Mining the Association Rules of Transcription Factor Binding Sites in Human Tandem Repeats Using Aprior Algorithm

Zhong-yu Liu, Yi-Yang*

School of Life Sciences, Sichuan University, Chengdu 610064, China

*Corresponding author: Yi-Yang, School of Life Sciences, Sichuan University, Chengdu 610064, China, Tel: 85418768; E-mail: zhongyujohn@gmail.com

Received May 05, 2009; Accepted June 12, 2009; Published June 13, 2009


Abstract

Tandem repeats (TR) 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 phylogenetic study. This study attempts to mine rules on how combinations of individual binding sites are distributed tandem repeats in human genome (http://www.trbase2.cn). 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 tandem repeats 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.

Keywords: Human tandem repeats; TRANSFAC database; Transcription factor binding sites; Data mining; Association rules

Introduction

Repetitive DNA sequences have been identified in large quantities in both eukaryotic and prokaryotic genomes (Buard, et al., 1994; Van Belkum et al., 1998). There are various databases which shows TR characteristics, for instance, ABCC GRID database (Collins et al., 2003), Minisatellite database (le Flèche et al., 2001), PlantSat database (Macas et al., 2002), Representative Sequences DataBase (RSDB) (Horng et al., 2002), the Microsatellite Analysis Server (MICAS) for prokaryotic genomes (Sreenu et al., 2003) and the Tandem Repeats Finder (TRF) (Benson, 1999). Boby et al., (2004) built a perfect and imperfect TR database, TRbase, relating TR to disease genes for the human genome (Boby, 2004). 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., (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 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 counter set, 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 and Srikanth, (1993) introduced the problem of mining association rules over basket data. An example of an association rule is given below. The work stated ‘50% of transactions that contain beer also contains 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 our study, to fill in the gap of the original TRbase, our first step was to extend the database. According to Cawley et al., (2004) conclusion (TF binding sites are distributed randomly in human genomes), we utilized Agrawal and Srikanth, (1993) algorithm to explore the combinations of TF binding sites in TR sequences in human genomes rather than in a specific region. We then identified all combinations rules, which were pruned by the chi-square test ($\chi^2$) and subjected to testing for independence and correlation (Mukherjee et al., 2004). The pruned combination rules of TF binding sites will be far-reaching for biologists researching gene expressions and regulatory elements.

Methods

Framework

Fig. 1 illustrates the framework of our method. The project started with extension of the original TRbase, followed by statistical analysis and data mining including generation of association rules.

The steps of the proposed approach are summarized as follows:

- Extension of the original TRbase to include all human chromosomes.
- Determination of the number of item sets of the TF binding sites in TRANSFAC.

In our study, to fill in the gap of the original TRbase, our first step was to extend the database. According to Cawley et al., (2004) conclusion (TF binding sites are distributed randomly in human genomes), we utilized Agrawal and Srikanth, (1993) algorithm to explore the combinations of TF binding sites in TR sequences in human genomes rather than in a specific region. We then identified all combinations rules, which were pruned by the chi-square test ($\chi^2$) and subjected to testing for independence and correlation (Mukherjee et al., 2004). The pruned combination rules of TF binding sites will be far-reaching for biologists researching gene expressions and regulatory elements.

**Figure 1:** System flow of our approach.

1: Extension of the original TRbase to the whole human genome. 2: Each TR in the TRbase was mapped to a transaction; all binding sites in TRANSFAC were mapped to itemsets; and Apriori and AprioriTid algorithms were then utilized to explore the TF binding site combinations in TR sequences. 3: Pruning and summarizing of the discovered associations.
Identification of TF binding sites.
Finding of the TF binding site combinations in TR sequences in the TRbase.
Application of the data mining approach to generate association rules.
Determination of interesting rules using a chi-square significance measure.
Pruning of redundant rules (Toivonen et al., 1995; Klemettinen et al., 1994).
Classification of the rules into direction setting (DS) rules and non-DS rules. Fig. 2 shows the conceptual flow of the technique of pruning and summarization (Liu et al., 1999).

TRbase Extension

DNA sequences and annotations in the original TRbase were retrieved for the completed chromosomes 4, 5, 6, 14, 16, 18, 19, 20, 21 and 22 (Boby, 2004); however this project required data on all those disease genes and their relevant information in the whole-human genome, indicating that the original TRbase needed to be extended to all human chromosomes prior to data preparation. DNA sequences and annotations of the remaining chromosomes were downloaded from GenBank. All TR were detected using the TRF program (version 3.01) (Benson, 1999) with parameters as in Boby, (2004) applied to DNA sequences extracted from GenBank in the FASTA format using the Seqret program of EMBOSS (Rice et al., 2000).

Data Mining

Association rule mining for TF binding sites in tandem repeats Association rule mining and Agrawal and Srikanth, (1993) algorithm

Association rule mining is important for extracting knowledge from many repetitive sequences. Agrawal and Srikanth, (1993) developed the Apriori algorithm for mining association rules. The Apriori algorithm (Agrawal and Srikanth, 1994) accepts as inputs two thresholds, \textit{min-supp} and \textit{min-conf}, and mines (finds) all association rules having \textit{support} and \textit{confidence} greater than or equal to those thresholds. The Apriori algorithm mines association rules using a two-stage process. The first stage generates all the sets of items that satisfy the \textit{min-supp} constraint, called frequent itemsets. The second stage constructs all the association rules that satisfy the \textit{min-conf} constraint from those frequent itemsets. Details of this algorithm are contained in Agrawal and Srikanth, (1994); Liu et al., (1999).

TRbase and TRANSFAC

Tandem repeats sequences in TRbase

The extended TRbase, now covering the whole human genome, consists of perfect and imperfect TR For more details of the features of the data in the TRbase; refer to Boby (2004).

TRANSFAC database (release 7.0) contains 7915 site sequences, and 6133 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.

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

Characteristics of the data in TRANSFAC are introduced as follows:

Example 1:

WTATYCAT R02160

This example indicates that ATATCCAT, ATATTCAT, TTATCCAT, TTATTCAT are all matched to the same site.
Mapping of tandem repeats and transcription factor binding sites

Let \( I = \{i_1, i_2, \ldots, i_m\} \) be a set of TF binding sites retrieved from TRANSFAC (http://www.gene-regulation.com/pub/databases.html), called the item set. Let \( D \) be a set of TR, where each TR sequence, \( d \), corresponding to a transaction, includes a set of items. Example 2 is used to illustrate the mapping between TR and the TF binding sites.

Example 2:

\[
\begin{align*}
&> ID17f \\
&AAAAAAGAAAAAAGAAAAGG \\
&AAAAAAG R02248 \\
&GG R04365, R04367, R04690
\end{align*}
\]

In this example 2, ‘AAAAAAGAAAAAAGAAAAGG’ is a TR sequence in the TRbase Database. We mapped it to a transaction whose ID is ID17. The repeat sequence has three consensus patterns, i.e., ‘AAAAAG’, ‘GG’. The consensus pattern ‘AAAAAG’ has an accession number R02248. However, the other consensus pattern ‘GG’ has three accession numbers: R04365, R04367 and R04690.

In our experiments, the minimum support is set to 10%. Association rules are generated only if they have higher support, i.e., \( \geq 10\% \) and confidence, i.e., \( \geq 90\% \). Apriori and AprioriTid algorithms (Agrawal and Srikanth, 1994) are then applied to mine association rules.

Pruning and summary of association results

Pruning the discovered associations

It is well known that many discovered associations are redundant or minor variations of others. Their existence may simply be due to chance rather than true correlation. Thus, those spurious and insignificant rules should be removed. This is similar to pruning of overfitting rules in classification (Klemettinen et al., 1994) Rules that are very specific (with many conditions) tend to overfit the data and have little predictive power. Although association rules are not normally used for prediction, rules that only capture the irregularities and idiosyncrasies of the data have no value and should be removed. An example of such a rule is shown below. Example: We have the following two rules:

R1: Job = yes \( \rightarrow \) Loan = approved [sup = 60\%, conf = 90\%]
R2: Job= yes, Credit_history=good \( \rightarrow \) Loan= approved [sup = 40\%, conf = 91\%]

If we know R1, then R2 is insignificant because it gives little extra information. Its slightly higher confidence is more likely due to chance than to true correlation. It thus should be pruned. R1 is more general and simple. General and simple rules are preferred. In this work, we measure the significance of a rule using chi-square test (\( \chi^2 \)) for correlation from statistics (Mills, 1955).

Summarizing the unpruned rules

Pruning can reduce the number of rules substantially. However, the number of rules left can still be very large. This step finds a subset of the rules, called direction setting rules (or DS rules), to summarize the unpruned rules. Essentially, DS rules are significant association rules that set the directions for non-DS rules to follow. The direction of a rule is the type of correlation it has, i.e., positive correlation or negative correlation or independence, which is also computed using \( \chi^2 \) test. Let us see an example. Here, it is presented as a post-processing method. In implementation, it is combined with rule mining. Example 2: We have the following discovered rules:

R1: Job = yes \( \rightarrow \) Loan = approved [sup = 40\%, conf = 70\%]
R2: Own_house = yes \( \rightarrow \) Loan = approved [sup = 30\%, conf = 75\%] \( \chi^2 \) analysis shows that having a job is positively correlated to the grant of a loan, and owning a house is also positively correlated to obtaining a loan. Then, the following association is not so surprising to us:

R3: Job = yes, Own_house = yes \( \rightarrow \) Loan = approved [sup = 20\%, conf = 90\%] because it intuitively follows R1 and R2. We can use R1 and R2 to provide a summary of the three rules. R1 and R2 are DS rules as they set the direction (positive correlation) that is followed by R3. In real-life data sets, a large number of associations are like R3. From the example, we see that the DS rules give the essential relationships of the domain. The non-DS rule is not surprising if we already know the DS rules. However, this, by no means, says that non-DS rules are not interesting. Non-DS rules can provide further details about the domain. For example, the non-DS rule above (R3) gives a higher confidence, which may be of interest to the user. Using DS rules to form a summary is analogous to summarization of a text article. From the summary, we know the essence of the article. If we are interested in the details of a particular aspect, the summary can point us to them in the article. In the same way, the DS rules give the essence of the domain and points the user to those related non-DS rules. Non-DS rules are basically combinations of DS rules. In the above examples, R3 is a combination of R1 and R2.

It is well known that many discovered associations are redundant or minor variations of others. Their existence may simply be due to chance rather than true correlation. Thus, spurious and insignificant rules should be removed. In the
present study, we measure the significance of rules using the chi-square test ($\chi^2$) for correlation from statistics (Mills, 1955). The chi-square statistical test ($\chi^2$) is frequently adopted to test independence and correlation.

**Results and Discussions**

**Extended TRbase**

The original TRbase can be accessed via the website http://trbase.ex.ac.uk (Boby, 2004). Similarly, the new extended TRbase can be accessed via http://www.trbase2.cn. Compared with the original TRbase, our extended TRbase is unabridged.

**Data Mining**

**Association rule mining**

We retrieved all human TF binding site consensus sequences (3,447) from the TRANSFAC database (Ali et al., 1997; Bulyk, 2003), and released those consensus sequences, which we found that they consisted of 455,979 site sequences. We then mapped the released sequences to itemsets in association rule mining. The 649,400 TR whose con_size was 10 or greater were selected as transaction datasets—as above described, the minimum confidence and support were set to 90% and 10%, respectively. The 44 combinations of TF binding sites (supplementary file 2) were worked out. Liu et al., (1999) describe the pruning and summarizing approaches and definitions. The 44 combinations of TF binding sites (supplementary file 1) were, respectively, classified into 3 groups: positive correlation rules, negative correlation rules and independent rules. In our experiment, only the positive correlation rules were considered significant, and were further classified into DS rules and non-DS rules. The rules can be used to find genes in complete genomes and cluster TR once they are verified. Some classification rules for the human genome are as follows.

R0046->R00938 independent  
R00046->R00707 positive-correlation, DS  
R0046->R00705, R00707 positive-correlation, DS  
R0046->R00705, R00938 positive-correlation, DS  
R0046->R00707, R00938 positive-correlation, DS  
R00707->R00938, R004293 positive-correlation, DS  
R0046, R00938->R00705 positive-correlation, Non-DS  
R0046, R00938->R00707 positive-correlation, Non-DS  
R0046, R00938->R00705, R00707 positive-correlation, Non-DS  
R0046, R004293->R00705 positive-correlation, Non-DS  

To explore human TR, we employed both statistical and standard data mining methods. The rule, R00046->R00707 positive-correlation, DS, means that the consensus sequences of more than 10% TRs in human genome contains the consensus sequences corresponding to TFBS RO0046, RO0707, furthermore, 90% of the cases, they are at the same time in the same TR. In most cases, each tandem repeat in TRbase2 corresponds to a disorder, the corresponding disorders are linked at certain sense by means of the combination of RO0046 and R00707, thus, those combinations will provide a bridge to research human disease correlations. Future studies should investigate other methods to detect those combinations mined.

**Conclusion**

The extended TRbase provides a platform to study the associations between disease genes and previously uncharacterized TR in the whole human genome. We successfully updated the original TRbase to include the whole human genome (http://www.trbase2.cn). Moreover, the extended TRbase now holds more comprehensive knowledge concerning human TR by inclusion of statistical analysis. The extended TRbase provides users not only with query tools but also useful statistical data. The authors hope that the statistical results will help biologists discover a whole new area of biology, including predicting transposons. This study identified TF binding site combinations in TR sequences of the TRbase. The TF binding sites in TRANSFAC were first preprocessed due to their complex characteristics. The Apriori and AprioriTid algorithms (Agrawal and Srikant, 1994) were then applied to mine the associations from the TF binding site combinations in repeat sequences. Some association rules were generated. Chi-square tests were used to remove the insignificant rules. Finally, redundant rules were pruned and the remaining rules were classified into DS and non-DS sets. The discovered rules can also be used to find useful genes in complete genomes as well as partially cluster the TR in the extended TRbase.

**References**

   » Google Scholar
   » Google Scholar


