KP206

Automated Classification of Quaternary Structure of Protein Based on Domain Pattern

Hiroyuki OGAWA, and Hiroaki KATO

Department of Knowledge-based Information Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi 441-8580 JAPAN

kato @ tutkie.tut.ac.jp

1. Introduction

It is well known that the three-dimensional (3D) structure of a protein is closely related to its function. A spherical unit called the protein domain is formed by a combination of secondary structure. These domains are assumed as the functional and evolutional unit of protein molecules [1]. Tertiary structure is the overall shape of a single polypeptide chain, and sometimes contains of several domains. Furthermore, some proteins such as *hemoglobin* possess a quaternary structure, which is resulted from the interaction of more than one protein chain. These structures are also important to understand more protein complex functions [2]. To understand protein structure, the structure classification database SCOP and CATH are widely used. However they were based on the tertiary structure level or domain level. In the present work, the authors have investigated the relationship between domains and constructed the quaternary structure classification database.

2. Construction of quaternary structure classification database

2.1 Dataset

In the present work, the authors have referred the domain information defined by SCOP database (Ver. 1.73) [3]. There are 92,927 domains and the corresponding 34,495 PDB entries. Although the SCOP classified protein structures on several hierarchical levels, the top level “class” was used. and retrieved only four major classes which contained all-α, all-β, α/β, and α+β (denoted as class-a, b, c, and d, respectively). The number of proteins which consisted by only these four major classes are 30,995 and the total number of domains is 85,685.

2.2 Classification by the number of domains

At first, we have investigated the distribution of the number of domains for each protein. As the result, 13,271 proteins contain exactly one domain. On the other hand, there are some proteins which contained many domains, such as coat protein of virus, or RNA-protein complex in ribosome. In the present work, the proteins which consisted of 2-12 domains were focused and classified based on the number of domains (Fig. 1).

2.3 Classification by the domain patterns

These proteins consisted multiple domains were divided two major classes: homo and hetero complexes. Here, we used four major SCOP classes to district domains. For example, 4-domain proteins were home complex (e.g. *a-a-a-a*) and hetero complex (*a-b-c-d*). Although there are 35 classes for 4-domain proteins from a combination, only 31 classes were found corresponding entries (Fig. 2).

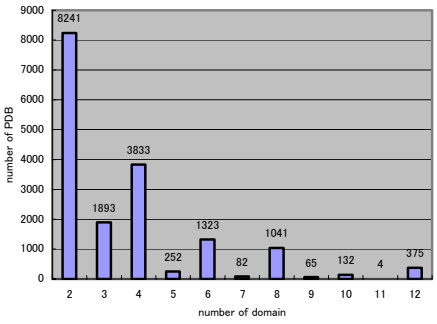

Fig. 1 The distribution of the number of domains.

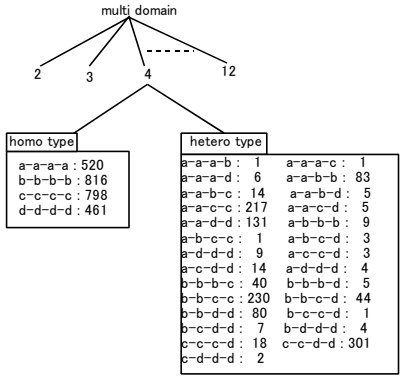

Fig. 2 Basic concept of the quaternary structure classification in the present work.

3. Structural feature analysis using the classification database

As mentioned above, *human hemoglobin* (PDB-ID: 2HHB) is consisted of four chains and belonged to four all-α domains. As shown in Fig.2, the corresponding entry of our database, "*a-a-a-a*" class, contains 520 proteins. From the database, *plectin* (1SH5), *chaoerinin* (1A4O), and *peroxidase* (1PAX) were identified which have similar domain pattern with *hemoglobin*. The common structural feature between them could not detect using the former approaches based on the tertiary structure level, or on the number of chains. These results show the potential applicability of this method for quaternary structural feature analysis classification of proteins. Geometrical information between domains is also used in the database.

Hemoglobin (2HHB) *Plectin* (1SH5) *Chaperonin* (1A4O) *Peroxidase* (1PAX)

Fig. 3 Structural view of 4-domain proteins (class *a-a-a-a*) .

References

[1] Branden, C. and Tooze, J., *Introduction to Protein Structure*, Garland Publishing (1991).

[2] Susan J., et al, *Proc. Natl. Acid. Sci,* **93**, 13-20, (1996).

[3] Murzin, A. G., et al., *J. Mol. Biol.*, **247**, 536-540 (1995).