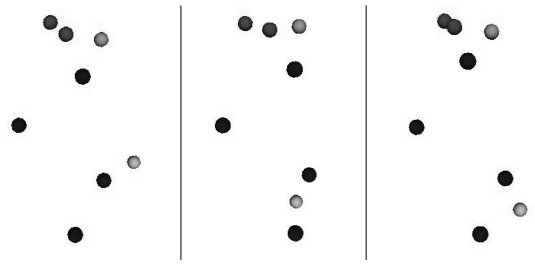

Figure 10: Output of retrieval using the profile

Acknowledgements

The authors wish to thank Prof. Norihisa Komoda, who offered valuable discussions related to this research. A part of this research was supported by BIRD of the Japan Science and Technology Corporation and Grant-in-Aid for Scientific Research.

References

[1] G. Kawamura, G. Nagakawa, and T. Ohkawa: "Development of Protein-Compound Interaction Database on Grid Data Service Using the Three-dimensional Structure Data of Complex," in *Abstracts of Pacific Symposium on Biocomputing 2004 (PSB2004)*, p. 87 (2004).

[2] M. Ishiguro and S. Imajo: "Modeling Study on Hydrolytic Mechanism of Class A β-Lactamases," *J Med Chem*, Vol. 39, pp. 2207–2218 (1996).

[3] R. C. Wilmouth, K. Edman, R. Neutze et al: "X-ray Snapshots of Serine Protease Catalysis Reveal A Tetrahedral Intermediate," *Nat Struct Biol*, Vol. 8, pp. 689–694 (2001).

[4] H. Kenner: Geodesic Math and How to Use It, University of California Press (2003).