Extraction of Substructures of Proteins Essential to their Biological Functions by a Data Mining Technique

Kenji Satou, Toshihide Ono[1], Yoshihisa Yamamura[1], Emiko Furuichi[2], Satoru Kuhara[1], and Toshihisa Takagi

Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shiroganedai, Minato-ku, Tokyo 108, Japan. {ken,takagi}@ims.u-tokyo.ac.jp

[1]Graduate School of Genetic Resources Technology, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812, Japan. {tosihide,yamamura,kuhara}@grt.kyushu-u.ac.jp

[2]Fukuoka Women's Junior College, 4-16-1 Gojo, Dazaifu, Fukuoka 818-01, Japan. emiko@grt.kyushu-u.ac.jp

Abstract

Correlation between the sequential, structural, and functional features of proteins is one of the most important open questions in the field of molecular biology. To this problem, we apply a technique known as data mining for discovering associations across protein sequence, structure, and function. We were able to find various association rules on the substructures essential to some protein functions. Moreover, structure-structure associations were found between proteins having different functions. The results suggest that data mining might be a powerful tool in protein analysis.

Introduction

Currently, the number of PDB(Bernstein et al. 1977) entries has increased to over five thousand, and hundreds of folding patterns are known to be different. It means that now the greater part of possible folding patterns are already-known. Actually, reports on the determination of protein structures with new folding patterns dissimilar to any in the PDB are big news in these days. In this situation, statistical and machine learning techniques can give full play to their ability for finding useful knowledge from the large amount of protein data. Now we have the opportunity of finding unknown relationships among different levels (sequence, structure, and function) of protein by the techniques with exhaustive database search against huge and heterogeneous protein data.

About this topic, we have been investigating the application of a technique of data mining called "Discovery of *association rules*". This technique, first developed by Agrawal et al(Agrawal, Imielinski, and Swami 1993), discovers a kind of propositional logic rule such as **bread, butter** ⇒ **milk** from collections of transaction data gathered by retail sellers. In this example, the association rule means that "If a customer buys bread and butter together, then the customer tends to buy milk also". Unlike example-based machine learning techniques including Inductive Logic Programming (Muggleton et al. 1992), association rule discovery does not need target concept to be learned, therefore it can point out unexpected information.

The technique has been applied to various fields with excellent results, however, there have been few applications in bioinformatics. Starting from a preliminary experiment on the co-occurrence of signals in mammalian promoter sequences(Shibayama, Satou, and Takagi 1995), we tried to apply the technique to the discovery of association rules spreading over sequence, structure, and function of proteins. Consequently, the technique turned out to be promising for finding knowledge from heterogeneous protein data(Satou et al. 1997). In this study, we 1) doubled the number of proteins which were fairly chosen based on reasonable criteria, 2) replaced the algorithm of finding association rules by a more powerful one, and 3) adopted additional criteria for eliminating unnecessary association rules. As the result of these enhancements, interesting associations were discovered and some of them were connected directly with active sites.

System and Methods

Basic Framework of Association Rule

The basic algorithm of association rule discovery can be outlined as follows.

Phase 1: Generation of large itemsets

As input, the algorithm takes a table of bit vectors like the following one:

trans_ID	bread	butter	rice	milk	sauce
1	1	1	0	1	0
2	0	1	0	0	1
3	1	0	0	0	1
4	1	1	0	1	1
5	1	1	1	0	0

where the attributes are the items in retail selling, row IDs are the transaction IDs, and the binary values tell whether an item was bought in a transaction. 1 (or 0) means "bought (or not bought)". From this table, for any combination of 2 items, the algorithm first counts the number of rows, called *support*, in which the 2 items were bought together. Here the 2-itemsets, whose supports are lower than a given threshold value called *minimum support*, are discarded. Then, from the n-itemsets that survive, $(n+1)$-itemsets are generated by adding new items to them. Again, the generated itemsets are tested by the threshold value, and unqualified ones are discarded. This iteration is continued until no larger itemset survive.

Phase 2: Generation and refinement of association rules

From all the itemsets that survive from phase 1, association rules are generated by choosing an item in an itemset as a head in an association rule. In the case where two or more items can occur in the head (to the right of the arrow), the rule is called *multi-headed*.

If the *confidence* of an association rule, that is, *(support of head and body)/(support of body)* is lower than a threshold value called *minimum confidence*, the rule is discarded. Furthermore, there exist statistical or user-defined ways of eliminating redundant or insignificant rules. Corresponding to the term "data mining", the elimination of such rules is called "refinement", which has been actively studied(Klemettinen et al. 1994)(Srikant and Agrawal 1996).

Figure 1 illustrates the experiment performed in the previous study using the above technique:

protein	sequence feature1	sequence feature2	structure feature1	function1	function2
p1	1	0	1	0	1
p2	0	0	1	1	0
p3	1	0	0	1	0
p4	1	0	1	1	1
p5	1	1	1	0	0

⇓ Data Mining

`sequence feature1, structure feature1 ⇒ function2 (support=2, confidence=66.6%)`

Figure 1: Sketch of data mining on data on proteins

Enhancements of System and Data for Mining

As in the previous study, we used PDB, SWISS-PROT (Bairoch and Apweiler 1996), and PROSITE(Bairoch, Bucher, and Hofmann 1995) entries as data sources for characterizing proteins from sequential, structural, and functional viewpoints. The data are related to each other by using PDB entry names as keys. The following table illustrates the data assembled for mining.

pdb code	ST=225	SPPR= FABP	EC3= 6.3.2	EC2= 4.2	SPKW= SIGNAL	···
1aaj	0	0	1	0	1	···
1aak	1	1	0	0	0	···
1abe	1	0	0	0	0	···
···	···	···	···	···	···	···

In this table, `SPKW=···` and `SPPR=···` represent a SWISS-PROT keyword and a PROSITE motif, respectively. `EC2=···` and `EC3=···` are level 2 and level 3 classifications of enzymes based on EC numbers. `ST=···` represents a set of 3-stranded substructures which are recognized as similar by a deductive database system PACADE (Satou et al. 1996). The similarity search function of PACADE and how it generates this kind of item are detailed in (Satou et al. 1996) and (Satou et al. 1997), respectively.

Though these settings are the same as the ones used in the previous study, we doubled the number of proteins for a more extensive experiment. Starting from the 360 PDB entries, 253 out of the 360 had corresponding SWISS-PROT entries, 193 had matching PROSITE motifs, and 159 were enzymes with EC numbers. Moreover, we fairly chose the 360 entries from PDB Release 76 based on the following reasonable criteria.

- All the entries have coordinate data.
- All the entries are neither NMR nor model data.
- All the entries have resolutions better than 2.5Å.
- All the entries are not DNA/RNA data.
- All the entries have under 70% sequence homology to each other.
- All the entries have more than 100 residues.

Furthermore, we enhanced our data mining system by adopting the Apriori algorithm(Agrawal and Srikant 1994) and the rule generation algorithm for multi-headed association rules. In addition to the minimum support and confidence, the following elimination criteria against redundant or insignificant rules were implemented in phase 1 and 2 to get interesting rules selectively.

maximum support

We experienced that existence of items with too high support is quite harmful in finding associations from protein data. To avoid a flood of insignificant variant rules caused by such items, we adopted a threshold value called *maximum support* in phase 1.

trivial rules

Suppose that the support of the head itemset is 10 times as high as the support of the body itemset in an association rule. In this case, the body might be one of the necessary conditions of the head, however, it explains only 10% of the whole. To avoid

this problem, we restricted the generation of a association rule in phase 2 if the support of its head is higher than the support of its body.

rules obvious from background knowledge

Suppose that we want to discover information by a comparison between two multi-leveled classifications, e.g. occupation and education of people. In this example, an association rule **Berkeley => Computer Engineer** may be suggestive, however, **Berkeley => University of California** and **Computer Engineer => Engineer** are not worthy of note since these implications are obvious from the classifications. We faced this sort of problem in relation to the items about EC numbers, and implemented a criterion for elimination.

Experimental Results

Using 4858 sequential, structural, and functional features on the 360 proteins described in the previous section, we performed association rule discovery with the following criteria.

- **For phase 1**
  - minimum support = 4 (proteins)
  - maximum support = 30 (proteins)
- **For phase 2**
  - minimum confidence = 65%
  - eliminate the rules including both of **EC2=X.X** and **EC3=X.X.X** items
  - eliminate the rules whose supports of their heads are greater than the ones of their bodies

As a result of the mining, 15324 multi-headed association rules were generated. Due to limitations of space, we describe only 3 expressive results in this section, instead of detailed discussion on all the 15324 rules.

Aspartic Endopeptidases

From a 5-itemset {**ST=383,ST=824,SPPR=ASP_PROTEASE,EC3=3.4.23, SPKW=ASPARTYL PROTEASE**}, 26 rules with support=4 and confidence=100% were generated. The following rule is an example.

```
ST=383,ST=824,SPPR=ASP_PROTEASE
   => SPKW=ASPARTYL PROTEASE,EC3=3.4.23
```

The proteins supported by the rules are CHYMOSIN B(1cms), CATHEPSIN D(1lya), PENICILLOPEPSIN(1ppm), and PEPSIN(4pep).

Since each item occurred equally between heads and bodies in the rules, it can be said that all of them are perfectly co-occurring and there is no explicit implication among the items. Figure 2 is a graphical display of two kinds of similar substructures **ST=383** and **ST=824** specific to aspartic endopeptidases. It is derived from this result that "a protein has both of these $\beta\beta\alpha$ type substructures iff it has a biological function of aspartic endopeptidase". Interestingly, the substructures

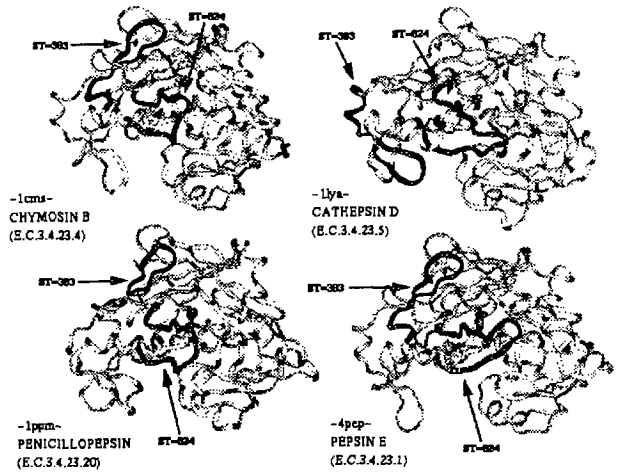

Figure 2: Aspartic endopeptidases

were not part of the well-known active site of aspartic endopeptidase. We think that the substructures extracted in this experiment might be props (skeletal substructures) making up the global structure of aspartic endopeptidase.

Calcium Binding Proteins

From a 4-itemset {**ST=128,ST=1174,SPPR=EF_HAND, SPKW=CALCIUM-BINDING**}, the following 4 rules with support=5 and confidence=83.3% were generated.

```
ST=128,SPKW=CALCIUM-BINDING,SPPR=EF_HAND
    => ST=1174
ST=128,SPKW=CALCIUM-BINDING
    => ST=1174,SPPR=EF_HAND
ST=128,SPPR=EF_HAND
    => ST=1174,SPKW=CALCIUM-BINDING
ST=128 => ST=1174,
   SPKW=CALCIUM-BINDING,SPPR=EF_HAND
```

The proteins supported by the rules are PARVALBUMIN B(1cdp), ONCOMODULIN(1rro), ALPHA-PARVALBUMIN(1rtp,5pal), and SARCOPLASMIC CALCIUM-BINDING PROTEIN (2scp). This result makes an excellent contrast with the one on aspartic endopeptidases in two points. First, the two extracted substructures, which were judged to be "essential and common to calcium binding proteins" by the data mining system, turned out to agree with well-known active sites with a sequential motif called an EF-hand (figure 3). Second, explicit implications among the items were observed, that is, the item **ST=128** occurs only in the bodies of the rules, while **ST=1174** only in the heads. This was caused by the existence of a protein TROPONIN-C(5tnc) which has **ST=128** but lacks **ST=1174**. Therefore, the above rules indicate that "a calcium binding protein does not always have **ST=1174** even if it has **ST=128**". Actually, TROPONIN-C(5tnc) has one more substructure with an EF-hand motif,

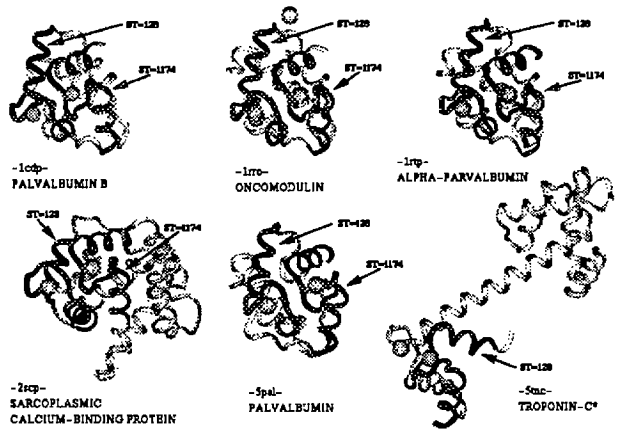

Figure 3: Calcium binding proteins

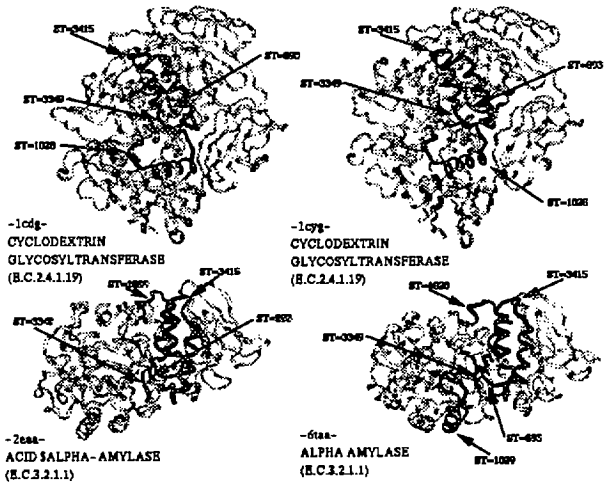

Figure 4: Similar substructures in proteins with different functions

which corresponds to **ST=1174**. However, since the substructure has a shape somewhat different from the ones in **ST=1174**, it was not recognized as similar to them by PACADE.

Similar Substructures in Proteins with Different Functions

From a 4-itemset {**ST=893,ST=1028,ST=3349,ST=3415**}, 10 rules with support=4 and confidence=100% were generated. The following rule is an example.

ST=893,ST=1028 => ST=3349,ST=3415

The proteins supported by the rules are CYCLODEXTRIN GLYCOSYLTRANSFERASE(1cdg), CARBONYL REDUCTASE(1cyd), and ALPHA AMYLASE(2aaa,6taa). As in the case of aspartic endopeptidases, all the items occurred equally between heads and bodies in the rules and explicit implications among them were not observed. However, this result is distinct from the above two results because the rules involve proteins with different functions and global structures (figure 4).

Conclusion

In this study, 3 expressive results concerning essential substructures common to sets of proteins were obtained by using a data mining system. In the case of calcium binding proteins, the extracted substructures corresponded to the active sites, though not in the case of the aspartic endopeptidases. Moreover, co-occurring substructures were found in the proteins with different shapes and functions.

Acknowledgments

We thank Mr. Hiroshi Noguchi at the Japan Advanced Institute of Science and Technology (JAIST) for providing the Apriori program which is used in this paper after modifications. This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science," from the Ministry of Education, Science, Sports and Culture in Japan.

References

Agrawal,R., Imielinski,T., and Swami,A.N. 1993. *ACM SIGMOD*, pp.207-216.

Agrawal,R. and Srikant,R. 1993. *VLDB*, pp.487-499.

Bairoch,A. and Apweiler,R. 1996. *NAR*, Vol.24, pp.21-25.

Bairoch,A., Bucher,P., and Hofmann,K. 1995. *NAR*, Vol.24, pp.189-196.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T., and Tasumi,M. 1977. *JMB*, 112, pp.535-542.

Klemettinen,M., Mannila,H., Ronkainen,P., Toivonen,H., and Verkamo,A.I. 1994. *3rd International Conference on Information and Knowledge Management*, pp.401-407.

Muggleton,S., King,R.D., and Sternberg,M.J.E. 1992. *Proc. of the Twenty-Fifth Hawaii International Conference on System Sciences*, Vol.1, pp.685-696.

Satou,K., Shibayama,G., Ono,T., Yamamura,Y., Furuichi,E., Kuhara,S., and Takagi,T. 1997. *Pacific Symposium on Biocomputing'97*, pp.397-408.

Satou,K., Furuichi,E., Hashimoto,S., Tsukamoto,Y., Kuhara,S., Takagi,T., and Ushijima,K. 1996. *Journal of Japanese Society for Artificial Intelligence*, Vol.11, No.3, pp.440-450.

Shibayama,G., Satou,K., and Takagi,T. 1995. *Proc. of Genome Informatics Workshop 1995*, pp.108-109.

Srikant,R. and Agrawal,R. 1996. *SIGMOD'96*.