Discovery of DNA Regulatory Motifs

Abigail Manson McGuire and George Church
Harvard University Medical School

Introduction

Short conserved sequence elements located upstream of the transcriptional start site are often binding sites for transcription factors that regulate a group of genes involved in a similar cellular function. Additional genes with this motif in their upstream regions are good candidates for involvement in the same cellular function. Thus, upstream regulatory motifs can provide powerful hypotheses about links in the genetic regulatory network. These upstream regulatory motifs can be discovered computationally by local alignment of upstream regions from coregulated sets of genes. Here we will focus on the AlignACE motif-finding algorithm (Roth *et al.*, 1998).

In order to discover a motif computationally using local alignment methods, an accurate list of coregulated genes, or regulon, is needed. The AlignACE program can tolerate some noise in the alignment process from extra sequence that does not contain the motif of interest, but too much extra sequence will prevent the motif from being aligned. Thus, we must employ additional experimental and computational methods to obtain accurate regulon predictions. mRNA expression data has proven useful for this task. Genes which are up-regulated or down-regulated in a single experimental condition are likely to be coregulated, as well as genes with similar mRNA expression profiles across many conditions or timepoints. There are also several computational methods for predicting regulons. Here we will discuss several methods for predicting regulons, motif-discovery algorithms, parameters for evaluating the significance of a motif, and applications of motif-finding using AlignACE in 17 bacterial genomes and in *Saccharomyces cerevisiae*.

Regulon prediction

Groups derived from mRNA expression data

A powerful method for predicting regulons is to look for patterns in mRNA expression data from DNA chips or microarrays. This has been done for several sets of experimental conditions in *S. cerevisiae* (Tavazoie *et al.*, 1999; Roth *et al.*, 1998). One particularly effective method for obtaining coregulated sets of genes from mRNA expression data is to cluster genes based on their expression profiles over a large number of experimental conditions, or over a large number of time points (Tavazoie *et al.*, 1999). Many known motifs, as well as a new motifs, have been found in *S. cerevisiae* by this method. Such studies will be performed soon in *E. coli* and other bacteria.

Groups derived from conserved operons in prokaryotes

Operons (tandem genes, usually functionally related, that are transcribed onto a single mRNA) are the main transcriptional regulatory units in bacteria. Often, several bacterial operons that share an upstream regulatory motif are coregulated to form a higher-order regulatory unit (a regulon). In different microbial organisms, the genes that make up a regulon are often assorted into operons in different combinations. Two genes which are spatially separated on the chromosome in one organism, but spatially close in several other microbial species, are good candidates for being coregulated. Such pairings between genes have been used to construct larger groups of genes which are predicted to be functionally coupled (Overbeek *et al.*, 1999; http://wit.mcs.anl.gov/WIT2/). Thus, regulons can often be reconstructed by looking at conserved operon composition across other genomes.

Several of the groups constructed in this manner (Overbeek *et al.*, 1999) correspond to known metabolic pathways, including the purine and arginine biosynthesis pathways in *E. coli*. Using the gene groups derived from conserved operons, the regulatory motifs for the DNA-binding proteins that regulate these two pathways (PurR and ArgR, respectively) can be found by AlignACE. Therefore, the purine and arginine biosynthesis regulons, as well as the binding motifs for PurR and ArgR, can be found "from scratch", with no input of biological data other than the genome sequence itself.

Groups derived from multidomain protein fusions and divergently transcribed genes in eukaryotes

In eukaryotes, operons are less common, and each individual gene must have an instance of the upstream regulatory motif in order to achieve coordinated transcription. However, two genes involved in the same pathway are often divergently transcribed from the same intergenic region and share an upstream regulatory site. In addition, genes which are found in the same operons in prokaryotes are often fused together into multidomain protein fusions in eukaryotes. Thus, genes which are found divergently transcribed, or as part of the same multidomain protein fusion in one organism, but scattered in the genome in another organism, are good candidates for co-regulation. Using this idea, predicted regulons can be constructed. The methods of predicting functionally coupled groups of genes based on both conserved operons and multidomain protein fusions can be combined together. These methods will become more powerful as the amount of available genome sequence increases.

Groups based on protein phylogenetic profiles

Pellegrini *et al.* (1998) describe a method for determining functionally related groups of genes based on the assumption that proteins that function together in the cell are likely to evolve in a correlated fashion. In each species, functionally linked genes (i.e. all of the proteins that make up a certain pathway) tend to be either all preserved or eliminated. They constructed groups of evolutionarily correlated genes by constructing a phylogenetic profile for each protein (a vector representing whether an ortholog to this protein is present or not present in each organism) and clustering these vectors based on their similarity.

Groups derived from metabolic and functional pathways

Metabolic and functional group categories, such as those compiled in the KEGG database (Ogata *et al.*, 1999; http://www.genome.ad.jp/kegg/), as well as those in the MIPS and YPD database for *S. cerevisiae* are useful for motif-finding. Since these groups are based on a compilation of experimental data in these organisms, the motifs found within these groups are biased towards known motifs. However, new motifs have also been found in the upstream regions of these groups.

Groups derived from homologs to footprinted regulons in other species

Groups based on known regulons in other organisms can also be used, such as orthologs to the *E. coli* genes known to be regulated by 55 DNA-binding proteins (http://arep.med.harvard.edu/ecoli_matrices/; Robison *et al.*, 1998). This allows for analysis of the conservation of known *E. coli* DNA motifs in other organisms, as well as identification of potential new mechanisms for regulating the same cellular process in more distantly related organisms.

Motif-finding algorithms

The Gibbs sampling algorithm is based on the statistical method of iterative sampling (Liu *et al.*, 1995; Lawrence *et al.*, 1993). This algorithm has been optimized for use with DNA sequence alignments in the program AlignACE (Roth *et al.*, 1998; http://arep.med.harvard.edu/mrnasoft.html). The AlignACE algorithm, along with applications of AlignACE in 17 bacterial genomes and in *S. cerevisiae*, will be discussed in more detail in the next sections.

Only a sampling of the large number of motif-finding algorithms available will be mentioned here. MEME (Multiple EM for Motif Elicitation) is a motif-finding algorithm that uses an expectation maximization algorithm to locate one or more repeated patterns in the input sequence (Bailey & Elkan, 1995; Grundy *et al.*, 1996; http://www.scsc.edu/CompSci/Biomed/MEME). MEME searches by building statistical models of motifs (matrices of discrete probability distributions), and maximizing the posterior probability of these models given the data. MACAW (Multiple Alignment Construction and Analysis Workbench) is a motif-finding algorithm that allows the user to construct multiple alignments by locating, analyzing, editing, and combining "blocks" of aligned sequence segments (Schuler *et al.*, 1991). Several methods based on detection of overrepresented oligonucleotide frequencies have been proposed, including a method by van Helden *et al.* (1998; http://copan.cifn.unam.mx/Computational_Biology/yeast-tools). This simple and fast motif-finding method defines the statistical significance of a site based on tables of oligonucleotide frequencies in non-coding regions. However, this method is limited to short, highly-conserved motifs.

AlignACE Algorithm and Motif analysis

Algorithm

For clarity in describing the Gibbs sampling algorithm used in AlignACE, we will describe the simple case of searching for a single pattern of fixed width W within each of N input sequences ($S_1..S_N$). However, the AlignACE algorithm allows for automatic detection of variable pattern widths as well as multiple motif instances per input sequence. The pattern is described by a probabilistic model of residue frequencies at each position, q_{ij}, where i ranges from 1 to W and j ranges from 1 to 4 (A,C,T, and G). We also consider the background frequencies for each base, p_j, where j ranges from 1 to 4. The p_{ij} are different for each organism. The start site for the motif within each input sequence is designated by a set of positions a_k, where k ranges from 1 to N. To choose the best pattern, the algorithm must choose values of a_k which maximize the ratio of the pattern probability to the background probability.

The first step is to choose random start positions (a_k). The next step (the predictive update step), is to choose one of the N sequences at random (z), and calculate the q_{ij} for the current positions, excluding z. The q_{ij} are calculated using the following equation:

$$q_{ij} = \frac{c_{ij}+b_j}{N-1+B}$$

where c_{ij} is the count of nucleotide j at position i, b_j is a residue-dependent "pseudocount" suggested by Bayesian statistical analysis for the purpose of pattern estimation, and B is the sum of the B_j. The next step (the sampling step) is to consider every possible segment x of length W within z, and to calculate the probability Q_x of generating each such segment x according to the current pattern probabilities q_{ij}, and the probability P_x of generating each segment according to the background probabilities p_j. A weight equal to $A_x = Q_x/P_x$ is assigned to segment x, and with each segment so weighted, a random position is selected. This new position becomes the new values for a_Z, and the algorithm is iterated. The most probable alignment is the one that maximizes the following equation:

$$F = \sum_{i=1}^{W}\sum_{j=1}^{4} c_{ij} \log \frac{q_{ij}}{p_j}$$

where c_{ij} is the count of nucleotide *j* at position *i*. As the pattern description calculated in the predictive update step becomes more accurate, the determination of the location in the sampling step also becomes more accurate.

Several optimizations of the Gibbs sampler algorithm were added in the AlignACE program for use in finding DNA motifs: both strands of DNA are now considered; the near-optimum sampling method was improved in order to obtain higher scoring alignments; simultaneous multiple motif searching was replaced with an iterative masking approach; and the model for base background frequencies was fixed to the background nucleotide frequencies in the genome being considered.

Indices for evaluating the significance of a motif

MAP score

The MAP score is the parameter used by AlignACE to determine the statistical significance of alignments sampled, given the composition of the input sequence (Liu *et al.*, 1995). This is analogous to the *F* parameter described for the simple situation above. MAP scores are normalized so that the score for an alignment of zero sites is assigned a score of zero. The MAP score is higher for similar motifs with greater numbers of aligned sites and for more tightly conserved motifs, and lower for an identical alignment of sites derived from a larger set of input sequences, motifs with more dispersed information content, and motifs enriched in nucleotides more prevalent in the genome.

Specificity score

The specificity score is a measure of how specific a motif is for the sequence in which it was aligned, compared to the genome as a whole. This is the most useful motif for selecting functional motifs. Two different specificity score measures have been described: the specificity score (*S*) is more useful for yeast and eukaryotes (Hughes *et al.*, 1999), while the site specificity score (S_{site}) is more useful for bacterial sequence and genomes containing operons and multiple motif instances within an upstream region (McGuire *et al.*, 1999). The main difference between these two specificity scores is that *S* is a measure of how many of the top predicted ORF targets in the genome were used to align the motif originally, whereas S_{site} measures the probability of obtaining the observed fraction of the top motif instances in the genome are found within the sequence input to AlignACE.

For each motif, we constructed a position-specific weight matrix and searched for additional instances of the motif in the whole genome. We used the Berg & von Hippel weight matrix (Berg & von Hippel, 1987), implemented in the ScanACE program (Hughes *et al.*, 1999) to perform this search. To calculate S_{site}, we look at the top 200 ScanACE hits and calculate the number of these sites located within the upstream regions used to align the motif. We calculate the probability of obtaining this number of sites within this subset of the genomic sequence by random using the following equation:

$$S_{site} = \sum_{i=x}^{\min(s_1,s_2)} \frac{\binom{s_1}{i}\binom{N-s_1}{s_2-i}}{\binom{N}{s_2}}$$

N is the total number of possible sites (equal to the number of base-pairs in the genome), s_1 is the number of ScanACE hits considered (200), s_2 is the total number of possible sites in the set of upstream regions used to align the motif (equal to the number of base-pairs in the sequence input to AlignACE), and i is the intersection of s_1 and s_2. To calculate S, we select the 100 ORFs which are the most likely targets for the factor binding the motif (the ORFs with the strongest sites in the region between -100 and 500 bp upstream of the translational start). This list of gene targets is compared to the list of genes in the group used to find the motif using the same equation as above; however, now N is the total number of ORFs in the genome, s_1 and s_2 are the numbers of ORFs in the group used to find the motif and in the list of target genes, respectively, and x is the number of ORFs in the intersection of the two lists.

<u>Positional bias</u>

This is a measure of bias in the locations of the top ScanACE hits in the genome relative to the start codon of the closest gene. The subset of the top 200 ScanACE hits that fall within a certain distance s of the start codon are considered (300 bp for bacteria and 600 bp for *S. cerevisiae*). A smaller window w (30 bp for bacteria and 50 bp for *S. cerevisiae*) was then chosen such that it contained the maximum number of hits within s. The probability of seeing the observed number of sites or greater in a window of size w out of a sample space of size s was calculated as follows (Hughes *et al.*,1999):

$$P = \sum_{i=m}^{t} \binom{s_1}{i}\left(\frac{w}{s}\right)^{i}\left(1-\frac{w}{s}\right)^{t-i}$$

<u>Similarities between motifs</u>

The CompareACE program (Hughes *et al.*, 1999), used to compare two motifs by finding the best alignment, calculates the Pearson correlation between position specific scoring matrices (Pietrokovski, 1996). Two motifs with a CompareACE score greater than 0.7 are considered similar. New motifs can be compared to databases of known motifs using CompareACE, including a database of footprinted *E. coli* motifs (Robison *et al.*, 1998). CompareACE scores can also be used as a metric to cluster motifs. After calculating a matrix of pairwise CompareACE scores, motifs can be clustered using a simple joining algorithm (Hartigan, 1975). Often when analyzing a large set of motifs, multiple instances of similar motifs will be present, and clustering will reduce the set of motifs to unique clusters.

<u>Palindromicity</u>

Many known bacterial motifs are palindromic motifs. To identify palindromic motifs, we use the CompareACE program to compare a motif and its reverse complement. We used the same CompareACE cutoff score as for comparing motifs to one other (0.7).

AlignACE applications

Applications of AlignACE in 17 bacterial genomes

AlignACE has recently been applied to three different kinds of potentially coregulated groups of genes in 17 microbial genomes (McGuire *et al.*, 1999): groups derived from conserved operons, groups derived from metabolic pathways, and groups derived from homologs in other species to members of known *E. coli* regulons. At least 30% of the known *E. coli* DNA regulatory motifs were found to be conserved in closely related bacteria. Specific, new motifs were found using each grouping method and in each organism. Identifying motifs in bacterial genomes is complicated by the presence of operons. It is difficult to locate the regulatory region for a gene within an operon (see Figure 1). In addition, there are fewer instances of most regulatory motifs because there is usually only one instance per operon instead of one instance per gene. Motif conservation was used to aid in finding new motifs be grouping upstream regions from closely related bacteria, thus increasing the number of instances of each motif in the sequence to be aligned.

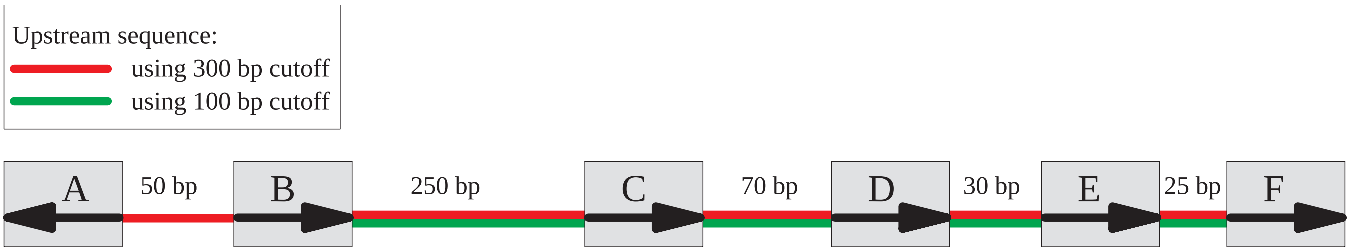

Figure 1. Example of upstream region prediction. If the gene lies within an operon, its promoter could lie several genes upstream. Thus, we must include several possible intergenic regions. Here an operon is defined as two tandem genes separated by less than a certain cutoff distance. We include up to 300 bp of noncoding sequence directly upstream of the head of the predicted operon, as well as the entire sequence of all of intergenic segments of length > 20 bp between the gene of interest and the operon head. This figure shows the predicted upstream region for gene *F*.

The effectiveness of AlignACE in finding known motifs was analyzed. Of the 32 *E. coli* footprinted regulons in our database with 5-100 known binding sites (Robison *et al.*, 1998), 26 regulons have motifs which can be found by AlignACE (81%). Figure 2 shows the MAP scores and S_{site} for the known *E. coli* motifs (colored symbols), as well as all of the rest of the motifs found by AlignACE. Motifs in the upper right hand corner of the plot (nonspecific motifs with good alignments) tend to be either repetitive elements (i.e. *E. coli* BIME elements) or common elements in the genome such as Shine-Dalgarno sequences. The fraction of motifs corresponding to known *E. coli* motifs (colored symbols) is highest closer to the upper left corner of the plot, where the highest-scoring motifs are located. Pooling together upstream regions from orthologs in several closely related organisms (purple diamonds) clearly improves the ability to discriminate known conserved motifs. These conserved motifs tend to have higher MAP scores and lower S_{site} than the same motif found in sequence from *E. coli* alone (green triangles), since there are more instances of the motif to align when you consider sequence from several organisms containing the conserved motif.

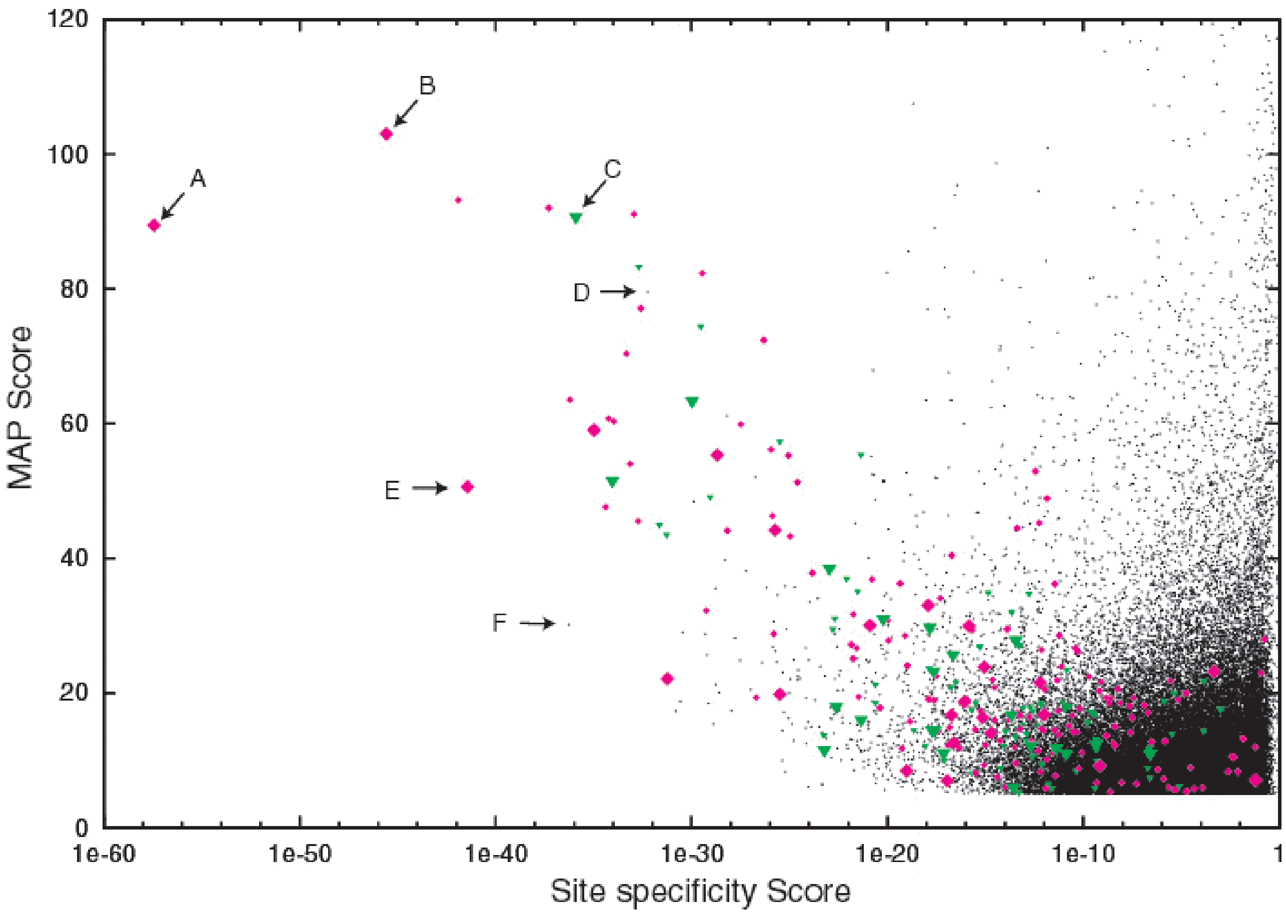

Figure 2. 104,282 motifs from AlignACE runs on 482 gene groups in all 17 organisms are presented here (black points). Known *E. coli* motifs are plotted with colored symbols: green triangles correspond to *E. coli* motifs found in alignments of upstream regions from genes making up regulons in *E. coli* only; purple diamonds correspond to conserved *E. coli* motifs found in sequence from *E. coli* together with sequence from *B. subtilis*, *H. influenzae*, or both. For each known motif, we have represented the motif instance with the lowest value for S_{site} with a large colored symbol, and all other instances of the same motif with a smaller symbol. a) PurR conserved in *E. coli* and *H. influenzae.* b) LexA conserved in *E. coli* and *H. influenzae*. c) LexA in *E. coli.* d) T-box in *B. subtilis*. e) ArgR conserved in *E. coli*, *H. influenzae*, and *B. subtilis*. f) A new motif found in the methane metabolism functional group in *M. thermoautotrophicum*.

The groups obtained from homologs in other species to members of known *E. coli* regulons were used to study conservation of *E. coli* DNA motifs in other organisms, as well as to identify potential new mechanisms for regulating the same cellular process in more distantly related organisms. Results from this analysis are shown in Figure 3. At least 30% of the known *E. coli* DNA regulatory motifs were found to be conserved in closely related bacteria. A new motif can indicate either a different mechanism for regulating a similar cellular process, or divergence of binding site residues in a conserved DNA-binding protein. Different but significant motifs were observed for Crp, LexA, and ArcA in *Bacillus subtilis*; four anaerobic regulons in *Archaeoglobus fulgidus* (NarL, NarP, Fnr, and ModE); and PhoB, PurR, RpoH, and FhlA in other archaebacterial species. The new motifs found in archaebacteria have not been previously documented in the literature.

	EC	HI	BS	MT	RP	HP	BB	TP	CT	CY	MP	MG	AA	MJ	PH	TH	AG
Crp	77*	33*	78*	43*	12	19	14	14*	7	33*	16	14	16	15	23	16	19*
NarL	41*	30*	20*	17*	8	4	1	3	5	11*	2	1	20	6*	7	6	25*
PhoB	32*	6*	34	37*	3	3*	7	3	2	33*	11	9	8*	10	13	11	9
NarP	29*	23*	10*	6*	7	2	1	3	2	7*	2	0	13	3	3	3	15
PurR	28*	19*	29*	22*	3	13	3	3	6	27	2	2	22	19	19	18	20
Fnr	26*	16*	21	16	3	4	1	1	4	8*	0	1	13*	7	6*	6	14
FruR	22*	11	18	7*	3	8	11	6	5	10	10	8	5	5	9	4	4
RpoN	21*	3	22*	11	4	8*	3*	4*	4*	10	5*	3*	16*	14*	14	18	8
LexA	20*	8*	12*	12*	9	8	8	10	7	13*	9	8	5	3	4*	8	3*
RpoH	19*	18*	20	19	17*	18	19	16	15	30	12	11	19	10	9*	13	14
CpxR	18*	8*	24*	23*	14*	9*	5	8	1	18*	0	0	4*	0	3	2	4
ArcA	15*	6*	20*	16*	10*	7	2	1	5	11*	3	1	8	5	6	8	9
Fur	14*	5*	10*	13*	0	2*	1	0	2*	5*	0	0	4*	1	2	2	1*
FhlA	14*	1	1*	6	0	4*	0	0*	1	6	0*	0*	4*	10	11	10	4*
GlpR	13*	12*	12*	6	3	1	3	3	4	5*	2	4	3	2	2	4*	4
ArgR	12*	5*	14*	10*	0	2	2	0	0	10	3	0	11	12	9	9	12
FlhCD	12*	0	9	2	0	8	7	8	8	0	0	0	7	2	0	0	2
TyrR	11*	6*	8*	0	0	1	0	3	5	1	1	0	0	2	1	2	1
ModE	11*	10*	9	11	0	7	0	0	0	8	0	0	10	5	8	8*	13*
GalR	10*	9*	19*	4	1	1	2	5	2*	3	3	2	1	5	6	4*	3
NagC	10*	4	14*	4	2	2	4	1	2	4	1	1	3	3	3	2	2
AraC	10*	1	13	1	0	1	1	2	1	2	1	0	0	0	1	0	1
SoxS	10*	4	8	5	1	2	2	2	4	3	2	1	2	0	3	2	0
TrpR	9*	8*	6	7	0	6*	0	3	5*	6	0	0	5	6	2	5	4
MalT	9*	2	10	8*	0	3	4	4	2	6	3	3	3	5	8*	2	2
CytR	8*	4	9*	5	0	3	1	3	0	2	4	4	1	2	2	2	5
Fis	8*	8*	5	2	7	7	5	3	6	6	0	0	6	4	4	3	3
PdhR	7*	3	5	4	3	0	0	0	3	3	2	2	2	0	5*	1	3
DnaA	6*	6*	3*	3*	6*	5*	2*	4*	5*	5*	1*	2*	3*	1	4*	0	1
MetJ	6*	5*	2	1	0	0	0	1	1	1	0	1	2	1	3	1	3
NtrC	6*	2	18*	5	4*	4*	3*	4*	3*	4	5	3	10*	4	3	8	4
FadR	6*	3*	5	14	1	1	2	1	0	1*	1	0	2	0*	3	2	14*
MetR	6*	4*	4	4	2	1	1	2	2	2	1	2	3	3	3	3	5
RhaS	6*	0	5*	0	0	1	1	0	0	1	0	0	1	0	1	0	0

Figure 3. Summary of results in each organism from AlignACE runs on groups derived from *E. coli* footprinted regulons. Known *E. coli* DNA-binding proteins that regulate at least five genes in *E. coli* are listed in the left column. The numbers indicate the number of regulon members present in that organism. An asterisk indicates that an ortholog to the *E. coli* DNA-binding protein was found.

A red or pink square indicates that a motif similar to the known *E. coli* motif (CompareACE > 0.7) is conserved in this organism and was found by AlignACE with MAP score > 5 and S_{site} < 1e-10. Red indicates that the motif was found by aligning sequence from this organism only; pink indicates that the motif was found by aligning sequence from this organism together with sequence from *E. coli* .

Yellow indicates that there is a new and significant motif found in this organism that is not similar to the *E. coli* motif. These motifs score better than one of two cutoffs: 1) S_{site} < 1e-25, MAP score > 10; 2) S_{site} < 1e-10, MAP score > 10, palindromicity > 0.7, %AT < 80%.)..

Many new motifs were also found in the groups derived from metabolic pathways and in the groups derived from conserved operons. The most significant new motif found in all of the AlignACE runs was an AT-rich motif found in the genes making up the methane metabolism and folate biosynthesis metabolic pathways in *M. thermoautotrophicum* (see Figure 4b). Because many groups of genes based on conserved operons contain a small number of genes, it is beneficial to pool together sequence from closely related organisms. Using this method, we found a highly palindromic, conserved motif that may regulate ferrous ion transport in three archaebacterial genomes (see Figure 4a).

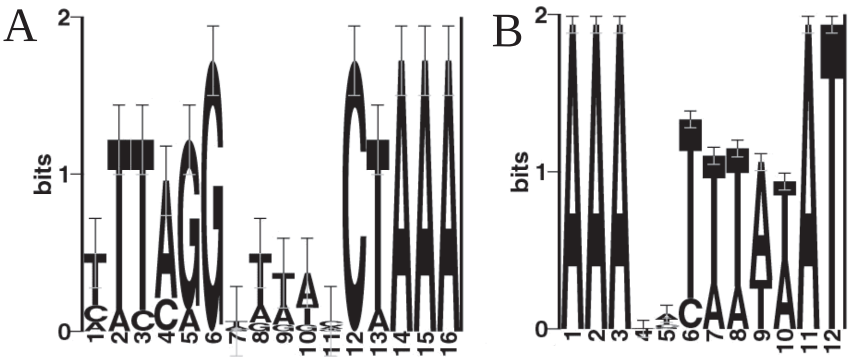

Figure 4. Motif logos (Schneider & Stevens, 1990) for new motifs. a) Palindromic motif predicted to regulate ferrous ion transport in archaebacteria . b) Motif predicted to regulate methane metabolism in *M. thermoautotrophicum*. The height of a stack of letters is proportional to the information content, and the relative frequency of each base is given by its relative height.

Applications of AlignACE in *Saccharomyces cerevisiae*

AlignACE has been used for motif-finding in *S. cerevisiae* with mRNA expression data (Roth et al., 1998; Tavazoie et al., 1999), as well as with functional group categories (Hughes *et al.*, 1999). Selecting upstream regions is more straightforward in yeast as there are no operons. Because intergenic regions are longer in yeast, and regulatory elements are often found further upstream than in bacteria, longer stretches of noncoding sequence have been used for motif-finding (up to 600 bp).

Roth *et al.* (1998) first used AlignACE to find motifs using regulons derived from experiments using Affymetrix GeneChips in the following three systems: galactose utilization, mating type, and heat shock response. By aligning the groups of genes with the largest fold-changes in these conditions using AlignACE, they were able to find the binding sites for many of the factors known to be involved in these processes.

Tavazoie *et al*. (1999) looked at a time-series of mRNA abundance data measured over two synchronized *S. cerevisiae* cell cycles. This data was used to cluster 3000 yeast genes into 30 clusters with similar temporal expression profiles. These groups were highly enriched for genes involved in the same cellular processes. By aligning the genes making up these clusters with AlignACE, they were able to find many known motifs, as well as several new motifs.

Hughes *et al.* (1999) used AlignACE to search for motifs upstream of the genes making up the MIPS and YPD (yeast protein database) functional groups. Many known and new motifs were found in this analysis, including a new motif found upstream of genes involved in proteasome formation. This motif has now been independently identified as the binding site for Rpn4 in the Church lab as well as by Mannhaupt *et al*. (1999). Several other potentially functional new *S. cerevisiae* motifs are also currently being experimentally tested.

Conclusions

The use of motif-finding algorithms with mRNA expression data is a powerful method for understanding the interconnections in genetic regulatory networks. Groups based on mRNA expression data have the drawback that they are dependent on the experimental conditions chosen. These groups can be complemented by theoretically obtained predicted regulons which are less dependent on the experimental conditions chosen. Prediction of such groups will become increasingly powerful as the number of complete genomes in the databases increases.

AlignACE has been used to predict new motifs in *S. cerevisiae*, as well as in 17 bacterial genomes. A new motif found by AlignACE upstream of proteasome subunits in *S. cerevisiae* has been confirmed experimentally. Numerous other new motifs should also be verified experimentally. AlignACE has also been used to analyze conservation and divergence of regulatory mechanisms between microbial genomes.

Acknowledgements

Jason Hughes, Jason Johnson, Martha Bulyk, Saeed Tavazoie, Pete Estep, John Aach, Jong Park, and all of the other members of George Church's lab (Tim Chen, Barak Cohen, Adnan Derti, Phil Juels, Felix Lam, Pam Ralston, Cindy Reyes, Wayne Rindone, Martin Steffen, Matt Tepel, Rob Mitra, Dereth Phillips, and Doug Selinger).

References

Bailey, T. L., & Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. *Proceedings of the Third International conference in Intelligent systems for molecular biology*, 21-29.

Grundy, W. N., Bailey, T. L., & Elkan, C. P. (1996). ParaMEME: A parallel implementation and a web interface for a DNA and protein motif discovery tool. *Comput. Appl. Biosci.* 12, 303-310.

Berg, O. G. & von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* 193, 723-750.

Hartigan, J. A. (1975). Clustering Algorithms. Wiley, New York.

Hughes, J. D., Estep, P. W., Tavazoie, S., & Church, G.M. (1999). Using AlignACE to search for motifs in the upstream regions of 248 groupings of genes in Saccharomyces cerevisiae: calibration of the method and discovery of new sequence motifs. in preparation.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262, 208-214.

Liu, S. J., Neuwald, A. F., & Lawrence, C. E. (1995). Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *Journal of the American Statistical Association*, 90, 1156-1170.

Mannhaupt, G. Schnall, R., Karpov, V., Vetter, I., & Feldmann, H. (1999). Rpn4 acts as a transcription factor by binding to PACE, a nonamer box found upstream of 26S proteasomal and other genes in yeast. FEBS Lett. 450, 27-34.

McGuire, A. M., Hughes, J. D., Church, G. M. (1999). Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. in preparation.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G., & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA*, 96, 2896-2901.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.* 27: 29-34.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci USA*, 96, 4285-4289.

Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucl. Acids Res.* 24, 3836-3845.

Robison, K., McGuire, A. M., & Church, G. A. (1998). Comprehensive Library of DNA-Binding Site Matrices for 55 Proteins Applied to the Complete *Escherichia coli* K-12 Genome. *J. Mol. Biol.* 284, 241-254.

Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol*. 16, 939-45.

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequence. *Nucl. Acids Res.* 18, 6097-6100.

Schuler, G. D., Altschul, S. F., & Lipman, D. J. (1991). A workbench for multiple alignment construction and analysis. *Proteins: Struct. Funct. Genet*. 9, 180-190.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., & Church, G.M. (1999). Systematic determination of genetic network architecture. *Nature Genet.* (in press).

Van Helden, J., Andre, B., & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* 281, 827-842.