EJB Electronic Journal of Biotechnology ISSN: 0717-3458
Vol.3 No.3, Issue of December 15, 2000

Received September 11, 2000 / Accepted November 17, 2000

RESEARCH ARTICLE

Predicting regulatory elements in repetitive sequences using transcription factor binding sites

Jorng-Tzong Horng*
Department of Computer Science and Information Engineering
National Central University
Taiwan
Tel: +886-3-4227151 Ext. 4519
Fax: +886-3-4222681
E-mail: horng@db.csie.ncu.edu.tw

Wen-Fu Cho
Applied Research Lab., Telecommunications Labs.
Chunghwa Telecom Co., Ltd.
Yang-Mei, Taoyuan, Taiwan
Tel: +886-3-4244197
Fax: +886-3-4244167

Financial Support: National Science Council of the Republic of China under Contract No. NSC 89-2213-E-008-061.
Keywords: binding sites, data mining, genomes, regulatory elements, transcription factors.

Repeat sequences are the most abundant ones in the extragenic region of genomes. Biologists have already found a large number of regulatory elements in this region. These elements may profoundly impact the chromatin structure formation in nucleus and also contain important clues in genetic evolution and phylogenic study. This study attempts to mine rules on how combinations of individual binding sites are distributed repeat sequences. The association rules mined would facilitate efforts to identify gene classes regulated by similar mechanisms and accurately predict regulatory elements. Herein, the combinations of transcription factor binding sites in the repeat sequences are obtained and, then, data mining techniques are applied to mine the association rules from the combinations of binding sites. In addition, the discovered associations are further pruned to remove those insignificant associations and obtain a set of discovered associations. Finally, the discovered association rules are used to partially classify the repeat sequences in our repeat database. Experiments on several genomes include *C. elegans*, human chromosome 22 and yeast.

An increasing number of genomes sequenced has ushered in the study of sequences. In this area, repetitive sequences have received considerable interest (Moyzis et al. 1989; Williams and Robbins, 1992; Alford and Caskey, 1994; A large amount of the subsequences continuously appears in a sequence. Repetitive sequences are the most abundant ones in extragenic region of genome, in which a large number of regulatory elements are located. These repeats may significantly affect the chromatin structure formation in nucleus and also provide valuable insight into genetic evolution and phylogeny. This study considers the repetitive sequences whose length extends from twenty to several thousands in the genomes. A database is also constructed for repetitive sequences. (http://dbl2b5.csie.ncu.edu.tw).

Many transcription factor binding sites have been collected in databases. TRANSFAC (Heinemeyer et al. 1998; Heinemeyer et al. 1999) is the most complete and well maintained database for transcription factor binding sites. Notably, consensus patterns or nucleotide distribution matrices can be used to describe transcription factor binding sites. While describing binding sites, Brazma et al. (Brazma et al. 1997) stated "The matrix representation is generally considered as the best available means for representing the consensus, however, at present most consensus descriptions are unreliable in the sense that they tend to give many false positives when compared against the genome sequences of even modest length". Therefore, this study describes the binding sites using consensus patterns. Brazma (Brazma et al. 1997) developed a general software tool to find and analyze combinations of transcription factor binding sites that occur often in gene upstream regions in the yeast genome. In addition to analyze the association rules in the combinations, their work focused on upstream and random regions, in which their ratio appears. Their tool can find all the combinations satisfying the given parameters with respect to the given set of upstream regions, its counterset, and the chosen set of sites. However, the tool is only used in yeast genome.

To face a large among of repeat sequences, data mining plays a prominent role in knowledge extraction. Agrawal (Agrawal et al. 1993) introduced the problem of mining

*Corresponding author

association rules over basket data. An example of an association rule is given below. The work stated '50% of transactions that contain beer also contain diapers; 5% of all transactions contain both of these items'. Where 50% is called the confidence of the rule, and 5% is the support of the rule. Data mining is crucial for extracting knowledge in a database. Frequently used data mining approaches, include association rules, statistical, neural network and genetic algorithms.

In statistics, Chi-square test statistics (c^2) is extensively applied for testing independence and correlation. Chi-square is based on comparing observed frequencies with the corresponding expected frequencies. The closer that observed frequencies are to expected frequencies, implies a greater weight in favor of independence. Let f_0 be an observed frequency, and f is an expected frequency, Chi-square is used to test the significance of the deviation from the expected values. The c^2 value is defined as follows:

$$\chi 2 = \sum \frac{\left(f_0 - f\right)^2}{f}$$

Where c^2 value of 0 implies the sites that are statistically independent. If it is higher than a certain threshold value, *e.g.*, 4.12 at the 97% significance level, we reject the independent assumption. We say that it is correlated.

Research of partial classification using association rules introduces two case studies for partial classification (Ali et al. 1997). The two case studies are medical diagnosis and telecommunications. Instead of attempting to predict future values, such research focuses on identifying characteristics of some of the data classes.

This study initially identifies the combinations of transcription factor binding sites in repeat sequences. Data mining techniques are then applied to mine the associations from the combinations of transcription factor binding sites that occur in repeat sequences.

The data mining technique can mine an enormous number of associations. The enormous number of associations makes it extremely difficult for a human user to identify those useful or interesting ones.

Next, the associations are used to remove insignificant ones and find a set of useful associations. In addition, the discovered associations are used to partially classify the repeat sequences in our repeat database. Our experimental genome sequences include *C. elegans*, human chromosome 22, yeast and several bacteria.

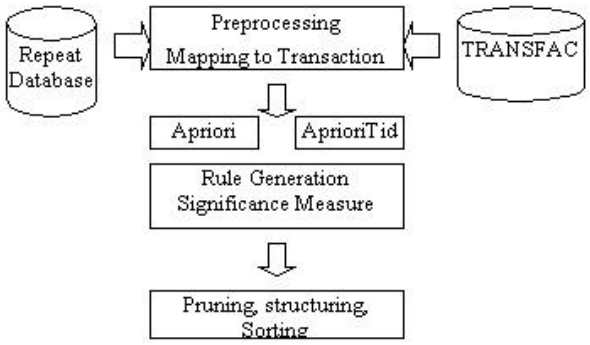

Figure 1. System flow of our approach.

Background

TRANSFAC database (release 4.0) contains 4965 site sequences, and 2837 factor entries. Most sites are also consensus patterns. The data in TRANSFAC has the following features. A transcription factor binding site accession number may have different consensus sequences. Different binding site accession numbers may have a same consensus sequence.

Wild characters such as 'M' or 'W' used in TRANSFAC make the sequences cover other sequences. Small consensus sequences may appear in larger ones. Our approach needs a preprocessing feature because complex characteristics of the transcription factor binding sites are encountered in TRANSFAC.

(a) Properties of repeat sequences in the repeat database

Repeat sequences in the repeat database can be categorized as belonging to the following three types:

- Minisatellite repeats: variable number tandem repeat (VNTR). Each repeat sequence of this type has a length ranging from ten to sixty base pairs. It repeatedly appears from five to fifty times in a sequence.
- Microsatellite repeats: each repeat of this type has a length ranging from one to four base pairs unit repeated 10-20 times.
- Interspersed genome-wide repeats.
    - Short Interspersed Nuclear Elements (SINEs). The length of each repeat is less than 280 base pairs. Repeats repeatedly appeared in genes.
    - Long Interspersed Nuclear Elements (LINEs). The length of each repeat ranges from 6 to 8k base pairs. They repeatedly appear from 50,000 to 100,000 times.

Repeat Database

TID	RID
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C'_1

TID	Set-of-RID
100	{{1},{3},{4}}
200	{{2},{3},{5}}
300	{{1},{2},{3},{5}}
400	{{2},{5}}

L_1

Itemset	Sup
{1}	2
{2}	3
{3}	3
{4}	3

C_2

Itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

C'_2

TID	Set-of-RID
100	{{1 3}}
200	{{2 3},{2 5},{3 5}}
300	{{1 2},{1 3},{1 5},{2 3},{2 5},{3 5}}
400	{{2 5}}

L_2

Itemset	Sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

Itemset
{2 3 5}

C'_3

TID	Set-of-RID
200	{{2 3 5}}
300	{{2 3 5}}

L_3

Itemset	Sup
{2 3 5}	2

Figure 2. Illustrative example of a mapping between a repeat sequence and its combinations of the transcription factor binding sites.

- Inverted repeats: Repeat sequences invert each other. For example, the following two repeat sequences are inverted.

  5’ GATTC---GAATC 3’
  3’ CTAAG---CTTAG 5’

The repeat sequences in our experiments include direct and inverted repeats whose length is equal or larger than twenty base pairs.

(b) Properties of the data in TRANSFAC

Genome sequences are a string of A, C, G or T. The symbols used in addition to A, C, G, or T also include the following:

W: A or T
R: A or G
K: G or T
B: C, G, or T
H: A, C, or T
N: A, C, G, or T
S: C or G
Y: C or T
M: A or C
D: A, G, or T
V: A, C, or G

Characteristics of the data in TRANSFAC are introduced as follows:

Example 1:

MATWAAT R04327

The illustrative example indicates that AATAAAT, CATAAAT, AATTAAT, CATTAAT are all matched to a same site identification.

Example 2:

R00018 TGCCCTAA
R00018 TGCCCTTG
R00018 TGCCTGG
R00018 TGGCAAAC

Example 2 indicates that site R00018 has four different binding site consensus sequences. In TRANSFAC, 71 site IDs belong to this type.

Example 3:

R01372 GGGGC
R01241 GGGGC
R01243 GGGGC

Example 3 indicates different binding sites having the same consensus sequence.

Example 4:

R02248 MAMAG
R08440 AAAG

The binding site R08440 is covered by the other R02248. In TRANSFAC, 3906 binding sites belong to this type. Each site may or may not have transcription factor names. 3006 accession numbers have transcription factor names.

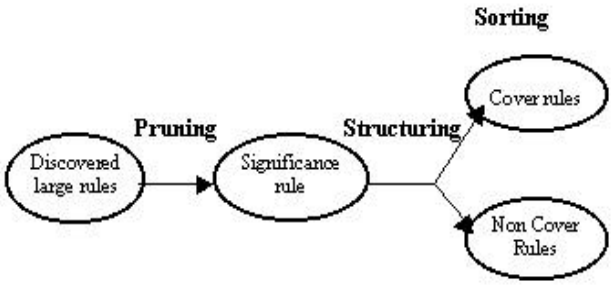

Figure 3. The pruning and structuring techniques.

Example 5 shows another situation. Different binding sites contain the same set of transcription factor names. For example, the binding sites R00303, R00304, R00305, R00306 have the same transcription factor names, *i.e.*, Oct-1C Oct-1B Oct-4 Oct-1A.

Example 5:

R00001 ISGF-3
R00002 ICSBP
R00003 ISGF-3
R00303 Oct-1C Oct-1B Oct-4 Oct-1A
R00304 Oct-4 Oct-1A Oct-1B Oct-1C
R00305 Oct-4 Oct-1A Oct-1B Oct-1C
R00306 Oct-1B Oct-1C Oct-4 Oct-1A

(c) Significance level

The significant measurement with correlated and independent is defined herein as follows (Liu et al. 1999):

Definition 1 (correlated): Where '*s*' is a minimum support, '*t*' is a significant level, A is a set of items and B is an item. Assume that the rule A=>B is correlated if it satisfies the following two conditions:

- The support exceeds '*s*'.
- The significant level exceeds '*t*'.

Definition 2 (independent): Let '*s*' be a minimum support, '*t*' be a significance level, A be a set of items, and B be an item. Assume that the rule A=>B is independent if it satisfies the following two conditions.

- The support exceeds '*s*'.
- The significant level does not exceed '*t*'.

The proposed approach

Figure 1 illustrates the proposed approach. The first component is a preprocessing and a mapping between the transcription factor binding sites in TRANSFAC and the repeat sequences in the Repeat Database. Next, apriori and aprioriTid (Agrawal et al. 1994) are applied to mine the association rules by combining the transcription factor binding sites in repeat sequences.

Then, Chi-square is used to select certain rules. Finally, the redundant rules are pruned and structured.

Summarized steps of the proposed approach:

- Determine the number of item sets of the transcription factor binding sites in TRANSFAC.
- For categorical binding sites, identification of a binding site is mapped to a set of transcription factor names.
- Find the combinations of transcription factors in repeat sequences.
- Apply the data mining approach to generate association rules.
- Determine the interesting rules using Chi-square significance measure.
- Prune redundant rules (Toivonen et al. 1995; Klemettinen et al. 1994).
- Classify rules to cover and non-cover sets.
- Partially classify repeat sequences using association rules mined.

Cover rules:

rule NF-1 => R04365\R04367 conf=0.942 sup=0.356
rule Sp1 => R04365\R04367 conf=0.967 sup=0.281
rule T-Ag => R04365\R04367 conf=0.977 sup=0.437
rule R03047 => R04365\R04367 conf=1 sup=0.417

Non-Cover rules:

rule R04365\R04367 => Sp1 conf=0.309 sup=0.281
rule R04365\R04367 => R01203 conf=0.346 sup=0.314
rule R04365\R04367 => NF-1 conf=0.392 sup=0.356
rule R04365\R04367 => R03047 conf=0.458 sup=0.417
rule R04365\R04367 => T-Ag conf=0.481 sup=0.437
rule R01203 => R04365\R04367 conf=0.929 sup=0.314

Figure 4. The partial classification rules for the human chromosome 22.

Results

(a) Preprocessing and mapping between the data in the Repeat Database and in TRANSFAC

The transcription factor binding sites in TRANSFAC above are first prepared due to the complicated situations described previously. This accounts for why the proposed approach requires preprocessing. Combinations of the transcription factor binding sites in the repeat sequences in our Repeat Database are then found. This work focuses mainly on the repeat sequences of the genomes *C. elegans*, human chromosome 22, yeast and several bacteria. Table 1 summarizes the results of the preprocessing. The abbreviations of the organisms in Table 1 are given in Appendix A.

Each row refers to a genome or bacteria that is experimented with. The column 'Average Factors' represents the average transcription factor binding sites found in a repeat sequence. As mentioned above, we find the combinations of transcription factors in repeat sequences. The 'Average Factors' is defined to be the sum of the transcription factor binding sites for all repetitive sequences over the sum of the repetitive sequences. The last column 'Ratio' denotes the number of repetitive sequences containing more than one binding site over the total repetitive sequences in a genome. For example, the ratio 77.17% in *C. elegans* indicates 77.17% repeat sequences, *i.e.*, 351,084 ones that will be used to mine associations.

Exactly how to mine associations from the combinations of the transcription factor binding sites found above is discussed as follows. Consider a large database with transactions, where each transaction consists of a set of items. An association rule is an expression as A=>B, where A and B are the sets of items. The mining of an association rule is that a transaction in the database that contains A also tends to contain B. For example, 90% of the people who purchase beer also purchase diapers. So, 90% is called the confidence of the rule. The support of the rule A=>B given herein is the percentage of transactions that contain both A and B.

The formal statement of the problem is described below. Let $I=\{i_1, i_2, \ldots, i_m\}$ be a set of sites, called 'item set'. Let D be a set of repeat sequences, where each repeat sequence S corresponding to a transaction contains a set of items such that $S \subseteq I$. Figure 2 presents an example of a mapping between the repeat sequences and the transcription factor binding sites, where TID is a number of a repetitive sequences and RID is a set of IDs of binding sites. In the proposed approach, only consider repetitive sequences that contain more than one binding site.

Example 6 is used to illustrate the mapping between a repeat sequence and the transcription factor binding sites.

Example 6:

```
>IDI0000000013
AGTTATTCAAACACGTATAA
    TTCAAA R02749
    TATAA  R00046 R00705 R00706 R03054
    TATA   R00671 R00689 R00938 R01128
    R01129 R01191 R04293
```

In Example 6, 'AGTTATTCAAACACGTATAA' is a repeat sequence in the Repeat Database. We map it to a transaction whose id is IDI0000000013. The repeat sequence has three consensus patterns, *i.e.*, 'TTCAAA', 'TATAA' and 'TATA'. The consensus pattern 'TTCAAA' has an accession number R02749. However, the other two consensus patterns 'TATAA' and 'TATA' have many accession numbers. The situation must be preprocessed. Example 6 is another case. Similarly, IDI0000000737 is a transaction ID mapped from a repeat sequence 'TTGAAATTTTGAAATTTAAA'. The repeat sequence has four consensus patterns.

Example 7:

```
>IDI0000000737
TTGAAATTTTGAAATTTAAA
TTGAA          R04347 R04360 R04369
  ATTTNNNNATTT  R02171
  TKNNGNAAK    R02216
      TTTAAA R01598
```

Example 7 presents the results after the mapping. Each list shows the factor name, consensus sequences and the identification of the binding site.

Example 8:

```
>IDI0000000737
TTGAAATTTTGAAATTTAAA
    DE unknown=TTTAAA>R01598
    DE unknown=TTGAA>R04347\R04360\R04369
    DE HiNF-A=ATTTNNNNATTT>R02171
    DE C/EBPbeta\C/EBPdelta=TKNNGNAAK>R02216
```

In Example 8, the repeat sequence (transaction) 'TTGAAATTTTGAAATTTAAA' contains four consensus patterns (items), *i.e.*, TTTAAA, TTGAA, ATTTNNNNATTT, and TKNNGNAAK.

Example 8 contains many interesting observations:

- One site and no factor: they resemble R01598.

- One site and one factor: they resemble R02171 with the factor HiNF-A.

- One site with many accession numbers: it is like R04347, R04360, and R04369 with the same consensus sequence TTGAA.

- One site and many factors: they resemble R02216 with factors 'C/EBPbeta' and 'C/EBPdelta'. Different factors or bindings are separated by the symbol '\'.

A transaction and the items it contains are represented in Example 9.

Example 9:

>IDI0000000737 R04347\R04360\R04369 HiNF-A C/EBPdelta\C/EBPbeta R01598

In Example 9, the transaction IDI0000000737 contains four items that are denoted R04347\R04360\R04369, HiNF-A, C/EBPdelta\C/EBPbeta, and R01598, respectively.

Assume that a repeat sequence 'S' contains 'A', a set of items of 'I', if $A \subseteq S$. An association rule is an implicate of the form A=>B, where $A \subset I$, $B \subset I$, and $A \cdot B=0$.

The rule A=>B holds in the repetitive sequence set 'D' with confidence conf if c% of transactions in 'D' contains 'A' and also 'B'. The rule A=>B has support sup in the repetitive sequence set 'D' if s% of repeat sequences in 'D' contained $A \cup B$.

In our experiments, the minimum support is set to 10%. The association rules are generated if the rule has a higher support and confidence than user specified. Apriori and aprioriTid (Agrawal et al. 1994) are then applied to mine association rules.

An enormous number of association rules are generated. The enormous number of association rules makes extremely difficult for human users to identify those interesting and useful ones. Therefore, Chi-square is applied to prune the discovered association rules to remove those insignificant association rules.

(b) Pruning and structuring association results

Herein, rules are generated using Chi-square significance test. The discovered rules are still large and unreadable after applying the process of Chi-square significance test. The redundant rules are pruned and the rules are structured to cover set and non-cover set.

Figure 3 presents the conceptual flow of the pruning and structuring, summarized as follows:

- Discovered rules may be not interesting for several reasons (Klemettinen et al. 1994). Rules can correspond either to prior biology knowledge or expectations.
- Rules can refer to uninteresting sites or sites combinations such as transcription factor binding sites on protein to *C. elegans.*
- Rules can be redundant.
  Three operations are used to process a large collection of rules.
  - Pruning: the operation may reduce the insignificant rules.
  - Structuring: the operation divides the rules into cover and non-cover sets.
  - Sorting: rank the rules by the use of confidence.

Chi-square significance is not hindered by simple redundancy and strict redundancy. For example, the rule AB=>C is redundant to A=>BC. The rule AB=>C is tested, while A=>BC is not. The strict rule A=>B is redundant of A=>BC, and A=>B is tested. The redundancy in our rules is similar to A=>B and AC=>B. The rule A=>B is kept and the rule AC=>B is pruned because AC=>B is covered by the rule A=>B.

For example, consider the rule MAMAG=>AAAG. Obviously, the binding site on the right-hand side is covered by that on the left-hand side because 'M' may be 'A' or 'C'.

Next, the rule is put into the cover set. Tables 2 and 3 present the association rules mined after applying Chi-square from Table 1. In Table 3, the significance level is set to 95%. In Table 2, the 'MiniSup' column refers to the minimum support used.

The 'Cover Rules' and 'Non Cover Rules' denote the number of rules in the cover and non-cover sets, respectively, after they are mined, pruned, and structured. The 'Total Rules' denotes the sum the rules in the cover and non-cover sets. The 'Ratio of Partial Classification' represents the ratio of the repeat sequences and are classified by the

'Total Rules'. For example, 47% repeat sequences of *C. elegans* are partially classified by the ten rules mined. Conversely, the situation also indicates that other 53% repeat sequences can't be classified by the rules. Therefore, the ratio can also be used to measure whether the rules mined are representative. Similarly, Table 3 summarizes the data for archaea, bacteria and virus. The minimum support is set to 10% and those with the '*' symbol in the precedence of the genome name is set to 20%.

Table 1. Combinations of transcription factor binding sites for *C. elegans*, human chromosome 22, yeast, archaea, bacteria, and virus.

Genome Name	Total Repeat Sequences	Match One	No Match	More Than One Match	Average Factors	Ratio
C. elegans	454927	73881	29962	351084	4.8	77.17%
Human chromosome 22	1347364	47159	22211	1277994	7.6	94.85%
Yeast	4329	305	338	3686	22.5	85.14%
Bsub	700	73	27	600	11.5	85.71%
Hinf	788	93	55	640	7.3	81.22%
Hpyl	713	98	25	590	8.3	82.75%
Hpyl99	721	88	33	600	6.3	83.22%
Mgen	373	26	16	331	6.7	88.74%
Mtub	4932	784	171	3977	5.1	80.64%
E. coli	1897	188	60	1649	8.8	86.93%
CP	135	14	8	113	7.3	83.70%
MP	1282	107	36	1139	7.5	88.85%
RP	98	8	2	88	5.8	89.80%
TP	102	7	4	91	15.3	89.22%
AP	398	62	7	329	7.4	82.66%
AR	779	48	21	710	7.8	91.42%
PA	277	20	4	253	5.1	91.34%
PH	401	17	4	380	6.5	94.76%
AA	299	20	7	272	6.9	90.97%
CT	27	4	1	22	14.5	81.48%
S	1580	78	34	1468	9.1	92.91%
TM	518	24	14	480	7.0	92.66%
UU	302	31	9	262	6.2	86.75%

Table 2. The association rules mined after applying Chi-square.

Genome Name	MiniSup	Cover Rules	Non Cover Rules	Total Rules	Ratio of Partial Classification
C.elegans	5%	4	6	10	47%
Human chromosome 22	28%	4	6	10	79%
Yeast	31%	5	5	10	77%

Cover rules:

rule c-Myb => R01514 conf=0.479 sup=0.052

rule NF-1 => R03553 conf=0.504 sup=0.086

rule c-Myb => R04347\R04360\R04369 conf=0.522 sup=0.057

rule TCF-1alpha\TCF-1\TCF-1F\TCF-1G\TCF-1E\TCF-1C\TCF-1B\TCF-1A\TCF-2alpha\LEF-1 => MNB1a conf=0.856 sup=0.106

Non-Cover rules:

rule R04347\R04360\R04369 => R01514 conf=0.327 sup=0.054

rule R04347\R04360\R04369 => c-Myb conf=0.347 sup=0.057

rule R03553 => NF-1 conf=0.463 sup=0.086

rule MNB1a => TCF-1alpha\TCF-1\TCF-1F\TCF-1G\TCF-1E\TCF-1C\TCF-1B\TCF-1A\TCF-2alpha\LEF-1 conf=0.692 sup=0.106

rule R01514 => c-Myb conf=0.701 sup=0.052

rule R01514 => R04347\R04360\R04369 conf=0.722 sup=0.054

Figure 5. The partial classification rules for the *C. elegans* genome.

Figures 4 and 5 present partial classification rules for the human chromosome 22 and *C. elegans* genome, respectively. These rules can be used to find genes in complete genomes and cluster repeat sequences once they are verified. Biologists at National Ynag-Ming University in Taiwan are verifying these results.

Discussion

To verify the association rules found in repetitive sequences also appear in their genomes. We experiment on several archaea and bacteria. This is because their sizes are shorter. The experimental results are shown in Table 4. The column 'Occurrences in Repeats' denotes how many copies of a repetitive sequence are found in a genome. The column 'Occurrences in Genome' represents how many associations are found in a genome. The 'Window' column indicates the offset of the transcription factors binding site, *e.g.*, the difference of the transcription factors binding site. For example, two of the rules YY1=R00231\R00232\R00335\R00668\R00669\R00761\R01081\R01345\R01445\R01446\R02955\R02957 and YY1=> R00388 are found in a repetitive sequence of the organism *Pyrococcus abyssi*. For more details of the two rules, the reader may refer to Appendix B.

The repetitive copies of the repetitive sequence are 39. We then go back to its genome scale and find the association YY1= R00388 also exist 48 different positions when the window is set 5. The result seems to be reasonable. The larger of the window is, the more associations are found. However, a huge amount of associations are found in a genome scale such as *Thermotoga maritima* even the occurrences of the repetitive sequence is not large. We can't conclude from these observations. We will further study the phenomenon in the future.

Concluding Remarks

This study finds combinations of transcription factor binding sites in the repeat sequences of the Repeat Database. Each repeat sequence is mapped to a transaction, and combinations of transcription factor binding sites are mapped to items of a transaction. The transcription factor binding sites in TRANSFAC are first preprocessed due to their complex characteristics. The apriori and aprioriTid (Agrawal et al. 1994) approaches are then applied to mine the associations from the combinations of transcription factor binding sites in repeat sequences.

An enormous number of association rules are generated. The enormous number of association rules makes it extremely difficult for a human user to identify those interesting and useful ones. In addition, Chi-square significance level is used to remove those insignificant rules. Finally, the redundant rules are pruned and then the remaining rules are classified into cover and non-cover sets. Moreover, experiments are conducted on many genomes including *C. elegans,* human chromosome 22, yeast and bacteria. Biologists at National Yang-Ming University in Taiwan have verified and found the rules mined to be interesting. The rules mined can also be used to find useful genes in complete genomes as well as partially cluster the repeat sequences in Repeat Database. Biologists are experimenting and verifying now.

Table 3. The association rules for archaea, bacteria and virus are mined after applying Chi-square.

Genome Name	Prune Rules	Non Cover Rules	Cover Rules	Total Rules
Bsub	63	103	55	158
Hinf	3	3	3	6
Hpyl	0	3	1	4
Hpyl99	18	11	21	32
Mgen	19	17	11	28
Mtub	0	5	1	6
Ecoli	0	1	1	2
CP	0	3	1	4
MP	0	3	5	8
RP	3	10	14	24
*TP	0	8	10	18
AP	31	24	26	50
AR	1004	74	15	89
PA	3	4	2	6
PH	55	8	12	20
AA	0	3	5	8
*CT	0	4	2	6
S	3	22	18	40
TM	55	20	6	26
UU	0	8	8	16

*The minimum support of the genome name is set to 20%.

Appendix A. Abbreviation of organisms.

Helicobacter pylori J99	HPJ9
Helicobacter pylori 26695	HP25
Mycoplasma genitalium	MG
Mycobacterium tuberculosis H37Rv	MT
Escherichia coli	EC
Hepatitis C virus	HCV
Human immunodeficiency virus type 1	HIV1
Japanese encephalitis virus	JEV
Aquifex aeolicus	AA
Aeropyrum pernix K1	AP
Archaeoglobus fulgidus	AR
Chlamydia pneumoniae AR39	CP
Chlamydia trachomatis	CT
Mycoplasma pneumoniae M129	MP
Pyrococcus horikoshii OT3	PH
Rickettsia prowazekii strain Madrid E	RP
Synechocystis PCC6803	S
Thermotoga maritima	TM
Treponema pallidum subsp. pallidum	TP
Ureaplasma urealyticum	UU
Pyrococcus abyssi	PA

Table 4. The association rules in a small scale (repetitive sequences) and genome scale.

Organism	Association Rules	Occurrences in Repeats	Occurrences in Genome		
			Window=1	Window =5	Window =10
Thermotoga maritima	c-Ets-2=>R03553	272	1506	1700	2019
	R03553=>R01230	220	0	56	332
	c-Ets-2=>R01230	218	0	66	206
Mycoplasma genitalium	TCF-1alpha\TCF-1\TCF-1F\TCF-1G\TCF-1E\TCF-1C\TCF-1B\TCF-1ª \TCF-2alpha\LEF-1=>MNB1a	208	3785	3954	4557
Treponema pallidum subsp. Pallidum	Sp1=>R03047	33	549	719	1219
	Sp1=>T-Ag	39	984	1285	1779
	Sp1=>GAL4	39	474	1150	1883
	GAL4=>R04141	39	0	1641	1853
	R01203=>R04398	33	0	602	817
	GAL4=>R03047	39	0	161	416
	R04398=>R00290\R01241\R01244	43	879	894	940
Ureaplasma urealyticum	YY1=>R01513	62	754	2003	2614
	YY1=>Pit-1ª	60	0	893	1859
	N-Oct-3=>Pit-1ª	64	179	2610	3230
	TCF-1alpha\TCF-1\TCF-1F\TCF-1G\TCF-1E\TCF-1C\TCF-1B\TCF-1ª \TCF-2alpha\LEF-1=>MNB1a	72	3202	3295	3650
	Pit-1ª =>R01598	50	0	1305	1621
	Pit-1ª =>YY1	60	0	893	1859
	R01513=>YY1	62	754	2003	2614
Pyrococcus abyssi	YY1=>R00231\R00232\R00335\R00668\R00669\R00761\R01081\R01345\R01445\R01446\R02955\R02957	39	0	34	105
	YY1=>R00388	41	0	48	175
	R00388=>R00231\R00232\R00335\R00668\R00669\R00761\R01081\R01345\R01445\R01446\R02955\R02957	37	0	37	64
Synechocystis PCC6803	NF-1=>R03553	356	6328	9307	12568
	TCF-1alpha\TCF-1\TCF-1F\TCF-1G\TCF-1E\TCF-1C\TCF-1B\TCF-1ª \TCF-2alpha\LEF-1=>MNB1a	449	12871	13209	14597
	NF-1=>R00291	469	696	3506	5305
Rickettsia prowazekii	YY1=>TFIID	16	335	551	975
	N-Oct-3=>ETF	14	445	1334	1728
	YY1=>SEF4	22	872	1017	1275
	YY1=>R01513	24	1024	2265	3051
	Pit-1ª =>N-Oct-3	18	111	2571	2991
	R00671\R00689\R00938\R01128\R01129\R01191\R04293=>TFIID	14	2037	2382	2869
	R00671\R00689\R00938\R01128\R01129\R01191\R04293=>R00583	16	4769	5071	5716
	R00671\R00689\R00938\R01128\R01129\R01191\R04293=>R01513	18	0	2519	3374
	Pit-1ª =>R01598	18	0	869	1035
	ETF=>TFIID	14	2724	2754	2982

Appendix B. The details of the two rules of the organism *Pyrococcus abyssi.*

YY1=>R00231\R00232\R00335\R00668\R00669\R00761\R01081\R01345\R01445\R01446\R02955\R02957

YY1=>R00388

YY1	TATTT CCWTNTTNNNW CATTA CATTT
R00388	TCAAT
R00231\R00232\R00335\R00668\R00669\R00761\R01081\R01345\R01445\R01446\R02955\R02957	ATTGG

Acknowledgments

The authors would like to thank Professors Cheng-Yen Kao at National Taiwan University and Ueng-Cheng Yang and Dr. Yu-Chung Chang at Yang-Ming University for their helpful suggestions.

References

Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining associations between sets of items in large databases. In: ACM SIGMOD Int'l Conference on Management of Data, Washington D.C., 207-216.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In: 20th Int'l Conference on Very Large Databases, Santiago, Chile, Sept. 1994. Expanded version available as IBM Research Report RJ9839, 487-499.

Alford, R.L. and Caskey, C.T. (1994). DNA analysis in forensics, disease and animal/plant identification. Current Opinion in Biotechnology 5:29-33.

Ali, K., Manganaris, S. and Srikant, R. (1997). Partial classification using association rules. Knowledge Discovery and Data Mining. pp. 115-118.

Bennatzen, J.L. (1996). The contributions of retroelements to plant genome organization, function and evolution. Trends in Microbiology 4:347-353.

Brazma, A., Vilo, J., Ukkonen, E. and Valtonen, K. (1997). Data mining for regulatory elements in yeast genome. In: International Conference Intelligent Systems for Molecular Biology, 5th. Halkidiki, Greece, June. pp. 65-74.

Brown, T.A. (1999). Genome anatomies. In: Genomes. BIOS Scientific Publishers Ltd., Oxford, UK. pp. 135-141.

Heinemeyer, T., Chen, X., Karas, H., Kel, A. E. , Kel, O. V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F. and Wingender, E. (1999). Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. Nucleic Acids Research 27:318-322.

Heinemeyer, T., Wingender, Reuter, E., Hermjakob, I. H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A., Podkolodny, N. L. and Kolchanov, N. A. (1998). Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. Nucleic Acids Research 26:362-367.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. and Verkamo, A. I. (1994). Finding interesting rules from large sets of discovered association rules. In: Conference on Information and Knowledge Management. Gaithersburg, Maryland, November. pp. 401-407.

Liu, B., Hsu, W. and Ma, Y. (1999). Pruning and summarizing the discovered associations, In: ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) International Conference on Knowledge Discovery and Data Mining. San Diego, CA, USA, May. pp.125-134.

Mitas, M. (1997). Trinucleotide repeats associated with human disease. Nucleic Acids Research 25:2245-2253.

Moyzis, R.K., Torney, E.C., Meyne, J., Buckingham, J.M., Wu J.R., Burks C., Sirotkin K.M. and Goad, W.B. (1989). The distribution of interspersed repetitive DNA sequences in the human genome. Genomics 4:273-289.

Primrose, S.B. (1998). The organization and structure of genomes. In: Principles of genome analysis. Blackwell Science Ltd., Massachusetts, USA. pp.17-44.

Toivonen, H., Klemettinen, M., Ronkainen, P., Hatonen, K. and Mannila, H. (1995). Pruning and grouping discovered association rules. In: MLnet Workshop on Statistics, Machine Learning, and Discovery in Databases. Heraklion, Crete, Greece, September. pp.47-52.

Warren, S.T. (1996). The expanding world of trinucleotide repeats. Science 271:1374-1375.

Williams, S.M. and Robbins, L.G. (1992). Molecular genetic analysis of drosophila rDNA arrays. Trends in Genetics 8:335-340.