LETTER

Predicting Gene Regulatory Elements in Silico on a Genomic Scale

Alvis Brāzma,[1] Inge Jonassen,[2] Jaak Vilo,[3,4] and Esko Ukkonen[3]

[1]European Molecular Biology Laboratory (EMBL) Outstation–Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK; [2]Department of Informatics, University of Bergen, Høyteknologisenteret, N5020 Bergen, Norway; [3]Department of Computer Science, FIN-00014 University of Helsinki, Helsinki, Finland

We performed a systematic analysis of gene upstream regions in the yeast genome for occurrences of regular expression-type patterns with the goal of identifying potential regulatory elements. To achieve this goal, we have developed a new sequence pattern discovery algorithm that searches exhaustively for a priori unknown regular expression-type patterns that are over-represented in a given set of sequences. We applied the algorithm in two cases, (l) discovery of patterns in the complete set of >6000 sequences taken upstream of the putative yeast genes and (2) discovery of patterns in the regions upstream of the genes with similar expression profiles. In the first case, we looked for patterns that occur more frequently in the gene upstream regions than in the genome overall. In the second case, first we clustered the upstream regions of all the genes by similarity of their expression profiles on the basis of publicly available gene expression data and then looked for sequence patterns that are over-represented in each cluster. In both cases we considered each pattern that occurred at least in some minimum number of sequences, and rated them on the basis of their over-representation. Among the highest rating patterns, most have matches to substrings in known yeast transcription factor-binding sites. Moreover, several of them are known to be relevant to the expression of the genes from the respective clusters. Experiments on simulated data show that the majority of the discovered patterns are not expected to occur by chance.

Completely sequenced genomes, together with the emerging DNA microarray technologies enabling the measurement of the gene expression levels in cell cultures (Schena et al. 1995; for a survey, see Ramsay 1998), are opening new possibilities for studying gene regulation. The sequencing of the first eukaryotic genome (the yeast *Saccharomyces cerevisiae*) was completed in 1996 (Goffeau et al. 1996; Mewes et al. 1997). Data about the expression levels of almost all of the ~6000 yeast genes have been obtained (DeRisi et al. 1997; Velculescu et al. 1997; Wodicka et al. 1997) during 1997. In particular, DeRisi et al. (1997) measured the relative expression levels of the yeast genes at seven consecutive time points (in 2-hr intervals) during a shift from anaerobic to aerobic metabolism (diauxic shift). They showed that some of the genes that are known to be involved in metabolic pathways related to the diauxic shift underwent a very significant change in their expression level during the shift. By treating the expression measurements as a time series, it is possible to cluster genes according to similarities in their expression profiles. It may be hypothesized that at least some of the genes in a cluster are regulated by similar mechanisms.

The transcription regulation mechanisms in eukaryotic genomes are not well understood. Evidently, however, an essential role is played by transcription factors, which can bind to particular DNA sequences, called transcription factor-binding sites, believed to be about 5–25 bp long. In yeast, these sites are usually within several hundred base pairs upstream of the respective ORFs (Mellor 1993).

Regular expression type patterns, as well as nucleotide distribution matrices, have both been used for describing transcription factor-binding sites, (e.g., see Bucher 1990; Ghosh 1990; Chen et al. 1995; Wingender et al. 1996). Inference of such descriptions from the sequences that are assumed to contain a site for a particular transcription factor is a difficult problem as the consensus of the different binding sites of the same transcription factor is often rather weak. Algorithms have been proposed for inferring such descriptions from sets of relatively small number of sequences (about 20) in which all

[4]Corresponding author.
E-MAIL vilo@cs.helsinki.fi; FAX 358 9 708 44441.

or almost all of the sequences are known to contain the site for the respective transcription factor (e.g., see Stormo and Hartzell 1989; Wolfertstetter et al. 1996; van Helden et al. 1998). More recently, van Helden et al. (1998) and Yada et al. (1998) have proposed methods for the discovery of putative transcription factor-binding sites from larger data sets. Yada et al. (1998) applied their method to analyze about 400 human promotor sequences.

Apparently, an even more difficult problem is identifying potential binding sites or other regulatory elements from sets of sequences only suspected to contain such elements. In this report, we consider the case when only a small portion of the sequences in the given set may actually contain a common regulatory element, and the total number of sequences may be up to thousands. In this setting, it may not be possible to infer precise binding site descriptions; still, if the number of sequences containing a common regulatory element is larger than would be expected by chance, it may be possible to obtain hints about sequence properties of such an element and in which particular sequences it may be present.

An obvious difficulty in attacking this problem is the computational complexity of the algorithmic problem of discovering interesting sequence patterns in a large collection of sequences only some of which may contain a common pattern. Ultimately the results of such discoveries should be taken as predictions that must be verified by independent, that is, wet biology, means. Still, some validation can be obtained by comparing the discovered site descriptions to the transcription factor database entries, or by statistical means by comparing the distribution of the discovered patterns to the distribution in simulated data.

Pattern discovery methods basically fall into two groups; sequence-driven and pattern-driven methods (for a survey, see Brāzma et al. 1998a,b). Algorithms in the first group normally work by combining the results of pairwise sequence comparisons to form patterns that match the subsets of the sequences. These algorithms are too slow to find patterns that occur in arbitrarily sized subsets of thousands of sequences. Pattern-driven algorithms work by enumerating or searching a predefined pattern class to find patterns and their occurrence frequencies. In these methods, one needs a very fast method for locating all matches of each pattern from the search space. Special data structures and pattern occurrence lists have been used for this purpose, but the methods have been limited to the analysis of smaller data sets.

We have developed a new, more powerful, pattern discovery algorithm that is able to discover various subclasses of regular expression type patterns of unlimited length common to as few as ten sequences from thousands. We used this algorithm for predicting regulatory elements from gene upstream regions in the yeast *S. cerevisiae*.

We considered two cases. First, we looked for patterns that occur more frequently in the gene upstream regions than in randomly chosen regions in the yeast genome. For each pattern present in at least 10 sequences (from >12,000), we calculated a score equal to the ratio of the number of upstream regions that contain the pattern divided by the number of random regions (of the same length and number) that contain the pattern, and rated the patterns according to this ratio.

In the second case, we used information from the yeast genome expression data (DeRisi et al. 1997) to cluster the genes according to their expression profiles. After clustering the upstream regions (treating the expression measurements as time series) we selected characteristic clusters according to some rigorous criteria. We hypothesized that some of the genes in a cluster may contain binding sites for the same transcription factors or other common regulatory elements. We used our algorithm to look for patterns that are over-represented in each cluster as compared with other upstream regions.

We systematically compared the high-scoring patterns that we discovered to the transcription factor-binding site descriptions for the yeast in TRANSFAC database (Wingender et al. 1996). We found that most of the discovered patterns (both from the total set of upstream regions and from the clusters) have matches to substrings of genome regions that contain transcription factor-binding sites. We also compared the distribution of patterns present in upstream regions to the distribution of the patterns that can be discovered in random regions of the genome and showed that the distributions are rather different. The comparison with the TRANSFAC database as well as the overall statistics of the discovered patterns suggest that many of the discovered patterns can be important for the expression profile of the particular clusters of genes or for the transcription or translation initiation in general.

RESULTS

First, we describe the pattern discovery in the complete set of yeast gene upstream regions, then the clustering of the yeast gene expression data, and finally, the results obtained by pattern discovery

from within the subsets of upstream regions of genes sharing similar expression profiles.

We considered three different types of patterns: (P1) substring patterns (i.e., words in the alphabet A, T, G, C); (P2) substring patterns with wild cards (of fixed length); and (P3) patterns with character groups [such patterns can be represented as words over IUPAC code (Corhish-Bowden 1984) characters; here we will use a more explicit notation].

We denote wild-card positions by a dot in the pattern (e.g., TA.A), and the group positions by enlisting all possible characters in square brackets (e.g., T[AT]A). A wild-card position is group position [ATCG], that is, all characters are allowed. For instance, pattern A[TG].C matches all strings that contain a substring beginning with A, followed by either T or G, followed by any character, followed by C. In practice, for reasons of efficiency, we restrict ourselves to various subclasses of these pattern classes (e.g., limiting the number of possible wild cards or group symbols). The implementation of the algorithm, results, data, and additional images are available on the worldwide web at http://www.cs.Helsinki.FI/~vilo/Yeast/.

Discovering Patterns from the Total Set of Upstream Regions

We extracted upstream regions relative to all ORFs, as annotated in the MIPS Yeast genome database (Mewes et al. 1997). Concretely, we extracted seven sets of upstream regions of length 100 from the positions −100 to 0, −150 to −50, −200 to −100, −250 to −150, −300 to −200, −350 to −250, and −400 to −300, a set of regions of length 300 from positions −300 to 0, and a set of regions of length 600 from positions −600 to 0 (all positions are relative to the start codon of the ORF; see Methods). Also we extracted two sets of sequences of the same number and length from randomly selected locations of the same chromosome. These sets of random regions were used as random samples of the yeast genome sequences (the nucleotide and dinucleotide distribution in the random regions reflected that in the genome in general) (1) to compare the upstream regions to random regions for identifying patterns that are more frequent in upstream regions than in the genome in general and (2) to compare the two random sets against each other for testing whether the pattern occurrence statistics resulting from the comparison of upstream and random regions can be explained by chance.

We analyzed these data sets for occurrences of patterns. We presented each pattern that occurred at least 10 times in upstream or random regions as a dot in a two-dimensional plot (see Fig. 1, left column). The vertical axis shows the number of upstream regions, and the horizontal axis the number of random regions, where the pattern is present.

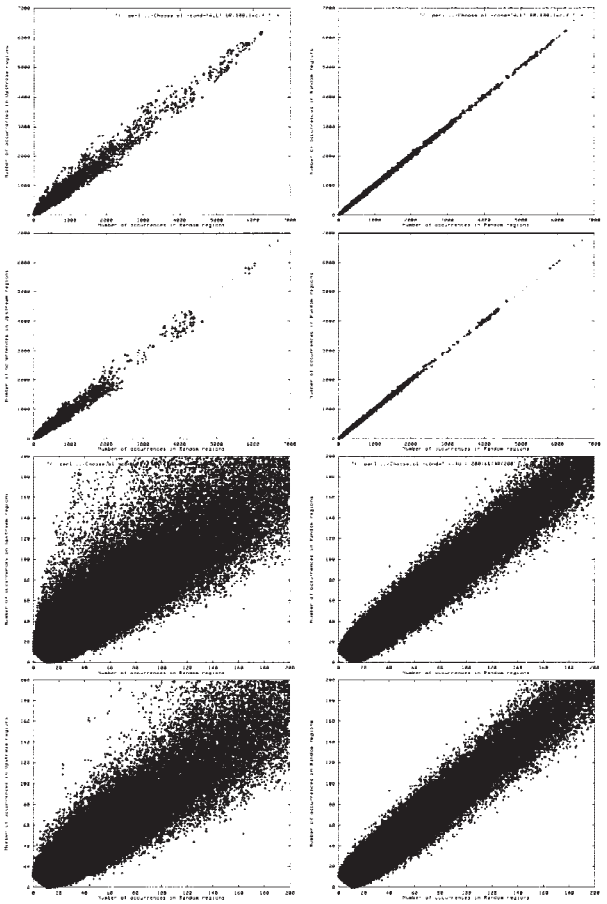

Figure 1 The distribution of all patterns (of unrestricted length) with at most one wild-card symbol in the regions −250 to −150 (upstream from the ORFs) and randomly chosen genomic regions of length 100 bp. Dots in graphs in the *left* correspond to patterns that occur in *x* sequences from the random regions (along horizontal axis) and *y* sequences from the upstream regions (vertical axis). In graphs on the *right*, the upstream regions are replaced by another set of random regions; therefore, these plots show the expected statistics if the regions are chosen at random. (*Top row*) All patterns with at least 10 occurrences. (*Second row*) Subset of top row with all patterns containing at least two characters C or G and not containing any of the substrings AAAA, TTTT, ATAT, or TATA. (*Bottom two rows*) Same plots as in the first two rows, but only including patterns with at most 200 occurrences in upstream or random regions (i.e., zoomed to the lower left corner).

Hence a dot in plot location (x; y) indicates that there is a pattern that occurs in x random regions and y upstream regions. The patterns deviating from the diagonal, and particularly, being above the diagonal, are the ones that can distinguish the upstream regions from the random regions (and, therefore, are likely to distinguish the upstream regions from the genome in general), in contrast to the patterns that fall close to the diagonal and thus occur with the same frequency in upstream and random regions. The dots farthest above the diagonal correspond to the patterns that are potential candidates for regulatory elements. For each pattern we calculated a score as defined by equation (2) in Methods, which is essentially the number of occurrences in the upstream regions divided by the sum of the number of occurrences in the random regions and a correcting constant.

A control experiment (right column in Fig. 1) was done to estimate whether the difference in pattern frequencies observed for upstream versus random sequence segments could be explained by chance. In the control experiments, we compared two sets of random regions. The pattern occurrence statistics obtained when comparing the upstream regions to the random regions is rather different from the statistics obtained when comparing two sets of random regions. We also tested that this considerable difference can be explained neither by higher AT content in the upstream regions, nor by poly(A), poly(T), or poly(AT) patterns. To achieve this goal, we plotted the patterns containing at least two characters C or G and not containing any of the substrings AAAA, TTTT, ATAT, or TATA. The difference between the plots remained essentially as strong (see Fig. 1). Therefore, we conclude that the distribution of patterns in the upstream regions differs from the distribution in regions. In particular, there are some specific patterns that occur considerably more often in upstream regions than in random regions.

The best distinction (as judged by visual inspection) between upstream and random regions by substring patterns was achieved for upstream regions of length 100 when counting matches only on the gene's strand. [The use of only one strand can be justified because of the very distinct distribution of different bases in a region of 300 bp upstream from the start of the gene (see Fig. 3, below, in Methods).] Similar differences were observed for all considered lengths and region relative positions. We also experimented with the three sets of sequences of length 600 and 300 bp, analyzing substring patterns on either strand; and the sequences of length 100, analyzing the patterns that contain up to one wild card. Some results for patterns with at most one wild-card symbol from regions of length 100 bp at upstream positions −250 to −150 are shown in Figure 1.

Many of the top-scoring patterns, particularly, for the region −250 to −150, are effectively poly(T) sequences. Still, as mentioned above, these trivial poly(T) patterns cannot explain the differences in the pattern occurrence statistics compared with random genomic regions; therefore, overall, the patterns not containing poly(T) sequences are significant. We removed from the list of discovered patterns the ones that contain substrings TTTT or AAAA (and additionally the patterns ending in the wild-card—we call the remaining patterns nontrivial) and the list of the 20 remaining highest scoring patterns are given in Table 1 (the numbering of the patterns is given for the total list of patterns including the trivial ones).

We compared the groups of highest scoring nontrivial patterns from each of the seven regions of length 100 bp of various distances with the respective ORFs. We used the program Pratt (Jonassen 1997) to try to find patterns that would be a consensus for a substantial number of patterns for each group. More concretely, we took the 20 highest scoring patterns and used Pratt to discover patterns matching at least 6 patterns. It turned out that only for regions −150 to −50, the highest scoring pattern groups have a relatively good consensus pattern GATG.G.T, the region −200 to −100 has two consensus patterns, T.ACCCG and CGGGT.A, which are mutually symmetric, and the region −250 to −150 has the consensus ACCCG (note that it is a subpattern of T.ACCCG). No significant consensus patterns have been found for other regions.

We also matched the 50 highest scoring nontrivial patterns for each of the regions against all the transcription factor-binding site descriptions given in the TRANSFAC (Wingender et al. 1996) database for the yeast. The results of the exact matches are given in the Table 2 (by an exact match, we mean that the discovered pattern exactly matched a substring in the binding site description). Note that although the highest scoring patterns from neighboring regions are not necessarily similar themselves, the number of coinciding binding sites (from TRANSFAC) matched by patterns from two regions show a considerable correlation with the distance between the positions of the regions.

The complete list of the discovered patterns is available on the World Wide Web.

Table 1. Highest Scoring Nontrivial Patterns with (at Most) One Wild-Card Symbol

No.[a]	Pattern	Score[b]	$N+$[c]	$N-$[d]
	A. Regions −100..0			
2	AAG.AAACAAA	6.54	37	1
6	A.TAAGAACA	5.79	27	0
8	A.AATAGGA	5.61	43	3
9	AAGAAA.CAAA	5.58	26	0
12	GTAACAA.C	5.36	25	0
13	AAA.AACTTA	5.36	25	0
20	ACAAC.TAA	5.09	39	3
21	AG.AAACAAA	5.06	64	8
23	ACAAACAA.A	4.97	48	5
26	AATAGTA.A	4.92	77	11
32	AATAGTATA	4.77	27	1
34	TCACTAC.T	4.72	22	0
35	CAAACA.ACA	4.72	22	0
37	ACA.ATAGA	4.72	55	7
42	AGAGA.ATA	4.63	54	7
47	AATAAACAA.A	4.59	26	1
50	AAAG.ACAAG	4.57	35	3
52	CTAAGAA.A	4.55	53	7
56	A.AAGGGAAG	4.51	21	0
57	CAAA.TAAC	4.50	48	6
	B. Regions −250..−150			
14	TTACCCGC	6.22	29	0
58	GT.ACCCG	5.59	54	5
71	T.ACCCGC	5.48	42	3
126	CGGGTA.T	5.06	64	8
141	G.TACCCG	4.97	48	5
165	CGGGTAA.A	4.87	47	5
178	GTTACCCG	4.83	37	3
305	TACAT.TATA	4.43	65	10
353	TTTCTC.TTT	4.32	46	6
372	TTACCCG	4.30	119	23
379	TTTCCTGT.T	4.29	20	0
405	CTCATCTC.T	4.24	24	1
425	TCACGTGA	4.20	28	2
427	T.ATATATTC	4.20	28	2
454	CGGGTAA	4.12	114	23
460	TGTGT.GAT	4.08	19	0
465	ATTACCCG.A	4.08	19	0
474	G.ACATATAT	4.06	23	1
485	TA.GTAAAC	4.05	27	2
500	TTTCTCT.TT	4.03	47	7

Matches were only allowed on the *W* (gene) strand.
[a]No. of the pattern enumerating them decreasingly by scores (before trivial patterns were removed).
[b]From equation 2.
[c]No. of upstream regions matching the pattern.
[d]No. of random sequences matching the pattern.

Clustering the Gene Expression Data

DeRisi et al. (1997) studied the relative expression rate changes of yeast genes during the diauxic shift. They inoculated yeast cells from an exponentially growing yeast culture into fresh medium and after some initial period, harvested samples at seven 2-hr intervals, isolated their mRNA, and prepared fluorescently labeled cDNA. Two different fluorescent moieties were used—one for cells harvested in each of the successive time points, the other for reference, from cells harvested at the first time point. The cDNAs from each time point, together with the reference cDNA were hybridized to the microarray with ~6400 DNA sequences representing ORFs of the yeast genome. Measurement of the relative fluorescence intensity for each of the ~6400 × 7 elements reflect the relative abundance of the corresponding mRNA in each cell population. The measurement data is available on the Internet.

We used the data from these yeast gene expression studies (DeRisi et al. 1997) and clustered all the genes by similarities in their expression profiles in several alternative ways. To achieve this goal, we developed and implemented a simple algorithm based on discretizing the time series of the measurement space into a simplified form and then clustering these simple time series. Some rigorous selection criteria were used for defining good clusters (for details, see Methods). This produced 32 different clusters containing from 10 to 77 ORFs each and 11 clusters containing at least 25 ORFs (see Table 3).

The most significant changes in gene expression rates during the diauxic shift occurred during the last two time points. This significance is reflected in the clusters that we obtained (although some fluctuations at earlier time points occur for smaller groups of genes, which may be due to noise). Many of the constructed clusters strongly overlap. From the 11 clusters of at least 25 ORFs each, in 8 clusters, the expression level is increasing in the time point 6, in 2 it is decreasing, and in 1 it is unchanged.

Discovering Patterns from the Gene Clusters

We studied whether clusters of genes with similar expression profiles can help to discover sequence patterns putatively describing transcription factor-binding sites. For each cluster, we compared the corresponding upstream regions of length 300 bp against all other upstream regions. The algorithm was used to find the highest scoring patterns containing up to three wild cards. The patterns were

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.