Data Mining for Regulatory Elements in Yeast Genome

Alvis Brāzma *
abra@cclu.lv
Institute of Mathematics and Computer Science
University of Latvia
29 Rainis Bulevard
LV-1459 Riga, Latvia

Jaak Vilo
Jaak.Vilo@cs.Helsinki.FI
Department of Computer Science
P.O.Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki
Finland

Esko Ukkonen
Esko.Ukkonen@cs.Helsinki.FI
Department of Computer Science
P.O.Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki
Finland

Kimmo Valtonen
Kimmo.Valtonen@cs.Helsinki.FI
Department of Computer Science
P.O.Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki
Finland

Abstract

We have examined methods and developed a general software tool for finding and analyzing *combinations* of transcription factor binding sites that occur relatively often in gene upstream regions (putative promoter regions) in the yeast genome. Such frequently occurring combinations may be essential parts of possible promoter classes. The regions upstream to all genes were first isolated from the yeast genome database MIPS using the information in the annotation files of the database. The ones that do not overlap with coding regions were chosen for further studies. Next, all occurrences of the yeast transcription factor binding sites, as given in the IMD database, were located in the genome and in the selected regions in particular. Finally, by using a general purpose data mining software in combination with our own software, which parametrizes the search, we can find the combinations of binding sites that occur in the upstream regions more frequently than would be expected on the basis of the frequency of individual sites. The procedure also finds so-called association rules present in such combinations. The developed tool is available for use through the WWW.

Keywords: data mining, promoters, yeast, transcription factors, complete genome, MIPS, IMD, TRANSFAC.

Introduction

The first complete genomes have recently been sequenced and published, including the first eukaryotic genome of yeast *Saccharomyces Cerevisiae* (Goffeau *et al.* 1996) with length of more than 12 million base-pairs (Mb). This gives enormous amount of information for the studies of how the genome as the whole is organized and how it functions. However, extracting knowledge from this information may be even more challenging task than the genome sequencing. The data mining and machine learning techniques will probably play an essential role in this knowledge extraction by finding interesting, statistically unexpected patterns and thus, generating hypotheses for further investigation by biologists.

The genes in an eukaryotic genome have each a particular combination of binding *sites* for sequence-specific transcription factors that activate or repress their transcription (for survey see for instance (Goodbourn & King 1996; Mellor 1993; Mitchell & Tijan 1989)). Usually these sites are specific DNA sequences of length from about 5 to 25 nucleic acids, and they are arrayed within several hundreds base pairs predominantly upstream from the transcription initiation site in the *promoter* region, though some elements can exert control over much greater distance. We will call the genome regions that control the gene transcription the *transcription regulation units* (TRUs). We will be particularly interested in TRUs for protein coding genes, which are transcribed by polymerase II.

The spacers between the sites in TRUs may be much longer than the sites themselves, may have highly variable length, and usually they have no obvious sequence similarity. It is believed that only few genes in the organism are regulated by individual pathways and that the number of TRUs with a very similar organization of sites is small - probably between 10 to 50. Therefore, it seems that the detection of similarity between dif-

* The results were obtained while the author was working at the Department of Computer Science, University of Helsinki. Currently at European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK

ferent TRUs based solely on the traditional alignment methods may be difficult.

Many transcription factor binding sites have been collected in databases (Chen, Hertz, & Stormo 1995; Ghosh 1990; Wingender *et al.* 1996). Individual binding sites can be generalized and described by consensus patterns or so-called position weight or nucleotide distribution matrices. 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.

It is yet an open question how reliable in principle the prediction of individual binding sites can be made, since in reality the transcription factors usually operate in combinations and even perfect binding sites might have no effect if they are isolated. The efforts to look for rules of how combinations of individual binding sites are distributed in a genome have been very rudimentary so far (e.g., (Kel *et al.* 1995b; Prestridge 1995; Quandt, Grote, & Werner 1996)). Understanding of such combinations and their association rules would help in identifying gene classes regulated by similar mechanisms, as well as in prediction of regulatory elements. Our work is aimed towards this goal.

We have developed a tool for the analysis of the upstream regions (of different length) of putative genes of the complete yeast genome taken from MIPS database (Goffeau *et al.* 1996) for all occurring combinations of transcription factor binding sites as given in Information Matrix Database (Chen, Hertz, & Stormo 1995). We are interested in finding combinations that occur relatively frequently in upstream regions and whose occurrence pattern in the genome is different from what might be statistically expected by chance. Such combinations may be parts of promoter classes. A combination of binding sites is characterized by the following parameters:

1. the number of its occurrences in upstream regions;

2. the ratio of the number of its occurrences in upstream regions vs. the number of occurrences in random regions (of the same length and number); and

3. the ratio of the number of its occurrences vs. the expected number of its occurrences based on the individual sites.

The combinations with high values for all these parameters can possibly define promoter classes. The sufficient value of parameter (1) ensures that the combination is present in at least a given number of upstream regions, parameter (2) ensures that the rate of the occurrences of the combination in upstream regions are not just a consequence of high rate of their occurrences in the genome as the whole, and parameter (3) that the rate of the occurrences of the combination is not only a consequence of the high rate of individual occurrencies of the participating binding sites.

For finding these combinations we use general purpose data mining tools in combination with our own software. We also analyze the association rules present in these combinations. As a side-effect we have demonstrated the applicability of a general purpose data mining software to attacking problems in molecular bioinformatics.

The paper is organized as follows. In the next section we give some essential information regarding the available data on which we are relying. In section 3 we define the basic definitions and notions that we use. In section 4 we describe our methods and the information gathering and preprocessing phase, and in section 5 we describe the data mining tool TFCD (Transcription Factor Combination Discoverer) that we have developed and some sample data mining results in the yeast genome. Finally, we will discuss the results and the possible future research directions.

Background

The information about the 16 chromosomes of *S.Cerevisiae* are publicly available in MIPS database (Goffeau *et al.* 1996). MIPS database essentially contains two types of objects: the chromosome sequences, and their annotations providing the information, for instance, about the positions of the predicted genes. Totally 6275 open reading frames (ORFs) have been annotated. The annotations usually give "confidence levels" describing the confidence with which the given ORF has been predicted - from *known protein* to *questionable ORF*. The last category contains 390 ORFs, leaving 5885 ORFs as likely candidates for protein genes.

It is widely assumed that in yeast TRUs rarely extend more than 1500 bp upstream from the coding region of a gene and is often contained within 500 bp (Mellor 1993). By creating an appropriate software for parsing the annotations it is possible to extract all the sequences of a specified length upstream to all putative ORFs of a given level of confidence. Unfortunately this is not a completely trivial problem as there is no published formal grammar describing the MIPS annotations.

So far most of the bioinformatics research regarding TRU regions has been aimed towards describing individual factor binding sites (e.g., (Berg & von Hippel 1988; Bucher 1990; Cardon & Stormo 1992;

Penotti 1990; Schneider, Stormo, & Gold 1986)). For instance, algorithms for constructing consensus description of binding sites from a set of sequences known to contain the site are given in (Frech, Herrmann, & Werner 1993; Quandt *et al.* 1995; Wolfertstetter *et al.* 1996). Several algorithms have been presented for searching the given consensus sites in sequences (Frech, Herrmann, & Werner 1993; Prestridge 1991; Quandt *et al.* 1995), but an acknowledged problem in using these search algorithms is the high rate of false positives. As noted in (Chen, Hertz, & Stormo 1995) - some of the site sequences in the databases may be longer than the actual binding sites, while some others may be shorter.

A list of known yeast proteins, which include the transcription factors, is given in MIPS database. A compilation of *S.Cerevisiae* transcription factors has recently been published also in (Svetlov & Cooper 1995). However, an easier electronic access to yeast transcription factor binding sites can be obtained through specialized transcription factor databases. Two transcription factor databases: TFD (Ghosh 1990) and TRANSFAC (Wingender 1994; Wingender *et al.* 1996), and the transcription site information matrix database IMD (Chen, Hertz, & Stormo 1995) are widely known. TFD database (release 7.5) contains 45 transcription factors and 179 site sequences of *S.Cerevisiae.* TRANSFAC database (release 3.0) contains 116 factor entries and 308 site entries with 279 site sequences for this organism. Most of the site sequences are obtained by *in vitro* experiments, though there are also consensus patterns included.

The Information Matrix Database IMD (Chen, Hertz, & Stormo 1995) is a database containing the position weight matrices constructed from TFD and TRANSFAC as well as from the original references in the following way. The union of all factors from TFD and TRANSFAC that have at least two binding site sequences has been taken, and a position weight matrix has been constructed for each such factor from the given binding sites by maximization of the information content (Hertz & Stormo 1994). Totally there are 39 matrices for yeast factors. A program MATRIX SEARCH 1.0 for searching IMD matrices in a given sequence is also provided with the database.

The research aimed towards the analysis of the "second order" features of TRUs has been started only very recently. Several databases containing information about TRUs basically collected from the literature have been published: EPD (Bucher 1996), COMPEL (Kel *et al.* 1995b), and TRRD (Kel *et al.* 1995a), but none of these databases contain a class of entries specifically for yeast. Analysis of the distance correlation between ORF start positions and sites in several yeast chromosomes by GenomeInspector (Quandt, Grote, & Werner 1996) has been reported in (Werner 1996). Attempts to predict the promoter regions are reported in (Cai & Chen 1995; Pedersen *et al.* 1996; Prestridge 1995). However, no systematic analysis of frequent combinations of transcription sites and their distributions in the genome sequences has been reported. In this paper we are trying to close this gap by using data mining techniques.

Data mining is the most nontrivial step of an automated knowledge discovery process in databases. The task of data mining (see (Fayyad, Piatetsky-Shapiro, & Smyth 1996)) is to extract potentially interesting, statistically unexpected phenomena from the data, and in this way to generate hypotheses for exploration by domain experts. One of the approaches is finding so-called *association rules* (Imielinski & Mannila 1996; Toivonen 1996) (a typical example of an association rule is that in supermarkets almost everybody who buys beer and mustard, buys also sausages). The tools for automatic discovery of interesting facts in databases have been rapidly developing lately, and one of the goals of this paper is to study the applicability of some data mining tools to knowledge discovery in biodatabases.

Definitions

Given a genomic *sequence* (i.e., a string over the alphabet $\Sigma = \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$), and a set of *ORF positions* (i.e., pairs of integers) and *strands* (i.e., W or C), we can select the set of all ORF *upstream sequences* of some given length l. From this set we can select the subset of upstream sequences that do not overlap with genes. We will call them *strictly upstream sequences.* For more precise definitions of these notions see Appendix A.

Given such a set of upstream sequences we can use it for machine learning or data mining algorithms searching for interesting rules of some given type. We are interested in rules that are related to occurrences of yeast transcription factor binding sites. Although, we will mainly use *position weight matrices* (for definition see Appendix B) for describing binding sites, our method does not depend on the particular representation. What matters is only that given a genomic sequence α and a binding site σ we can tell whether the site σ is *present* (or *matches*) at a given position of α or not. When the sequence α is short, in some cases it is sufficient to know simply whether the given site σ is present at any position of α or not - in this case we will say that σ is *present* or *matches* α, or that the sequence α *contains* the site σ.

If we are given a set of n sequences $A = \{\alpha_1, \ldots, \alpha_n\}$ and a site σ, we define the *coverage* $c(\sigma, A)$ of σ in A as the number of sequences in A containing σ. We defined the *support* $s(\sigma, A)$ as the ratio of the coverage of σ to the total number of sequences in A, i.e.,

$$s(\sigma, A) = \frac{c(\sigma, A)}{|A|}$$

If we are given two sets of sequences, a "good" set A and a "bad" set B, and a site σ, then we can define the *goodness ratio* $g(\sigma, A, B)$ of σ in A vs. B as follows

$$g(\sigma, A, B) = \frac{s(\sigma, A)}{s(\sigma, B)}$$

In our case we will use the set of all strictly upstream sequences of a fixed length as the "good" set – we denote it by U – and a set of the same number of sequences of the same length taken from random positions as the "bad" set. We will call the "bad" set the *counterset* and denote it by R.

If we are given a combination (i.e., a set) of k different sites $C = \{\sigma_1, \ldots, \sigma_k\}$ and a sequence α, then we will say that α contains the combination C if it contains all the sites of C. Given a combination C and a set of "good" sequences A and (possibly) a set of "bad" sequences B, we can define *coverage* $c(C, A)$, *support* $s(C, A)$ and *goodness ratio* $g(C, A, B)$ of this combination in the same way as for an individual site. In our practical applications the fact that $g(C, U, R) > 1$ will mean that the combination C occur more frequently in upstream regions than in random regions. If such a combination is discovered, then it is possible (but is not guaranteed), that there is an evolutionary pressure to conserve this combination in upstream regions.

Note that given a site and a set of sequences A, the support $s(\sigma, A)$ of the site can be regarded as the probability that an arbitrarily chosen sequence from A will contain this site. Similarly, given a combination $C = \{\sigma_1, \ldots, \sigma_l\}$ of sites, $s(C, A)$ is the probability that an arbitrarily chosen sequence will contain C. We can use this fact for testing the statistical independence of occurrences of the sites in a set A, since if the sites are independent then we should expect that for sufficiently large $|A|$, the support $s(C, A)$ approximately equals $s(\sigma_1, A) \cdot \ldots \cdot s(\sigma_l, A)$. Let us define the *expected support* $e(C, A)$ of a combination C as

$$e(\{\sigma_1, \ldots, \sigma_l\}, A) = s(\sigma_1, A) \cdot \ldots \cdot s(\sigma_l, A)$$

and the *unexpectedness ratio* as

$$u(C, A) = \frac{s(C, A)}{e(C, A)}$$

In our applications $u(C, U) > 1$ means that the combination C occurs more frequently than expected if the sites were statistically independent.

Thus $g(C, U, R)$ and $u(C, U)$ are two different means of evaluating the deviation of the occurrences of the site combination C in the genome from what should be expected if these occurrences were random. We will be interested in finding combinations with high support, goodness and unexpectedness.

By an *association rule* we understand an implication of the type

$$C_1 \Rightarrow C_2[conf, cov] \tag{1}$$

where C_1 and C_2 are combinations such that $C_1 \cap C_2 = \emptyset$, $0 < conf \leq 1$ is a real number called *confidence*; and *cov* is a positive natural number called the *coverage* as defined earlier. The presence of rule (1) in a set A means that the coverage of the combination $C_1 \cup C_2$ is *cov*, and in every sequence of A where C_1 is present, with the "probability" *conf* the combination C_2 is also present.

Information Gathering and Preprocessing Phase

Our aim was to develop a tool for finding the transcription factor binding site combinations that have high support, goodness and unexpectedness ratios in the yeast genome.

We retrieved the sequences and annotations from MIPS (Goffeau *et al.* 1996) database (the latest update was done on January 29, 1997). From inspecting the annotations and by a trial and error method we found a regular grammar by which in the majority of cases the annotation files can be parsed (we do not describe the grammar here as in itself it does not present a scientific novelty). After having the grammar we created a simple software for retrieving all ORF positions (and strands) that are annotated in the classification field as at least "similar to unknown protein" (thus excluding the classes corresponding to "no similarity" and "questionable ORF") and the positions of all strictly upstream sequences (i.e., these upstream sequences that do not overlap with other ORFs, see Appendix A) of the lengths 50, 100, 150, ..., 1000. For instance, for length 600, there are 2391 strictly upstream sequences. In fact, we do not store the sequences themselves, but only the positions of these sequences in the chromosomes. For each of these sets of upstream sequences we also constructed a random counterset.

Next, we retrieved all the binding sites associated with the organism yeast in TFD (release 7.5) and TRANSFAC (release 3.0) databases, and matched these sites (together with their reverse complements) against the yeast genome. We found, for instance, that in TRANSFAC 3.0 from the given 279 yeast transcription factor binding sites, 130 sites have exactly 1 match, 56 sites have from 2 to 100 matches, 23 sites from 101

IMD sites in yeast genome									
site	matches in total genome	support $l = 300$	goodness ratio $l = 300$	goodness ratio $l = 600$	site	matches in total genome	support $l = 300$	goodness ratio $l = 300$	goodness ratio $l = 600$
ADR1	235	0.0061	0.75	0.62	MCM1	432	0.0157	1.64	1.7
AP-1	52	0.0014	0.71	2.75	MIG1	1676	0.0527	1.27	1.82
ARGRII	6	0.0009	∞	∞	MSN4	4209	0.071	0.72	0.81
BAF1	2125	0.1169	2.16	1.65	NBF	1936	0.0547	1.12	1.2
BAS1	1	-	-	-	PHO2	34752	0.5477	0.96	0.97
BUF	882	0.0134	0.55	0.81	PHO4	919	0.0247	0.9	1.13
CBF1	347	0.0207	2.29	2.54	PUT3	3813	0.0748	0.86	0.91
CCBF	1371	0.0477	1.21	1.16	RAP1	4850	0.1146	1.13	1.17
CUP2	2845	0.0809	1.25	1.19	RC2	45985	0.6271	0.92	0.96
CYP1	8910	0.2103	1.02	1.05	REB1	4891	0.1705	1.62	1.39
DAL82	2	0.0003	∞	0.67	SKO1	908	0.0268	1.46	1.53
DBFA	424	0.0119	0.91	1.23	SRF	300	0.0093	1.1	1.36
GAL4	50	0.0009	0.6	1.33	STE12	10246	0.1914	0.85	0.9
GAL80	7	0.0006	∞	∞	SWI5	56	0.0012	4	0.75
GCN4	15971	0.3229	0.99	0.98	TAF	728	0.0477	2.83	2.01
HSTF	1437	0.03	1.01	0.83	TFIID	22827	0.6501	1.69	1.32
MAL63	4	-	-	3.98	URSF	59	0.0026	2.26	1.86
MATa1	648	0.0145	0.86	0.91	galR	0	-	-	-
MATa2	7570	0.1911	1.08	1.05	CAR1	1335	0.0332	0.91	1.34
MCBF	692	0.0337	1.81	2.05					

Table 1: Table characterizing matches of IMD sites in yeast genome. Factor names are given as in IMD database, except CCBF standing for CCBF/SW4+SW6, CYP1 standing for CYP1.HAP1, MCM1 standing for MCM1.PRTF, RAP1 standing for RAP1/SBF-E/TUF, and CAR1 standing for CAR1represor in IMD denotations. The ∞ symbol is used in cases when there were no detected matches in the chosen random regions. The lengths $l = 300$ and $l = 600$ are of the upstream and random regions.

to 1000 matches, 20 sites from 1001 to 10000, and 10 sites have more than 10000 matches (40 sites did not have any matches, but this is possible, because they can belong to a different yeast strand or can be from the mitochondrial genome). The matches by TFD sites gave similarly wide spectrum. This statistics supports the observations discussed in (Chen, Hertz, & Stormo 1995) that these site entries have rather varying properties. Therefore for further research we decided to use the matrix representation of IMD database (release 1.0) and the related software MATRIX SEARCH 1.0 (Chen, Hertz, & Stormo 1995). We run MATRIX SEARCH against the complete yeast genome and marked the positions of matches of each of the given 39 yeast transcription factor matrices.

The spectrum of the numbers of matches of IMD sites is still wide (see Table 1), nevertheless it is more even than that for TRANSFAC or TFD site matches, and therefore more appropriate for studies of site combinations. Some properties of the individual sites can be noticed from these matches. For instance, the two of the sites with the highest occurrence rates in the genome - RC2 and PHO2, both have goodness ratio slightly less than 1 for most region lengths, meaning that these sites occur slightly less frequently within upstream regions than in random regions. The reason for this may be the unreliability of the descriptions of these sites, therefore, we excluded them from some of the further experiments. Note that for the region length 300 (600) only 11 (resp. 12) sites have goodness ratio more than 1.5, most of which have low support. In general, for almost all lengths there is a slight preference of 5% to 10% of total occurrences of sites in upstream regions vs. random regions.

Next we transformed the output of MATRIX SEARCH into the format that can be used by the data mining program of (Toivonen 1996) to find frequent combinations of sites present in upstream regions, and the association rules between these sites. (In principle a relatively straightforward enumeration of combinations is also possible, but it would be much less efficient and technically more difficult as using efficient general purpose software for this aim. Also, the straightforward enumeration would become infeasible after the number of known binding sites will increase).

We ran the data mining program on these sets and found all combinations of sites that are present in at least one upstream or random region, and their supports. Based on these outputs we additionally calculated the goodness and the unexpectedness ratio of each combination. All this information was compiled into separate files for each length of the region (with the total size of more than 100 MB). Note that implementing of data mining via storing precomputed information in a specially created database, and thus, effectively reducing later data mining to mere querying the new database has been proposed for instance in (Imielinski & Mannila 1996).

Knowledge Discovery Phase

We developed a program Transcription Factor Combination Discoverer (TFCD) for extracting the combinations with specified parameters from the created files. The TFCD is implemented as a `WWW` site (`http://www.cs.Helsinki.FI/~vilo/Yeast/`) allowing queries via simple user interface on the Web. The user can choose the upstream region length (from 50 to 1000), the minimal support, goodness and unexpectedness ratios and some other parameters, and (optionally) a factor which he wants to be included in the combinations. The system returns all the combinations satisfying the given parameters. For instance, for the regions of length 600, if the minimal unexpectedness ratio is set to 3 (meaning that the combinations should occur in upstream regions at least three times more frequently than expected if the sites were statistically independent), the minimal goodness ratio is set to 5 (meaning that the combinations should be at least five times more frequent in upstream regions than in the random regions), the minimal coverage of the combination to 10 (meaning that at least 10 of the upstream regions should contain this combination), and the number of sites in the combinations from 1 to 4, and if it is required that the factor TFIID should participate in the combination, then we get the result that is given in Figure 1.

Clicking on "Show occurrences in full genome" returns all positions in the total genome where this combination has occurred within the window of the given length. For instance, one of the occurrences for the combination 2 in the Figure 1 is shown on Figure 2.

The closest ORFs are reported. The integers between the factors are the distances between their occurrences. By clicking on the ORF we obtain the annotation for this ORF in MIPS database. Links are provided also to IMD and TRANSFAC databases for the information about the factors.

Clicking on "Show ... association rules" the user can find all the association rules that can be combined from occurrences of this particular combination. For instance, the rules found for the combination 2 of Figure 1 with confidence 0.75 are given in Figure 3.

The first rule means that if in a window of size 600 the sites for `BAF1 CBF1 RAP1/SBF-E/TUF` are present, then always also the site for `TFIID` is present, and this happens totally in 12 upstream regions.

In this way the user can generate hypotheses (containing the factor of his interests) for combinations that might be a part of promoter classes and obtain additional information of different kinds about them.

Discussion and Future Research

The task of data mining is to generate potentially interesting, statistically unexpected hypotheses. We propose a tool for extracting the combinations of transcription factor binding sites that are frequent in upstream regions and have unexpected occurrence pattern. Our tool is guaranteed to find all the combinations satisfying the given parameters in respect to the given set of upstream regions, its counterset, and the chosen set of sites. Therefore, the "guarantees" are only as good as the chosen upstream regions and chosen sites.

The reliability of the chosen upstream regions basically depends on the reliability of the predicted ORFs, and on the way of determining the relative position of the TRU in respect to the ORF. Regarding the ORF positions we entirely rely on the annotations given in MIPS. Regarding the TRU positions relative to ORFs, there are basically three problems: can we assume that the transcriptions start point is close to the ORF, how long is the upstream region with functional importance for the transcription regulation, and how to deal with the upstream regions that overlap with the coding regions of other genes? As far as the transcription start point is concerned, in yeast it is assumed usually to be relatively close to ORF (there are no known long transcribed untranslated regions in yeast). It is particularly unclear how to deal with the upstream regions overlapping with other genes. We are following a naive approach here - we simply discard such regions. To minimize the overlapping, and to maximize the reliability of the predicted genes, we also use only these ORFs that are annotated as at least as coding a protein similar to some other protein.

Note that, the longer the upstream regions, the more overlapping with the coding regions they have. This creates a certain dilemma - either to have different sets of upstream regions for regions of different length, or to discard many nonoverlapping upstream regions when taking the regions of smaller length. As we are in-

```
Transcription Factor Combination Discoverer (TFCD)

region lenght = 600
minimal unexpectedness ratio = 3
minimal goodness ratio       =  5
minimal coverage of the combination = 10 : (0.41%)
minimal coverage of individuals = 20 : (0.83%)
minimal combination lenght = 1
maximal combination lenght = 4
contains patterns:  TFIID

combination:   MCM1.PRTF SRF TFIID
 goodness ratio =  5.02
 upstream coverage: expected =  0.89 in fact = 15  ratio = 16.89
 random   coverage: expected =  0.29 in fact = 3  ratio = 10.26
 unexpectedness ratios upstream vs random =  1.65  +
 Show occurrences in full genome and association rules containing this combination

combination:   BAF1 CBF1 RAP1/SBF-E/TUF TFIID
 goodness ratio = 1000.00
 upstream coverage: expected =  3.41 in fact = 12  ratio =  3.52
 random   coverage: expected =  0.52 in fact = 0  ratio =  0.00
 unexpectedness ratios upstream vs random =  0.00  +
 Show occurrences in full genome and association rules containing this combination

combination:   BAF1 CCBF/SW4+SW6 TAF TFIID
 goodness ratio = 10.95
 upstream coverage: expected =  2.40 in fact = 11  ratio =  4.57
 random   coverage: expected =  0.47 in fact = 1  ratio =  2.13
 unexpectedness ratios upstream vs random =  2.14  +
 Show occurrences in full genome and association rules containing this combination

combination:   BAF1 CYP1.HAP1 TAF TFIID
 goodness ratio =  5.12
 upstream coverage: expected = 10.29 in fact = 41  ratio =  3.99
 random   coverage: expected =  2.24 in fact = 8  ratio =  3.58
 unexpectedness ratios upstream vs random =  1.11  +
 Show occurrences in full genome and association rules containing this combination

total nr. of combinations satisfying the conditions = 4
```

Figure 1: A sample output of a TFCD query. By clicking on "show occurrences" the user can obtain the information where in the genome this combination is present and which gene is downstream from it. Similarly user can view the association rules where the factors of the combination are present.

terested not so much in creating completely unbiased statistics, as in creating as many interesting hypotheses as possible from the complete genome data, we have chosen the first approach.

It seems that the bottle-neck regarding the quality of our hypotheses is the reliability of the identification of the individual sites, which is known not to be very high. Also, the sites are not available for all factors. We are entirely relying on the identification of sites by MATRIX SEARCH and IMD. Although our method may actually help in filtering out some false positives, still improvement in the reliability of site identification is probably necessary for raising the quality of hypotheses generated by our method.

So far our approach has been naive in the sense that we are looking for the combination of sites disregarding their occurrence order and structure. To some degree we overcome this when we match the identified combinations against the whole genome - these matches each return the order and the structure of the combination in the particular match. However a different approach would be to look for frequent combinations of sites in a particular order from the very beginning. A data mining software that can be adjusted for ordered combination discovery is so-called episode analysis tool (Mannila, Toivonen, & Verkamo 1995; Hätönen *et al.* 1996). We are currently implementing a transcription site combination analysis tool based on

```
Find the occurrences of the factors within close range (600)

Factors considered: BAF1 CBF1 RAP1/SBF-E/TUF TFIID
...
Chromosome: VII
...
RAP1/SBF-E/TUF-28-BAF1-74-CBF1-95-TFIID-35-TFIID-103-ORF:(905927-908764 W YGR204w)
...
```

Figure 2: The information about where in the genome this combination (`BAF1 CBF1 RAP1/SBF-E/TUF TFIID`) is present and which gene is downstream from it. It is further possible to look for all possible factors in addition to the selected ones.

```
Find all rules that contain the factors: RAP1/SBF-E/TUF TFIID BAF1 CBF1
        with confidence at least 0.75
Rules are computed from 2391 strictly upstream regions of length 600.

FACTORS  => FACTORS  [Confidence,coverage]

BAF1 CBF1 RAP1/SBF-E/TUF ==> TFIID  [1.00000,12]
BAF1 CBF1 RAP1/SBF-E/TUF ==> GCN4 TFIID  [0.83267,10]
BAF1 CBF1 RAP1/SBF-E/TUF TFIID ==> GCN4  [0.83267,10]
BAF1 CBF1 GCN4 RAP1/SBF-E/TUF ==> TFIID  [1.00000,10]
BAF1 CBF1 MSN4 RAP1/SBF-E/TUF ==> TFIID  [1.00000,6]
BAF1 CBF1 RAP1/SBF-E/TUF REB1 ==> TFIID  [1.00000,6]
```

Figure 3: The association rules that involve the preselcted factors and have support higher than the treshold.

this.

Finally, although our method is currently aimed at yeast and tied to MIPS database, its principles are very general and applicable to any genome or parts of a genome. The implementation of such general data mining tool will be more efficient after the specification of the representations of complete genomes in general databases EMBL or GenBank is published.

Acknowledgments

The authors wish to thank Dr. Chen Qing, Dr. Alfred Zollner, Dr. Tom Flores, Dr. Mary Ann Tuli, Dr. Inge Jonassen, Dr. Hilkka Turakainen, and Dr. Hannu Toivonen for valuable consultations and discussions. A. Brāzma was supported by the Academy of Finland, the Finnish Centre for International Mobility and the Council of Research of Latvia.

Appendix A: Strictly upstream regions

We define an *annotated chromosome* as a pair $\mathcal{C} = (\alpha, \mathcal{A})$, where α is a genomic sequence and A is a set of annotations. An *ORF annotation* is a quintuple $\pi = (I, b, e, d, c)$, where I is a character string called the *identifier*, b and e are integers such that $b < e$, called the *beginning* and *end* positions, d is the *strand* and its value is either "W" (meaning so-called "Watson" strand), or "C" (meaning the "Crick" or complementary strand), and c is an integer $1 \leq c \leq 6$ called the *confidence level.* In practice the ORF annotations are obtained from annotation files given in MIPS database. The confidence level correspond to different values of the classification field in MIPS annotations. The level 1 means "known protein", the level 6 means "questionable ORF". We also assume that each next lower confidence level includes all the higher ones. Thus, the level 6, in fact is, either "questionable ORF", or any of the levels from 1 to 5. In the experiments we basically use genes of the confidence level 4 or 5, meaning at least "similarity to unknown protein" and, respectively, "no similarity".

Given a chromosome $\mathcal{C} = (\alpha, \mathcal{A})$ and an identifier I_0 such that $\pi = (I_0, b, e, d, c)$ is in A, and an integer l, we define the *gene upstream sequence of length* l as the substring $\alpha[b - l..b - 1]$, if $d =$"W"; and reverse(complement($\alpha[e + 1..e + l]$)), if $d =$"C", where reverse means the reverse string and complement means the string obtained by mutually substituting `A` for `T` and `C` for `G`. Given an annotated chromosome (or set of such chromosomes), a length l, and a confidence level c_0, we can define the set of all *upstream regions* of length l with at least the confidence level c_0. From this set we can choose the subset of all upstream regions that are not overlapping with any coding regions (i.e., with substrings at the positions $b..e$ given in some annotations). We call this subset the set of all *strictly* *upstream regions.*

Appendix B: Position weight matrix

A *position weight matrix* is a matrix (array) M of n columns and 4 rows of nonnegative integers such that the sum over each column is the same number d for all columns. We assume that the rows are indexed by $\mathtt{A,C,G,T}$, and write $M[\mathtt{A..T}, 1..n]$, by $M[x, i]$ meaning the element in the i-th column in the row indexed by x. The central notion in using position weight matrices is the *similarity score* s of a matrix $M[\mathtt{A..T}, 1..n]$ against the string α (over Σ) of length n, describing the similarity between the matrix and the string. This notion can be defined in various ways - we use the definition from (Chen, Hertz, & Stormo 1995):

$$t = \sum_{i=1}^{n} \log_2 \frac{M(\alpha[i], i) + 0.01}{p(\alpha[i]) \cdot (d + 0.01)}$$

where $p(a)$ is the probability (the relative frequency) of the character a in the total sequence (or in the database). If we are given a real number, called *cut-off score* or threshold, then for each position in a sequence we can determine whether it is matched by the matrix or not, i.e., if the matching score at this position is at least the cut-off score.

References

Berg, O. G., and von Hippel, P. H. 1988. Selection of DNA binding sites by regulatory proteins. *Journal of Molecular Biology* 200:709–723.

Bucher, P. 1990. Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology* 212:563–578.

Bucher, P. 1996. The eukaryotic promoter database EPD. Technical Report EMBL Nucleotide Sequence Library, Release 47, European Bioinformatics Institute, Hinxton Hall, Hinxton, Cambridge, CB10 1RQ, UK.

Cai, Y., and Chen, C. 1995. Artificial neural network method for discriminating coding regions of eukaryotic genes. *Computer Applications in Biosciences* 11:497–501.

Cardon, L. R., and Stormo, G. D. 1992. Expectation maximization algorithm for identifying protein-binding sites with variable lenghts from unaligned DNA fragments. *Journal of Molecular Biology* 223:159–170.

Chen, Q. K.; Hertz, G. Z.; and Stormo, G. D. 1995. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Computer Applications in Biosciences* 11:563–566.

Fayyad, U.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM* 39:27–34.

Frech, K.; Herrmann, G.; and Werner, T. 1993. Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Research* 21:1655–1664.

Ghosh, D. 1990. A relational database of transcriptional factors. *Nucleic Acids Research* 18:1749–1756.

Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsen, P.; Tettelin, H.; and Oliver, S. G. 1996. Life with 6000 genes. *Science* 274:546–567.

Goodbourn, S., and King, P. 1996. Multi-factorial and combinatorial models for specific transcriptional regulation. In Goodbourn, S., ed., *Eukaryotic Gene Transcription*, 132–164. Oxford University Press.

Hätönen, K.; Klemettinen, M.; Mannila, H.; Ronkainen, P.; and Toivonen, H. 1996. Knowledge discovery from telecommunication network alarm databases. In *12th International Conference on Data Engineering (ICDE'96)*, 115 – 122.

Hertz, G. Z., and Stormo, G. D. 1994. Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps. In *Proc. of Third International Conference on Bioinformatics and Genome Research.* http://www.scri.fsu.edu/mimi/bio94/hertz/hertz.html.

Imielinski, T., and Mannila, H. 1996. A database perspective on knowledge discovery. *Communications of the ACM* 39:58–64.

Kel, O. V.; Romachenko, A. G.; Kel, A. E.; Naumochkin, A. N.; and Kolchanov, N. A. 1995a. Database of transcription regulatory regions on eukaryotic genomes. In *Proc. of the 28th Annual Hawaii International Conference on System Sciences, Biotechnology and Computing*, 42–51. IEEE Computer Society Press.

Kel, O. V.; Romaschenko, A. G.; Kel, A. E.; Wingender, E.; and Kolchanov, N. A. 1995b. A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Research* 23:4097–4103.

Mannila, H.; Toivonen, H.; and Verkamo, A. I. 1995. Discovering frequent episodes in sequences. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, 210 – 215.

Mellor, J. 1993. Multiple interactions control the expression of yeast genes. In Broda, P.; Oliver, S. G.; and Sims, P. F. G., eds., *The Eukaryotic Genome, Organisation and Regulation*, 275–320. Cambridge University Press.

Mitchell, P. J., and Tijan, R. 1989. Transcription regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* 245:371–378.

Pedersen, A. G.; Baldi, P.; Brunak, S.; and Chauvin, Y. 1996. Characterization of prokaryotic and eukaryotic promoters using Hidden Markov Models. In *Proc. of Fourth International Conference on Intelligent Systems for Molecular Biology*, 182–191. AAAI Press.

Penotti, F. E. 1990. Human DNA TATA boxes and transcription initiation sites - a statistical study. *Journal of Molecular Biology* 213:37–51.

Prestridge, D. S. 1991. SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements. *Computer Applications in Biosciences* 7:203–206.

Prestridge, D. S. 1995. Predicting Pol II promoter sequence using transcription factor binding sites. *Journal of Molecular Biology* 249:923–932.

Quandt, K.; Frech, K.; Karas, H.; Wingender, E.; and Werner, T. 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research* 23:4878–4884.

Quandt, K.; Grote, K.; and Werner, T. 1996. GenomeInspector: Basic software tools for analysis of spatial correlations between genomic structures within megabase sequences. *Genomics* 33:301–304.

Schneider, T. D.; Stormo, G. D.; and Gold, L. 1986. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology* 188:415–431.

Svetlov, V. V., and Cooper, T. G. 1995. Review: Compilation and characteristics of dedicated transcription factors in Saccharomyces Cerevisiae. *Yeast* 11:1439–1484.

Toivonen, H. 1996. *Discovery of Frequent Patterns in Large Data Collections.* Ph.D. Dissertation, University of Helsinki, Department of Computer Science, P.O. Box 26, FIN-00014 University of Helsinki, Finland.

Werner, T. 1996. A new approach towards systematic theoretical analysis of gene regulatory regions in genomic sequences. In *Collection of abstracts of International Workshop on Computational Analysis of Eukaryotic Transcriptional Regulatory Elements.* German Cancer Resaerch Fund, Heidelberg.

Wingender, E.; Dietze, P.; Karas, H.; and Knuppel, R. 1996. TRANSFAC: a database of transcriptional factors and their DNA binding sites. *Nucleic Acids Research* 24:238–241.

Wingender, E. 1994. Recognition of regulatory regions in genomic sequences. *Journal of Biotechnology* 35:273–280.

Wolfertstetter, F.; Frech, K.; Herrmann, G.; and Werner, T. 1996. Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Computer Applications in Biosciences* 12:71–80.