BIOINFORMATICS

Vol. 18 no. 5 2002
Pages 705–714

Extraction of knowledge on protein–protein interaction by association rule discovery

T. Oyama[1,*], *K. Kitano*[1], *K. Satou*[2] *and T. Ito*[3]

[1]*INTEC Web and Genome Informatics Corporation, 3-23, Shimoshin-machi, Toyama 930-0804, Japan,* [2]*School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan and* [3]*Cancer Research Institute, Kanazawa University, 13-1 Takara-machi, Kanazawa, Ishikawa 920-0934, Japan*

Received on June 10, 2001; revised on October 26, 2001; accepted on December 3, 2001

ABSTRACT

Motivation: Protein–protein interactions are systematically examined using the yeast two-hybrid method. Consequently, a lot of protein–protein interaction data are currently being accumulated. Nevertheless, general information or knowledge on protein–protein interactions is poorly extracted from these data. Thus we have been trying to extract the knowledge from the protein–protein interaction data using data mining.

Results: A data mining method is proposed to discover association rules related to protein–protein interactions. To evaluate the detected rules by the method, a new scoring measure of the rules is introduced. The method allowed us to detect popular interaction rules such as 'An SH3 domain binds to a proline-rich region.' These results indicate that the method may detect novel knowledge on protein–protein interactions.

Contact: oyama@isl.intec.co.jp

INTRODUCTION

Protein–protein interactions are fundamental biochemical reactions in the organisms and play an important role since they determine the biological processes. Therefore comprehensive description of protein–protein interactions would significantly contribute to the understanding of biological phenomena. The yeast two-hybrid system provides a powerful method to detect protein–protein interactions occurring in all possible combinations between given proteins, and was recently applied to the comprehensive analysis of the interactions among the budding yeast proteins (Ito *et al.*, 2001a,b; Uetz *et al.*, 2000). As a result, a large quantity of information about protein–protein interactions has been accumulated. However, general information or knowledge about interactions is poorly extracted from the numerous interactions identified using this method. In other words, we do not understand general rules so much about the relation between proteins interacting with each other, such as 'The protein having the feature α interacts with the protein having the feature β' or, more directly, 'This domain interacts with that domain' and so on. Thus, we are trying to discover such rules or knowledge from the interaction data through their analysis.

The technique to extract useful information or knowledge hidden in vast data, which is called 'data mining', attracts a great deal of attention, and much research applying data mining to bioinformatics have already been completed. Satou *et al.* applied 'Discovery of Association Rule' method (Agrawal *et al.*, 1993), which is a popular method of data mining, to find rules describing the association among heterogeneous genome data such as sequences, structures and functions of the proteins extracted from various genome databases (Satou *et al.*, 1997a,b). They found some association rules stating sequential, structural and functional aspects of two kinds of endopeptidases. They also illustrated three expressive results concerned with essential substructures common to sets of proteins. Moreover many studies to discover novel knowledge from profile data of microarrays using data mining techniques are reported (Zweiger, 1999; Lemkin *et al.*, 2000). As described above, data mining is effective for bioinformatics.

With regard to data mining using protein–protein interactions, Fellenberg *et al.* developed a system for the integrative analysis of protein interaction data (Fellenberg *et al.*, 2000). Using their system, they tried to predict the function of so far uncharacterized proteins. However, their system is intended to reveal the function of individual protein, not the general rule or knowledge about the relation between proteins interacting with each other.

We studied a method supporting the discovery of knowledge relating to protein–protein interactions from accumulated protein–protein interaction data, using the

*To whom correspondence should be addressed.

'Association Rules Discovery'. The method proposed here reveals the tendency or relation between mutually interacting proteins in the form of the association rule. The method detects a lot of rules of various types. They include some valuable and novel rules, which would lead to appropriate experimental validation and provide novel insights on protein interactions. One of the most desirable rules is a 'binding region predicting rule' that indicates the protein regions involved in the interactions. Of course, various other types of rules may well be detected.

In this paper, we describe the details of our method and the result of our attempt to apply the method to the interaction data of the yeast *Saccharomyces cerevisiae*.

SYSTEMS AND METHODS

Discovery of association rules

'Discovery of Association Rules' is one of the popular methods of data mining to detect the association between items. Suppose the following table gathered by a retailer.

Trans_id	Bread	Butter	Rice	Milk
1	1	1	0	1
2	0	1	0	0
3	1	0	0	0
4	1	1	0	1
5	1	1	1	0

In this table, there are four items, 'bread', 'butter', 'rice' and 'milk', and the value 1 (or 0) means that the item was bought (or not bought) by the customer corresponding to each transaction 'trans_id'. From this table, an association rule 'bread & butter → milk' is detected, which represents 'if a customer bought bread and butter, then he/she also bought milk'.

An association rule has two values, 'support' and 'confidence', representing the characteristics of the rule. Support represents the frequency of co-occurrence of all the items appearing in the rule. For the above rule, support is 2. And confidence represents the accuracy of the rule computed by dividing the support value by the frequency of co-occurrence of if-part items. For the above rule, confidence is 66.6% (=2/3). Discovery of association rule is the method to detect all the possible rules whose supports and confidences are larger than the user-defined threshold values called 'minimal support' and 'minimal confidence' respectively.

Although it takes an enormous amount of time to explore all the possible combinations among the items when the table is large, Agrawal *et al.* proposed a fast algorithm to detect all the rules satisfying the thresholds in a practicable time (Agrawal *et al.*, 1993).

Framework

Figure 1 illustrates the framework of our method. The data for mining is created from thousands of protein–protein interactions and features that characterize each protein appearing in the interactions. Then we extract novel knowledge on protein–protein interaction from the data via mining by association rule discovery algorithm.

Interaction data

An interaction is represented as a pair of two proteins that directly binds to each other. Protein pairs of the interactions are obtained from four sources as below.

YPD (Yeast Proteome Database) YPD (Hodges *et al.*, 1999) is a database of the yeast developed by Proteome Inc. It contains various data for over 6000 yeast proteins including interaction data. We extracted 1841 protein pairs of complexes composed of two proteins from the version as of June 2000 since they could be regarded to bind directly to each other.

MIPS (Munich Information Center for Protein Sequences) MIPS (Mewes *et al.*, 2000) presents a comprehensive database that summarizes current knowledge regarding more than 6000 ORFs encoded by the yeast genome, including lots of interaction data examined by biological experiments. From the version as of September 2000, we extracted 1064 protein pairs that could be identified to bind directly by the method used in biological experiments.

Two-hybrid experiment by Ito et al. Ito *et al.* executed comprehensive analysis using their system to examine two-hybrid interactions in all possible combinations between the 6000 proteins of the budding yeast *Saccharomyces cerevisiae* (Ito *et al.*, 2001a). We extracted 1482 protein pairs that are observed two times or more in the experiment.

Two-hybrid experiment by Uetz et al. Uetz *et al.* also examined protein–protein interaction by the two-hybrid method. They identified 957 interactions (Uetz *et al.*, 2000). We used all of them in our method.

Consequently, we obtained 4307 interaction protein pairs by omitting overlapped interactions.

Features of the yeast proteins

Each of the yeast proteins has various functions or characteristics, which are called 'feature' in this paper. Features of each protein are described in many publications and public genome databases. Thus we provide each protein appearing in the interactions with features according to several genome databases. The features that we used are defined from functional, primary structural and other various viewpoints, and are classified into seven types, which are listed below.

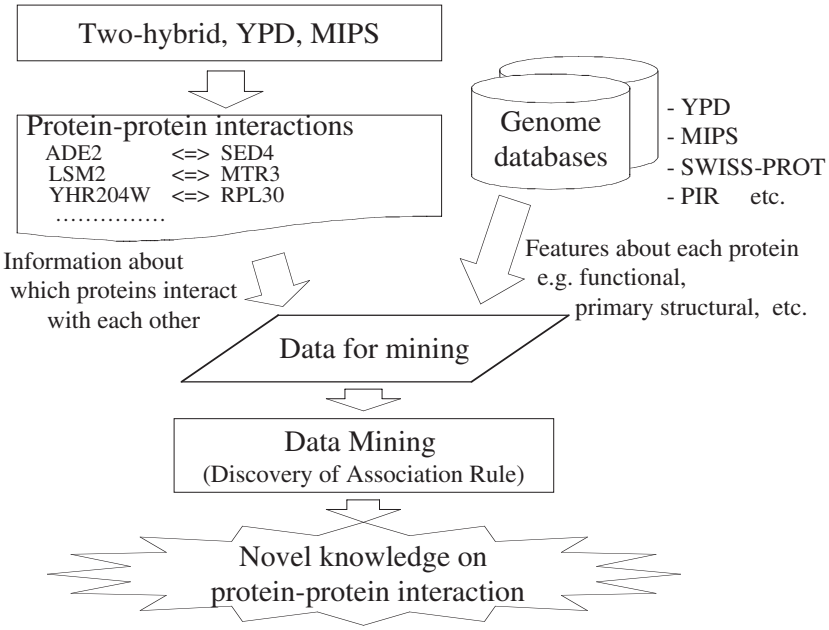

Fig. 1. Framework of our method to discover the knowledge related to protein–protein interactions. Data for mining is created from interaction data and genome databases.

YPD categories In YPD, several categorizations are defined based on some aspects such as cellular role, biochemical function and localization, and proteins are classified into dozens of categories. Features of the first type are defined by YPD categories to which the protein belongs. That is, if a protein belongs to category C, it is assigned a feature stating, 'This protein belongs to C'. Total number of the features of this type is 275, which is the sum of the categories of each categorization.

EC numbers Enzymes are classified based on their functions into four-level hierarchical categories, each of which is labeled by EC number. Some databases have descriptions of EC numbers. We extracted EC numbers described in SWISS-PROT (Bairoch and Apweiler, 2000) of the release 38 and PIR (Barker *et al.*, 2001) of the release 64. Features of the second type are defined by the EC numbers to which the protein corresponds. We used EC numbers in two ways. Namely we characterized each protein based on the second level categorization and third level one. If a protein has EC number 1.2.3.4, it is assigned a feature of 'EC 1.2' and 'EC 1.2.3'. Total number of the features of this type is 154.

SWISS-PROT/PIR keywords Many of the proteins correspond to the entries of SWISS-PROT and PIR, and each entry has some keywords representing functional, structural or other characteristics of the protein. Features of the third type are defined by SWISS-PROT/PIR keywords assigned to the protein. Total number of the features of this type is 491.

PROSITE motifs Features of the fourth type are defined by motifs in PROSITE (Hofmann *et al.*, 1999) of the release 16, found in the amino acid sequence of the protein. Total number of the feature of this type is 698.

Bias of the amino acids It is often observed in many proteins that particular amino acids are clustered in a narrow range on the sequence. Features of the fifth type are defined by such, so to speak, 'bias' of the amino acids. That is to say, if amino acids of a particular kind are rich in a short region on the amino acid residues, the protein is assigned a feature representing it. Total number of the feature of this type is 20.

Segment clusters We considered the features based on the homologous segments. Executing all-versus-all homology search among all yeast proteins with BLASTP (Altschul *et al.*, 1990), we obtained a lot of pairs of homologues segments, in other words partial sequences of amino acids on the protein. Next, we classified homologous segments into clusters based on the result of the homology search, on the condition that the clusters were allowed to be overlapped with each other. As a result, we got lots of clusters. Among them, we found many clusters composed of the segments that a single particular amino acid occupies the majority of the sequence, like 'IQRQQQQQFRHHVQIQQQQQKQQQQQQQ'. How-

ever, such segments are not appropriate to be regarded as homologous because they have been clustered only for the reason that they have many 'Q's, rather than their whole sequences are similar. So we eliminated the clusters containing such segments. Consequently we got 3589 clusters. Then, features of the sixth type are defined by the clusters into which the segments of the protein are classified. That is, if a segment of a protein belongs to a cluster, the protein is assigned to a feature representing it. Total number of features of this type is 3589, equal to the number of the clusters.

Amino acid patterns There exist many patterns of amino acid sequence already known to bind to a specific domain of the protein. For example, pattern 'RxxPxxP' and 'PxxPxR' are generally considered to bind to the SH3 domain. Features of the seventh type are defined by whether the popular amino acid patterns exist or not on the amino acid residues. Total number of the feature of this type is 14.

As a result of characterizing all yeast proteins, we got 5241 features of above seven types.

Data for mining

Data for mining is created combining protein pairs of the interactions and the features of each protein. In many cases of association rule discovery, physical entities such as proteins are regarded as transactions of mining data, like (a) of Figure 2. That is to say, each transaction represents a protein. However, association rules detected from so-called 'protein-based' transaction data merely represent the relations between the features of a single protein, not those between the features of the two proteins appearing in the interactions. Then the protein-based transactions are not appropriate to our purpose, and we decided to regard an interaction itself as a transaction. Namely, each transaction represents an interaction, and has the features of both left-hand side protein and right-hand side protein of the interaction protein pair, which are shown as 'LSP' and 'RSP' in (b) of Figure 2. Mining association rules from 'interaction-based' transactions would detect the rules such as 'if LSP has the feature 1, interacting protein with it, or RSP, also has the feature 2', which is equivalent to 'proteins having the feature 1 interact with proteins having the feature 2'.

Since an interaction is a non-directional binary relation, the interaction such that 'P1 and P2 interact with each other' can be represented in the form of either 'P1 $\Leftrightarrow$ P2' or 'P2 $\Leftrightarrow$ P1'. That is, P1 (or P2) can be placed either in the left side (LSP of Figure 2) or in the right side (RSP of Figure 2). In which side the protein is placed depends on the process of extracting interaction data from genome databases and the process of the two-hybrid experiments. However, this ambivalence of protein placement causes a problem since the result of mining is much influenced by the protein placement. Suppose the case that all the proteins having a particular feature are placed only in the left-hand side and the case that they are balanced on both sides. The results of mining in the two cases are different from each other. In order to avoid this problem, we added new transactions duplicated from the original transactions replacing the side of the proteins (see Figure 3). Then we finally obtained 8003 transactions by duplicating original 4307 interactions and deleting overlapped ones such as 'P1 $\Leftrightarrow$ P1'.

Focusing rules

Rules detected from interaction-based mining data are classified into several types as follows, where α, β and γ mean a feature or a conjunction of some features.

(1) LSP: $\alpha \rightarrow$ RSP: β
(2) RSP: $\alpha \rightarrow$ LSP: β
(3) LSP: $\alpha \rightarrow$ LSP: β
(4) RSP: $\alpha \rightarrow$ RSP: β
(5) Otherwise
(e.g. LSP: α & RSP: $\beta \rightarrow$ RSP: γ
LSP: $\alpha \rightarrow$ LSP: β & RSP: γ)

For example, type (1) means, 'If LSP has the feature(s) α, then RSP has the feature(s) β'. Type (5) is more complicated one containing the features about both LSP and RSP in the left side and/or in the right side of the rule.

The types (1) and (2) representing the relation between interacting two proteins are desirable and interesting for our work, while the types (3) and (4) representing the relation between the features of a single protein are trivial and uninteresting. On the other hand, type (5) is too complicated to interpret. Moreover, type (1) is essentially equivalent to type (2), because rules of both types are equally interpreted as 'proteins having α interact with proteins β'. Therefore we focus only on the rules of type (1).

RESULTS AND PROBLEMS

Applying our method under the condition of minimal support = 9 and minimal confidence = 75% in association rule discovery, 6367 rules were detected. The following two rules are examples of the detected rules, where the former number in brackets represents support and the latter number represents confidence.

(R1) LSP: Localization="Nuclear nucleolus" &
Keyword="nuclease"
$\rightarrow$ RSP: Localization="Nuclear nucleolus"
[14, 100.0%]

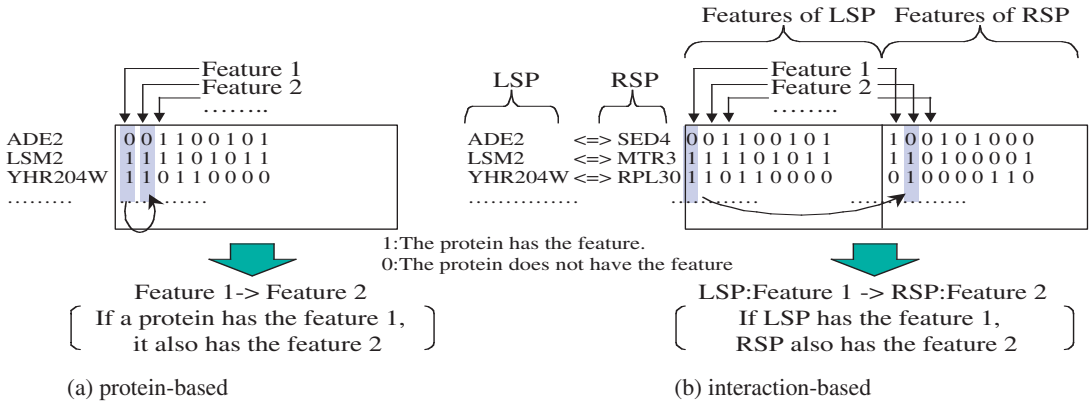

Fig. 2. Protein-based transactions are not appropriate since we focus the relation between interacting two proteins. Then we employed interaction-based transactions.

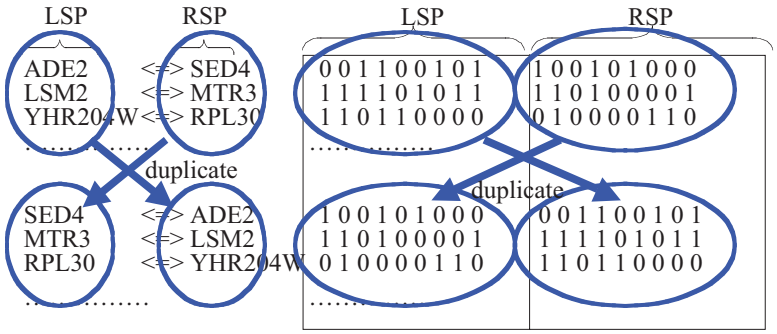

Fig. 3. Mining data is created duplicating original interactions and replacing the side of the protein in order to avoid the ambivalence of protein placement. By this operation, the same result can be obtained in spite of which protein is regarded as LSP (or RSP).

(R2) LSP: Keyword="DNA binding" &
Amino Acid Bias="Acidic(DE)" &
Motif="MADS_BOX_1"
→ RSP: Biochemical Function
="Transcription factor"
[9, 81.8%]

There were 1096 rules in which only the features of YPD categories, EC numbers and SWISS-PROT/PIR keywords appeared like (R1). Those rules are not so interesting since they are so general and not concerned with the features about amino acid sequences, namely since we cannot obtain which domain interacts.

The remaining 5271 rules had at least one feature about sequences such as PROSITE motifs, bias of the amino acids, segment clusters and amino acid patterns like (R2). The (R2) suggests that the 'MADS_BOX' domain may be somehow involved in the interaction with other transcription factors. Indeed, such a possibility was already pointed out in several papers (Mueller and Nordheim, 1991; Bruhn *et al.*, 1992; Bruhn and Sprague, 1994). This example would demonstrate the power of our method, since a phenomenon that seems to occur *in vivo* can be extracted as an association rule.

One of the most interesting rules for biologists is the one that pinpointing the regions involved in the direct protein binding on amino acid sequence. Such rules are interpreted as 'The domain A binds to the domain B'. Only the rules both proteins of which (LSP and RSP) have the features about the sequence patterns (namely, segment clusters, amino acid patterns or PROSITE motifs) can be such 'binding region predicting rules'. However, there were no rules that have sequence patterns for both proteins. This means that binding region predicting rules such as 'An SH3 domain binds to the proline-rich region', which is a popular one, could not be detected.

SH3 domain is registered as a PROSITE motif (PS50002), and the proline-rich patterns that an SH3 domain is generally considered to bind are 'RxxPxxP'

and 'PxxPxR'. Since mining data already contained the features corresponding to SH3 and the pattern 'RxxPxxP or PxxPxR', we considered that 'SH3 rule' should be found. Then we searched mining data for the SH3 rule. As a result, the following rule was detected.

(R3) LSP: Motif="SH3"
→ RSP: Amino Acid Pattern
="RxxPxxP or PxxPxR"
[28, 31.5%]

We found that the confidence of the (R3), 31.5%, was too low to be detected under the condition of minimal confidence = 75%. We wondered why the confidence of the already known rule was so low. Then we checked all the interactions LSP of which had a feature of SH3. There were 89 interactions that had SH3 in LSP. 28 interactions out of them also had proline-rich in RSP, while the rest of the 61 did not have it. 28 interactions having both SH3 and proline-rich supported the (R3), while 61 interactions having SH3 but not proline-rich contradicted the (R3). For the sake of convenience, we called the former interactions 'positive interactions' and the latter ones 'negative interactions' related to the corresponding rule.

Focusing BEM1 as LSP, there were 3 positive interactions and 7 negative ones. As shown in Figure 4, BEM1, which has SH3, interacts with 7 non-proline-rich proteins as well as 3 proline-rich proteins. This fact indicates that BEM1 has some binding domains other than their two SH3 domains, as was recently proved experimentally (Ito *et al.*, 2001b). Then we found the proteins that have more than one binding domain like BEM1 caused a fall of confidence. As many proteins have plural binding domains, they have so much influence on the confidence. However setting minimal confidence low so as to detect valid but low-confidence rules as (R3) would cause the explosion of trivial rules. Therefore it is necessary to use another scoring measure to indicate the validity of the rules instead of the confidence.

SCORING THE RULES

We defined new scoring measures indicating the validity of the target rule 'LSP: α → RSP: β', where α and β represent the features. We thought the new measure should satisfy the following four characteristics.

(C1) *The more positive interactions are, the higher the scores are*

There is no doubt about this characteristic because the reliability is getting higher as the interactions that support the rule increase.

(C2) *The number of negative interactions does not influence the scores*

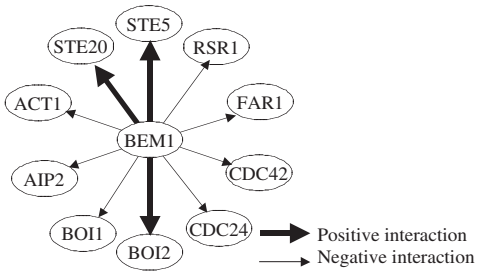

Fig. 4. Interactions related to BEM1 are shown. BEM1 has an SH3 domain. Besides, it intracts with three proteins that have proline-rich regions (positive interactions) and with seven proteins that do not have proline-rich regions (negative interactions)

As illustrated in Figure 4 of the previous section, BEM1 binds to 7 non-proline-rich proteins. It seems contradictory to the SH3 rule since the protein having an SH3 domain interacts with the proteins that do not have a proline-rich region. However such negative interactions do not always reject the SH3 rule. It is because a protein generally has some binding domains, namely because the negative interactions may be concerned with another domain of BEM1 other than an SH3 domain. Therefore we concluded that the existence of the negative interactions should not lower the score although many negative interactions exist.

(C3) *The more proteins concerned with the positive interactions are, the higher the scores are*

Suppose that the target rule is supported by three positive interactions shown by the case 1, 2 and 3 of Figure 5, which illustrates three cases of three positive interactions. In the figure, the circles represent all the proteins that have the target feature α or β and the arcs represent positive interactions. In case 1, six proteins are concerned with the positive interactions, while in the other cases only four proteins are concerned. The target rule is more reliable in case 1 than in cases 2 and 3 although the number of positive interactions in each case is equal. This is because case 1 is supported by more proteins than other cases are. It becomes clearer when thinking of the following question.
In which case, cases 1 or 2, will cause the reliability of the target rule to go higher when considering adding two more interactions to case 4?
In case 1, the interactions composed of entirely different proteins are added. Namely entirely different interactions (or evidences) support the target rule apart from the existing interaction. On the other hand, case 2 contains not entirely different

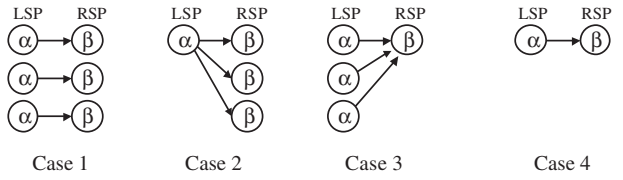

Fig. 5. Although the former three cases equivalently have three positive interactions, case 1 should be given the highest score among them.

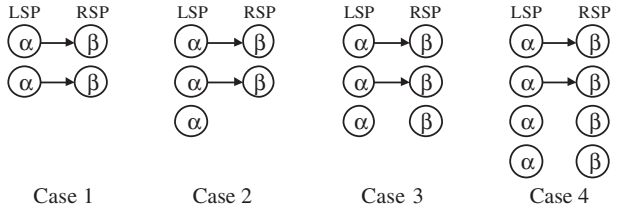

Fig. 6. As the proteins not concerned with the positive interactions increase, the score should get lower.

interactions. So case 1 should be given a higher score than case 2. Thus we concluded that the increase of the concerned protein should force the score higher.

(C4) *The more proteins not concerned with the positive interactions but having the target features are, the lower the scores are*

Suppose the case that the target rule is supported by two positive interactions shown in Figure 6, which illustrates four cases of two positive interactions. In case 1, all the proteins are concerned with the positive interactions. In other words all the proteins support the rule, or no protein rejects the rule. On the other hand, some proteins reject the rule in the other cases. Especially in case 4, half of the proteins reject the rule. Generally speaking, it is natural that the more proteins reject the rule, the lower the reliability of the rule. Therefore we concluded that increase of the non-concerned proteins should force the score lower.

By the way, this characteristic is equivalent to 'The more proteins having the target features, the lower the scores' on the assumption that the number of the concerned proteins does not change.

According to the characteristics described above, the following five values should be related to the definition of the scoring measure.

N: The number of the positive interactions for the target rule

$\hat{\alpha}$: The number of LSP concerned with the positive interactions

$\hat{\beta}$: The number of RSP concerned with the positive interactions

α: The number of the LSP having the feature α

β :The number of the RSP having the feature β

For example, each value is 3, 1, 3, 1 and 3 respectively for case 2 of Figure 5, and 2, 2, 2, 4 and 4 respectively for case 4 of Figure 6. Then using the above five values, characteristics can be translated as follows

(C1′) *The larger the N, the higher the score*

(C3′) *The larger the $\hat{\alpha}$ and $\hat{\beta}$, the higher the score*

(C4′) *The larger the α and β, the lower the score.*

According to (C1′), (C3′) and (C4′), we can easily define a scoring measure S as follows.

$$S = \frac{\hat{\alpha}\hat{\beta}N}{\alpha\beta}$$

However, using this definition, the scores of the cases 1, 2 and 3 of Figure 5 are the same values of 3. It is opposed to the characteristics (C3). The score of case 1 should be larger than that of cases 2 and 3. Then we modified the definition so that the effect of $\hat{\alpha}$ and $\hat{\beta}$ may become larger than that of α and β as follows.

$$S = \frac{\hat{\alpha}\hat{\beta}N}{\sqrt{\alpha\beta}}$$

DISCUSSION

Evaluating the scoring measure with various interaction patterns

To evaluate the scoring measure, we applied the measure to various interaction patterns and examined whether the scores are agreeable for our intuition or not. Figure 7 shows 12 interaction patterns (P1) to (P12) whose scores are mentioned below the graphs.

The scores of (P11) and (P12) are maximal, and the score of (P10) is minimal. That agrees with our intuition.

Now, let us consider the order of (P1) to (P6). Because of symmetry, the scores of (P2) and (P3) should be equal. Comparing (P3) and (P4), since (P3) has two more proteins that reject the rule than (P4), (P4) should have larger score than (P3) according to the characteristics (C4). In the same way, since (P1) has two more proteins that support the rule than (P4), (P1) should have larger score than (P4) according to the characteristics (C3). Moreover, (P6) should have a larger score than (P1) because of (C1), (P5) should have a larger score than (P2) because of

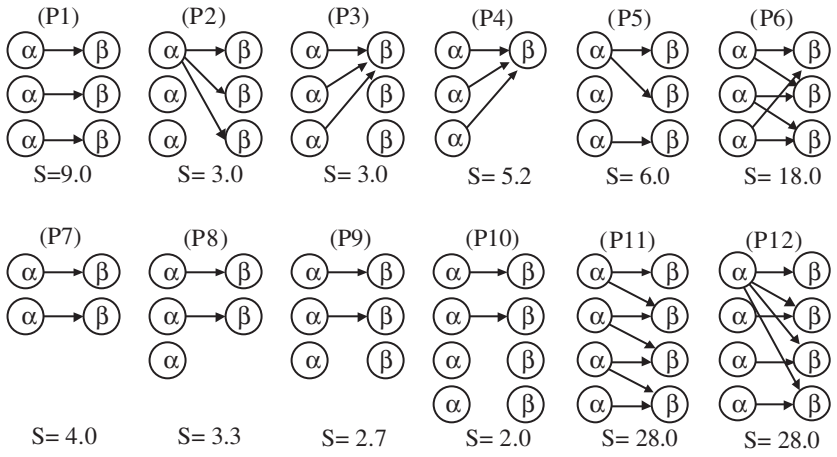

Fig. 7. Various interaction patterns are illustrated. The scores of the patterns are mentioned below the pattern graphs. Generally speaking, (P11) and (P12) should have the largest score, and the actual scores are agreeable with our intuition.

(C3) and (C4), and (P1) should have a larger score than (P5) because of (C3) and (C4). To sum up, the following equations should be satisfied.

$$(P2) = (P3) < (P4) < (P1) < (P6)$$
$$(P2) < (P5) < (P1)$$

We think these equations are agreeable for our intuition and the actual scores of the six patterns also satisfy the equations. However, it is difficult to determine which should be larger (P4) or (P5). The actual scores of (P4) and (P5) are 5.2, and 6.0 respectively, so the order is (P4) < (P5). We consider the order and the values for (P4) and (P5) do not disagree with the intuition significantly.

Next, consider the order of (P7) to (P10). It is easy to determine the order among these patterns. According to our intuition, the following equation should be satisfied and the actual scores also satisfy this equation.

$$(P10) < (P9) < (P8) < (P7)$$

The order among all the patterns with actual scores is as follows.

$$(P10) < \underline{(P9) < (P2) = (P3) < (P8) < (P7) < (P4)}$$
$$< (P5) < (P1) < (P6) < (P11) = (P12)$$

Although we think this order mainly agrees with our intuition, it is difficult to evaluate the adequacy of the order between some cases, which is underlined in the above order. It is a sensitive problem which one should be larger. We consider that adequate order cannot be determined while the actual proteins are not determined. Moreover the difference of the scores is not significant. Therefore we judged it is as not fatal as drawback.

Finally, we compare (P11) and (P12). The two patterns are obviously different although the scores are the same values. The interactions of (P12) are biased, while those of (P11) are uniform. However, it is also difficult to determine which should have a larger score. Therefore we decided not to introduce a viewpoint of bias of the interactions into the definition of the scoring measure.

As a result of examination with several simple interaction patterns, we found our scoring measure agrees with our intuition well.

Applying the scoring measure to the detected rules

We again tried to detect association rules. This time, we used lower minimal confidence so as to detect low-confidence but valid rules. As a result of mining on the condition of minimal support = 9 and minimal confidence = 30%, 99 106 rules were detected. From the 99 106 rules, 4660 rules for both proteins (LSP and RSP), of which had the features about sequences, were selected. Then we applied the scoring measure to the selected rules.

SH3 rule, which was not detected in the prior section, can be detected with the 13th highest score 111.94. The rule of top score is represented by (R4) and the score is 2471.36. The features appearing in this rule are so general that the rule is not considered to be valuable. Many of the rules that have a higher score than SH3 rule were similar to the rule (R4) and hence trivial.

(R4) LSP: Localization="Peripheral membrane"
& Amino Acid Pattern="SP or TP"
& Cellular Role="Vesicular transport"
→ RSP: Localization="Peripheral membrane"
& Amino Acid Pattern="SP or TP"
[94, 40.5%, 2471.36]

Fig. 8. The alignment of the amino acid sequences of segment cluster '001120' is illustrated.

Now, as mentioned above, one of the most interesting things is to identify the binding region of the proteins, namely, the relation between the sequence patterns of interacting proteins, such as PROSITE motifs and conserved sequences. Based on the above interest, we determined the 'rule selecting strategy', that is, selected the rules that have the feature of either PROSITE motifs or segment clusters. According to this selecting strategy, we obtained the 4118 rules out of above 4660 rules. In higher rank of the selected 4118 rules, we found several rules related to the protein kinase like (R5).

(R5) LSP: Motif="PROTEIN_KINASE_ATP" &
Segment. Cluster="001120"
→ RSP: Motif="PROTEIN_KINASE_ST"
[16, 30.2%, 68.07]

Sequences of the segment cluster '001120' and the positive interactions of (R5) are shown in Figures 8 and 9 respectively. Judging from the positive interactions, these rules are considered to represent a part of the phenomena that several protein kinases bind to each other. However, we found the motifs and the conserved pattern in Figure 8 do not directly bind to each other. They are merely the regions commonly appearing among the proteins in Figure 9, but not protein binding regions. The conserved pattern is just an active site of protein kinases. So the (R5) is not a binding region predicting rule but a trivial one only reflecting the fact that some protein kinases interact with other protein kinases. The users have to eliminate such trivial rules based on the biological knowledge.

Besides the kinase rules like (R5), we often found similar rules merely indicating the commonly appearing patterns but not the binding regions *per se*. This is caused by a scarcity of the features that represent protein interaction domains. Therefore, further efforts are required to obtain experimental data on binding regions or to improve the segment clustering so that we can use many more features on binding regions to extract more useful rules.

CDC28 <=> CAK1
CDC28 <=> HSL1
CKA1 <=> CKA2
CKA2 <=> CKA1
FUS3 <=> FUS3
FUS3 <=> STE11
FUS3 <=> STE7
GCN2 <=> GCN2
HOG1 <=> PBS2
HOG1 <=> RCK1
HOG1 <=> RCK2
KSS1 <=> STE11
KSS1 <=> STE7
MKK1 <=> BCK1
MKK1 <=> PKC
MKK1 <=> SLT2

Fig. 9. These 16 interactions support the rule (R5).

CONCLUSION

In this paper, we described a method to detect association rules related to protein–protein interaction from the accumulated protein–protein interaction data using a method of data mining. The normal way to detect association rules using the confidence was not appropriate because of the harmful influence of the proteins that have plural binding domain. Then we employed a new scoring measure evaluating the validity of the rules. The new scoring measure agrees well with our intuition. Using the scoring measure, we demonstrated that the method could detect already known rules such as the SH3 rule. This fact encourages us that novel and valuable knowledge will also be discovered using this method. While no apparently novel 'binding region predicting rules' have been found at present, we expect such rules can be achieved by the proposed method if sufficient features about binding regions are provided. One of the most important works to prepare the required features would be to improve the segment clustering by adding more information such as secondary and tertiary structure of the proteins.

The proposed method also detects trivial and uninteresting rules much more frequently than valuable ones. This is inevitable, because it is impossible to detect only the valuable and interesting rules automatically. The users have to select interesting rules by examining all the detected rules according to their 'selecting strategy' based on the biochemical knowledge and interest of their own. Thus our method supports the discovery of novel knowledge on protein interactions.

The proposed method provides many similar rules that actually represent the same phenomenon but from various points of view. Consequently, it becomes difficult to grasp an outline of all detected rules. In many cases, such rules are supported by similar interaction sets. Therefore, a solution to this problem may be to cluster the rules that have similar interaction sets. Such clustering of rules by interactions ensures the grouping of the rules for the same phenomenon.

ACKNOWLEDGEMENTS

The authors are partly supported by the Institute of Bioinformatics Research and Development, Japan Science and Technology Corporation.

REFERENCES

Agrawal,R., Imielinski,T. and Swami,A. (1993) Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD*, 207–216.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Barker,W.C., Garavelli,J.S., Hou,Z., Huang,H., Ledley,R.S. McGarvey,P.B. *et al.* (2001) Protein Information Resource: a community resource for expert annotation of protein data. *Nucleic Acids Res.*, **29**, 29–32.

Bruhn,L. and Sprague,Jr.,G.F. (1994) MCM1 point mutants deficient in expression of alpha-specific genes: residues important for interaction with alpha 1. *Mol. Cell. Biol.*, **14**, 2534–2544.

Bruhn,L., Hwang-Shum,J.J. and Sprague,Jr.,G.F. (1992) The N-terminal 96 residues of MCM1, a regulator of cell type-specific genes in *Saccharomyces cerevisiae*, are sufficient for DNA binding, transcription activation, and interaction with alpha 1. *Mol. Cell. Biol.*, **12**, 3563–3572.

Fellenberg,M., Albermann,K., Zollner,A., Mewes,H.W. and Hani,J. (2000) Integrative analysis of protein interaction data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 152–161.

Hodges,P.E., McKee,A.H.Z., Davis,B.P., Payne,W.E. and Garrels,J.I. (1999) Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.*, **27**, 69–73.

Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001a) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Ito,T., Matsui,Y., Ago,T., Ota,K. and Sumimoto,H. (2001b) Novel modular domain PB1 recognizes PC motif to mediate functional protein–protein interactions. *EMBO J.*, **20**, 3938–3946.

Lemkin,P.F., Thornwall,G.C., Walton,K.D. and Hennighausen,L. (2000) The microarray explorer tool for data mining of cDNA microarrays: application for the mammary gland. *Nucleic Acids Res.*, **28**, 4452–4459.

Mewes,H.W., Frishman,D., Gruber,C., Geier,B. Haase,D. *et al.* (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **28**, 37–40.

Mueller,C.G. Nordheim,A. (1991) A protein domain conserved between yeast MCM1 and human SRF directs ternary complex formation. *EMBO J.*, **10**, 4219–4229.

Satou,K., Ono,T., Yamamura,Y., Furuichi,E., Kuhara,S. Takagi,T. (1997a) Extraction of substructures of proteins essential to their biological functions by a data mining technique. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 254–257.

Satou,K., Shibayama,G., Ono,T., Yamamura,Y., Furuichi,E., Kuhara,S. Takagi,T. (1997b) Finding association rules on heterogeneous genome data. *Proceedings of the Pacific Symposium on Biocomputing '97*. pp. 397–408.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R. Lockshon,D. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Zweiger,G. (1999) Knowledge discovery in gene-expression-microarray data: mining the information output of the genome. *Trends Biotech.*, **17**, 429–436.