

# 3

## Transcriptional Regulatory Network Prediction

JIMMY LIN<sup>1</sup>, DONALD J. ZACK<sup>1,2,3</sup> AND JIANG QIAN<sup>1\*</sup>

<sup>1</sup>*Wilmer Eye Institute,* <sup>2</sup>*Departments of Molecular Biology and Genetics,* <sup>3</sup>*Neuroscience, and Program in Human Genetics and Molecular Biology, McKusick–Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA*

### Introduction

With the development of whole genome sequencing, gene expression microarray technology, and large-scale protein level monitoring, the major post-genomic challenge is no longer data acquisition but analysis and interpretation of large amounts of data. The availability of comprehensive whole-genomic information on DNA, RNA, and protein allows researchers to begin unravelling the intricate relationships between different interacting levels of genomic information, as well as the underlying biological networks of living organisms.

Comparisons of different genomes have revealed that differences in phenotype cannot be sufficiently explained by the observed small differences in DNA coding regions. In other words, the crucial difference between organisms is present not simply at the level of sequence variations between protein encoding genes, but differential regulation and expression of genes. Large-scale experimental and computational methods are now available for researchers to start elucidating the control circuitry and networks of gene expression. The mapping and understanding of these transcriptional regulatory networks are now one of the major challenges of the post-genomic age.

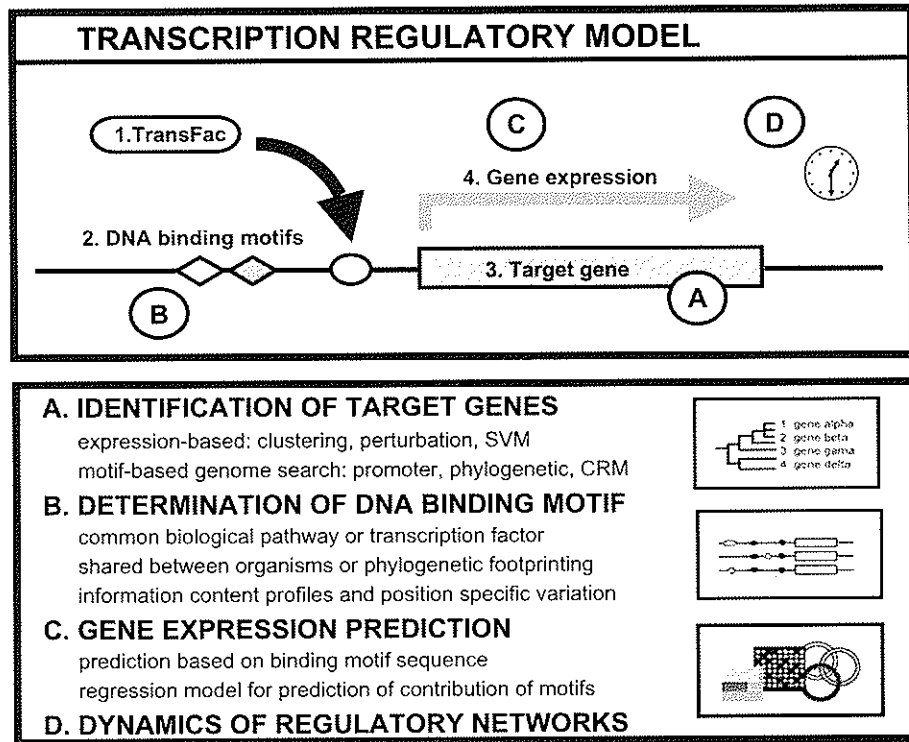
Regulatory networks consist of four basic units: 1) a transcription factor; 2) its DNA binding motif; 3) its target genes; and 4) the expressed target genes (see *Figure 3.1*). When a positively-acting transcription factor binds to the upstream DNA

---

\*To whom correspondence may be addressed (jiang.qian@jhmi.edu)

---

Abbreviations: COG, clusters of orthologous groups of proteins; CRM, *cis*-regulatory module; EPD, eukaryotic promoter database; EST, expressed sequence tag; KEGG, Kyoto Encyclopedia of Genes and Genomes; SVM, support vector machine; TSS, transcriptional start site; UTR, untranslated region.



**Figure 3.1.** Transcriptional regulation overview. In a transcriptional regulatory network, there are four basic units: 1) the transcription factor; 2) its DNA binding motif; 3) its target genes; and 4) its expressed genes. Each unit and their relationship are illustrated here. The transcription factor attaches to its corresponding DNA binding motif, and the target gene is then transcribed and expressed as messenger RNA.

This review focuses on four basic prediction models: A) identification of target genes; B) determination of binding motifs of transcription factors; C) prediction of gene expression; and D) the dynamics of the transcriptional regulatory network. Each is labelled in its corresponding location in the top panel and explained in more detail in the lower panel.

binding motif, downstream target genes are transcribed and expressed. *In silico* methods have been used to elucidate the process for each of the four components; in this review, we focus on the prediction and identification of the latter three, devoting sections to: 1) identification of target genes; 2) determination of binding motifs of transcription factors; and 3) prediction of gene expression. We will also discuss the effects of 4) dynamics on the transcriptional regulatory network.

## Overview

### IDENTIFICATION OF TARGET GENES

In a transcriptional regulatory network, the basic interaction is between the transcription factor and its target genes. An important step in understanding gene regulation is determining which genes are controlled by which transcription factors;

many efforts have been made to identify target genes of transcription factors (Kirmizis and Farnham, 2004). With the current explosion of data from whole-genome sequencing and microarray experiments, researchers are now beginning to be able to examine the effects of transcription factors on a large scale.

Of the many possible data sources for *in silico* prediction of transcription factor targets, our review focuses on two of the more popular approaches: 1) using expression pattern; and 2) binding motifs to predict targets.

#### DETERMINATION OF BINDING MOTIFS OF TRANSCRIPTION FACTORS

The interaction between transcription factors and their target genes is strongly related to the sequence of the DNA binding motifs (Stormo and Fields, 1998). The DNA binding motifs, also called *cis*-regulating elements, are usually located in the promoter region of the target genes. Binding motifs are now being determined not only through *in vivo* and *in vitro* experimentation, but also by *in silico* studies. The *in vitro* and *in vivo* approaches include affinity chromatography and related protein purification methods, yeast-one hybrid cloning, electrophoretic mobility shift assays (EMSAs), protein DNA cross-linking studies, DNaseI footprinting analysis, and more recently, chromatin immunoprecipitation (ChIP). The computational methods for binding motif determination are the major topic of the review.

Three *in silico* approaches to motif discovery are examined here: identification of sequences conserved 1) within a common biological pathway; 2) between organisms; and 3) within a gene family.

#### PREDICTION OF GENE EXPRESSION

With the availability of different types of whole-genome and large-scale data, the computational prediction of gene expression, although still imperfect, is now possible. Specifically, we will examine two examples of such prediction: 1) how gene expression can be predicted from DNA binding motifs; and 2) how different binding motifs contribute to the regulation of gene expression.

#### DYNAMICS

Finally, we will examine the transcriptional regulatory network not only as a static system, but also incorporate time in order to understand how regulation and expression change throughout different cellular processes and life cycles.

### **Identification of target genes**

#### EXPRESSION-BASED TARGET GENE IDENTIFICATION

With the advent of microarray technology, researchers can now measure gene expression level on a whole-genomic scale (Heller, 2002). Since global pictures of gene expression patterns are direct manifestations of the underlying transcriptional regulatory networks, one might be able to utilize gene expression information to recover the transcriptional regulatory networks.

The most common approach to analyse gene expression data is clustering, including methods such as hierarchical clustering, k-means, and self-organizing maps. The underlying principle is ‘guilt by association’, which assumes that genes sharing similar expression profiles also share common functional characteristics, and are similarly regulated. Therefore, unknown genes that share common expression profiles with known genes are likely to share some of the same transcription factors or be controlled by similar processes (Eisen *et al.*, 1998; Marcotte *et al.*, 1999; Gerstein and Jansen, 2000; Altman and Raychaudhuri, 2001).

Another popular approach for target gene identification is by genomic perturbation. This method compares the gene expression patterns of a wild-type organism with an engineered mutant or perturbed variant (i.e. the transcription factor of interest is either knocked out or over-expressed) (deRisi *et al.*, 1997; Hughes *et al.*, 2000b; Livesey *et al.*, 2000). Genes whose expression levels are significantly different in the mutant are often considered as putative targets of the transcription factor. However, with this method, it is difficult to distinguish primary from secondary or indirect effects; thus, the predicted putative genes may not be the direct targets of the transcription factor.

Genomic analyses of transcriptional regulatory networks have identified some interesting gene expression relationships. Yu and colleagues found that genes targeted by the same transcription factors tend to be co-expressed; furthermore, the degree of co-expression is correlated to the number of transcription factors shared (Yu *et al.*, 2003). However, there are often no obvious relationships between the expression profiles of a transcription factor and its regulated gene. The simple correlation coefficient, which is often the basis of popular clustering methods, is insufficient to capture the complicated expression relationships between transcription factors and their targets. This is likely due to the observation that gene regulation generally involves a complex combinatorial array of transcription factors, with positive and negative regulators, so correlations with individual factors can be limited.

Beyond simple correlation coefficients and syn-expression, a variety of methods have been proposed to detect the gene expression relationships between transcription factors and their targets. For example, to further characterize these relationships, Qian and co-workers have developed a local clustering method that was able to find time-shifted, as well as inverted, gene expression profiles (Qian *et al.*, 2001). This captures gene regulation relationships that may be delayed by time effects or have inhibitory effects, which are missed by simple correlation algorithms.

Other methods have focused on the changes in gene expression instead of the absolute level of gene expression. For instance, Kwon and co-workers proposed an ‘event-based’ method, which transforms the expression data into a string of events (e.g. ‘R’ for significant expression rising, ‘F’ for significant expression falling, and ‘C’ for constant or insignificant changes) (Kwon *et al.*, 2003). The expression events are then aligned to detect possible transcription factor-to-target relationships. Similarly, Filkov and co-workers proposed a similar method of ‘edge detection’, where the edge represents significant gene expression changes (Filkov *et al.*, 2002).

Instead of explicitly detecting the expression relationships between transcription factors and their targets, machine learning methods have been used to predict

transcription factor (TF) targets based on gene expression profiles. One implementation is with support vector machines. With this method, each gene pair (TF:target) is characterized by its gene expression patterns in various conditions (e.g. cell cycle, heat shock). In the training stage, the SVM algorithm attempts to find a hyperplane in high dimensional space that best separates the positive and negative examples. With the hyperplane obtained from the training stage, SVM could be used to predict the regulatory targets for transcription factors based on their gene expression (Qian *et al.*, 2003). The method was applied to the *Saccharomyces cerevisiae* genome by using the microarray expression data from many different physiological conditions. Overall, the prediction of the TF–target relationship achieved a success rate of 63% (Qian *et al.*, 2003).

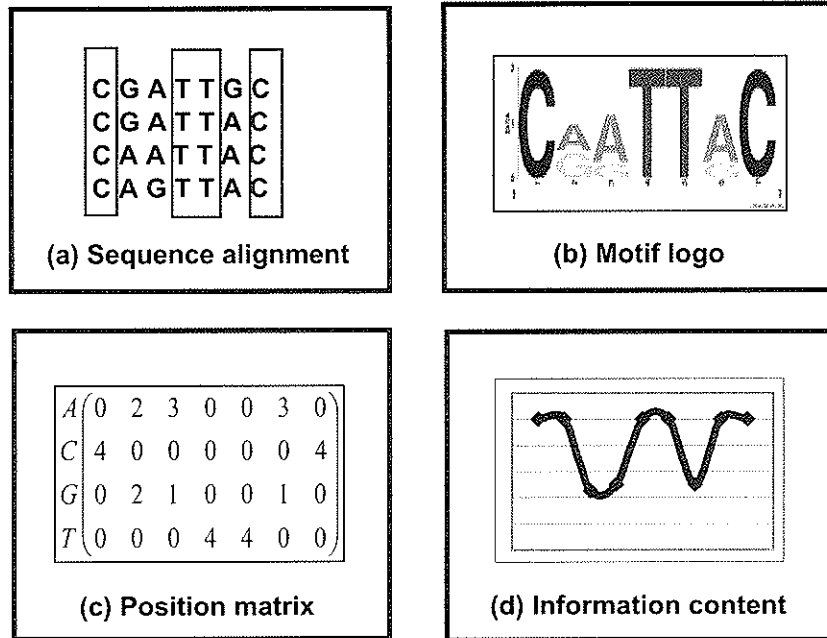
Although complex, ample data demonstrates that there is an intrinsic relationship between transcription factor targets and gene expression. With increasing amounts of available data and more transcription factor-based microarray experiments, it seems likely that researchers will be able to predict more accurately the targets of transcription factors based on gene expression patterns.

#### MOTIF-BASED TARGET GENE PREDICTION

An alternative approach to predict transcription factor targets is to take advantage of the sequences of the DNA binding motif. DNA binding motifs mediate the interaction between the transcription factor and the DNA sequence; these regulatory elements provide the bridge for transcriptional activation, and thus play a crucial part in transcriptional regulation. Therefore, one can search for DNA binding motifs and locate potential downstream target genes.

Before discussing approaches for motif searching, we will first describe some of the ways used to represent a sequence motif. In most cases, one transcription factor can recognize several similar sequences with different affinities. The binding site can often be represented by a sequence alignment to accommodate the sequence variation in some positions (*Figure 3.2a*). The motif logo provides a pictorial view of the binding site (*Figure 3.2b*). The binding sites can also be represented by a matrix, which is a very useful format for motif searching (*Figure 3.2c*). The information content profile (*Figure 3.2d*) summarizes the variation in each position and is defined as  $2 - \sum_{i=A,T,G,C} p_i \log_2 p_i$ , where  $p_i$  is the probability of occurrence of nucleic acid type  $i$  ( $i = A, T, G, C$ ) at this position.

In principle, for transcription factors with known binding sites, one could identify their regulatory targets by searching the whole genome for the presence of the DNA binding motif. Genes that have the DNA binding motif upstream of their coding region are potential target genes. However, the task has proved to be non-trivial. The challenge lies in reducing the high false-positive rate in the prediction, an outcome that stems from the fact that most transcription factor binding sites are short (6–12 bp). Sequences resembling, or identical to, known binding sites often appear and yet they are not biologically relevant regulatory sites. Increasing the accuracy of the predictions requires an understanding of not only the short transcription factor binding sites, but also the transcriptional mechanisms, potentially involving protein–protein interactions and chromatin structure. Despite limited knowledge about the transcriptional mechanisms that explain binding site specificity, several



**Figure 3.2.** Motif discovery models. The four representational models for motif discovery are shown here. a) The consensus sequence representations show the alignment of the sequences, highlight shared, highly conserved sites, and provide an information-intensive view of the motif. b) The motif logo summarizes the different sequences and provides a pictorial view of the motif. Each nucleotide is sized proportional to its dominance in the position, and the total size of the motif is representative of how conserved the site is. c) The position matrix model shows the information as a numeric matrix and enumerates the occurrence of each nucleotide in each position. Other possible matrices provide a normalized or percentage view. d) The information content model reduces each position to how conserved the site is based on the motif. Positions that do not vary between different sequences and that are most conserved have the highest information content.

strategies have been developed to address this problem of high false-positive rates in regulatory site prediction. Among these approaches are: 1) promoter prediction; 2) phylogenetic footprinting; and 3) *cis*-regulatory module (CRM) analysis.

#### *Promoter region identification*

In searching for DNA binding motifs in regulatory regions, researchers have focused on the upstream regions (even though, in eukaryotic genes, the regulatory gene can be upstream, downstream, or even in introns). This captures the most significant relationships, while decreasing the computational search space. However, identifying the upstream sequence of a gene can be non-trivial.

In current sequence databases, precise information about the 5' end termini is often not provided. One exception is the eukaryotic promoter database (EPD) (Schmid *et al.*, 2004), which provides a limited set of experimentally-determined transcriptional start sites (TSSs). High-throughput methods of oligo-capping are

now being used to determine the TSSs in the human genome (Suzuki *et al.*, 2002, 2004; Shiraki *et al.*, 2003). Through these studies, about 9000 full-length 5' UTR (untranslated region) sequences were obtained and the areas of upstream promoter regions were experimentally identified.

There are also some computational approaches for promoter prediction. One example is FirstEF (Davuluri *et al.*, 2001). This program optimizes probabilistic models to find potential first donor sites and CpG-related and non-CpG-related promoter regions on the basis of discriminant analysis. EST (expressed sequence tag) also can be used for finding the transcription start site. With more and more EST libraries available, the method will become more reliable. To map the cDNA sequence on the genome, sometimes one has to deal with long introns, which present a challenge to the conventional alignment programs. New methods, such as BLAT, have been developed to deal with this problem (Kent, 2002).

With the correct identification of the upstream promoter regions, the search for DNA binding motifs can be limited to a smaller search space, increasing the accuracy and computational speed. This greatly improves motif-based gene target predictions.

#### *Phylogenetic footprinting*

Besides limiting the search to specific gene regions based on promoter sequences, researchers have also incorporated information from multiple organisms to increase the accuracy of motif-based gene target predictions. In comparing different organisms, it has been noted that regulatory motifs are more likely to be located in the conserved non-coding regions, as important controlling processes of the transcriptional regulatory network tend to be conserved through evolution. Researchers have taken advantage of this observation and restricted their prediction to only DNA binding motifs in conserved non-coding regions. In this way, the false-positive rate can be reduced significantly, with only a modest cost of sensitivity. This idea has been applied to a variety of systems (Wasserman *et al.*, 2000; Krivan and Wasserman, 2001; Kellis *et al.*, 2003; Lenhard *et al.*, 2003; Thomas *et al.*, 2003). Recently, a similar idea, termed *phylogenetic shadowing*, was proposed to study the biology of *Homo sapiens*. An extensive set of Old World and New World monkeys and hominoids was used to identify functional regions in the human genome (Boffelli *et al.*, 2003).

With genome comparisons, one must also take into consideration the rate of evolutionary change that exists between organisms. Sequence similarity is often directly correlated to phylogenetic distance. Closely related organisms may not have had sufficient time for sequence divergence to occur, and this can result in the apparent conservation of sequences that are not functionally important. Therefore, when comparing closely related genomes, it is important to distinguish whether genome similarity is due simply to lack of change or due to actual functional conservation. This observation is compounded by the fact that much of phylogenetic footprinting research relies on highly related organisms. Rajewsky and colleagues were able to measure the mutational rates between organisms and show that the rates are high enough that the conservation between organisms is not just a function of lack of change, but also a function of selective pressure (Rajewsky *et al.*, 2002).

Another challenge of phylogenetic footprinting is locating orthologous genes between different genomes. Current methods rely mostly on sequence similarity (Clusters of Orthologous Groups of proteins, COG; Kyoto Encyclopedia of Genes and Genomes, KEGG) (Kanehisa, 2002; Tatusov *et al.*, 2003). Multi-genomic approaches are dependent on the identification of corresponding regions between genomes. With organisms that are phylogenetically more related, there is smaller variation between genomes and thus, whole-genome alignment is easier to accomplish. However, with genomes that are not closely related, there may be very little correlation between gene sequences. Since most organisms do not share the same gene composition (Lin and Gerstein, 2000), and divergent genomes may undergo significant gene rearrangement and duplication, there might not be clearly identifiable regions of orthology.

A number of different approaches have been taken to identify orthologous genes or regions between genomes. The most commonly employed methods utilize sequence similarity. With prospective genes, the target genomes are searched with programs such as BLAST (Altschul *et al.*, 1997). Gene pairs with the highest levels of sequence similarity are labelled as orthologous. This approach allows for a general method of comparison, even for diverse genomes. However, due to events such as gene duplication, deletion, and inversion, there is often no simple one-to-one reciprocal relationship between genomes.

One approach taken to reduce the computational complexity is to compare phylogenetically highly related organisms. This way, due to the relative higher levels of global gene conservation and gene sequence similarity, there is often a closer relationship between identified genes, and the identified relationships are often more biologically significant. After obtaining the orthologous genes, one can use sequence alignment tools, such as CLUSTALW or LANGAN, to align the promoter regions of these orthologous pairs (Thompson *et al.*, 1994; Brudno *et al.*, 2003). The tools related to gene regulation prediction are summarized in *Table 3.1*.

Thus, using phylogenetic footprinting, researchers can search for known DNA binding motifs in the upstream regions of the genes and limit their results to highly conserved, non-coding regions. Understanding the basis of orthology and phylogenetic relationships is crucial in interpreting the results obtained in order to make biologically sound and accurate predictions.







#### *Cis-regulatory module (CRM)*

Another method to increase the accuracy of motif-based target prediction is the combinatorial transcription model known as the *cis*-regulatory module (CRM). Whereas phylogenetic footprinting takes advantage of multiple organisms, CRM uses multiple transcription factors to increase the accuracy of target gene predictions.

To specify precisely when and where a gene will be expressed in a genome, multiple transcription factors must act in concert; usually, a single transcription factor is not sufficient to fulfil the task of precise regulation. A *cis*-regulatory module (CRM) is thus defined as the multiple binding sites for multiple transcription factors. This method searches along the genome and looks for windows (e.g. 200 bp) with



Table 3.1. Transcriptional regulation online tools

	TRANSFAC	<a href="http://www.gene-regulation.com">http://www.gene-regulation.com</a>
	SCPD	<a href="http://cgsigma.cshl.org/jian">http://cgsigma.cshl.org/jian</a>
	JASPAR	<a href="http://jaspar.cgb.ki.se">http://jaspar.cgb.ki.se</a>
<b>Genome sequences</b>		
	NCBI	<a href="http://ncbi.nlm.nih.gov">http://ncbi.nlm.nih.gov</a>
	EMSEMBL	<a href="http://www.ensembl.org">http://www.ensembl.org</a>
	UCSC genome browser	<a href="http://www.genome.ucsc.edu">http://www.genome.ucsc.edu</a>
<b>Gene expression data</b>		
	NCBI GEO	<a href="http://www.ncbi.nlm.nih.gov/geo">http://www.ncbi.nlm.nih.gov/geo</a>
	SMD	<a href="http://genome-www5.stanford.edu">http://genome-www5.stanford.edu</a>
<b>Phylogenetic</b>		
	rVISTA	<a href="http://rvista.dcode.org">http://rvista.dcode.org</a>
	UCSC genome browser	<a href="http://www.genome.ucsc.edu">http://www.genome.ucsc.edu</a>
<b>Alignment</b>		
	CLUSTALW	<a href="http://www-igbmc.u-strasbg.fr/BioInfo">http://www-igbmc.u-strasbg.fr/BioInfo</a>
	LAGAN	<a href="http://lagan.stanford.edu">http://lagan.stanford.edu</a>
<b>Motif discovery</b>		
	MEME	<a href="http://meme.sdsc.edu">http://meme.sdsc.edu</a>
	AlignACE	<a href="http://atlas.med.harvard.edu">http://atlas.med.harvard.edu</a>
	CONSENSUS	<a href="http://bifrost.wustl.edu/consensus">http://bifrost.wustl.edu/consensus</a>
	MDscan	<a href="http://mdscan.stanford.edu">http://mdscan.stanford.edu</a>
	oligo-analysis	<a href="http://rsat.ulb.ac.be/rsat">http://rsat.ulb.ac.be/rsat</a>

There are an increasing number of online tools for the many different aspects of transcriptional regulatory network prediction. This table provides a summary of more popular tools for each of the major functional categories, along with the web address. The appropriate reference sources are provided in the text.

high occurrences of motifs. The benefit of this CRM approach is that it not only helps us to understand the interactions between the transcription factors, but also significantly reduces the false-positive rate in our predictions. The method has been successfully applied in several studies, including predicting the regulatory regions for *Drosophila* embryo development (Berman *et al.*, 2002) and liver-specific and muscle-specific gene expression (Wasserman and Fickett, 1998; Krivan and Wasserman, 2001).

To achieve the best performance with CRM-based methods, one needs to consider several factors, such as window size, motif score cutoff, density of motifs in the window, and the combination of transcription factors (Lifanov *et al.*, 2003). Various models have been trained with sets of optimized parameters. With careful setting of the parameters of CRM, the combination of multiple transcription factors allows for a biologically-based model to improve the accuracy for target gene identification

and provides an additional tool for understanding transcriptional regulatory networks. To further improve the CRM method, we believe that the density of motifs in a window might not be sufficient. Finding the grammar of combinatorial regulation (i.e. constraint on orientation, or distance between these motifs) could yield a more powerful approach.

In summary, because searching for DNA binding motifs in the genome produces a great number of false positives, additional information must be incorporated for accurate gene target predictions. By restricting the search to: 1) upstream promoter regions; and 2) conserved non-coding regions; and taking into account the combinatorial effects of multiple transcription factors, researchers are now able to identify more accurately the target genes of transcription factors, based on the DNA binding motifs. With the ability to make predictions of target genes based on both gene expression data and DNA binding motifs, researchers now have many potential tools for understanding the interaction between transcription factors and their targets. This provides a basis for constructing and elucidating transcriptional regulatory networks.

### **Prediction and discovery of DNA binding motifs**

Besides predicting transcription factor targets, the understanding of transcriptional regulatory networks also involves an appreciation of the mechanisms that bridge the gap between transcription factors and the final pattern of gene expression. Binding motifs are important in bridging this gap and thus play a key role in controlling gene regulation. Different initiatives have been taken to understand and predict the exact sequence of DNA binding motifs.

The goal of motif discovery is to find the regulatory motifs (usually 6–12 bp long) that are shared by a set of promoter sequences. In other words, motif discovery programs try to pinpoint these short regulatory motifs from much longer promoter sequences that harbour the motifs. The data source for motif discovery could be the genes that share a similar gene expression pattern, the promoter sequences detected by ChIP–chip approach (Horak *et al.*, 2002; Lee *et al.*, 2002), or the genes that share the same function or the same pathway. Because of technical noise and imperfection of the assumption (i.e. the genes having the same expression patterns share the same transcription factors), the desired motifs may be absent in some input sequences, and good programs are expected to be tolerant to this.

As noted in the Introduction, the three approaches to motif discovery that we will review are methods based on: 1) motifs used within a common biological pathway; 2) those shared between organisms; and 3) those conserved within a gene family. The common paradigmatic assumption for all three approaches of motif discovery is that a group of sequences that have similar properties of gene control will tend to have a common motif. Detailed examination of the similarity and differences between these sequences will allow researchers to understand what the motif is and which positions are conserved. The specific details of motif discovery are discussed below.

#### COMMON BIOLOGICAL PATHWAY OR TRANSCRIPTION FACTOR

Many motif discovery methods are based on the assumption that genes sharing similar expression patterns or the same function are likely to be regulated by some

common transcription factors. Therefore, their promoter sequences should contain common short sequence motifs that are the binding sites of the transcription factors. Many motif-discovery software approaches pinpoint these motifs from a set of much longer promoter sequences (e.g. MEME, CONSENSUS, AlignACE, Gibbs sampler, and MDscan).

These programs can be classified into two types: 1) local sequence alignment; and 2) over-representative words.

*Local sequence alignment* programs include CONSENSUS (Hertz *et al.*, 1990; Hertz and Stormo, 1999), MEME (Bailey *et al.*, 1997), AlignACE (Hughes *et al.*, 2000a), and Gibbs sampler (Lawrence *et al.*, 1993). From a set of input promoter sequences, these programs determine the best local sequence alignment and compare the result with the likelihood of obtaining these local alignments by chance. The empirical scoring schemes for these programs vary, but are mostly based on similar factors: sequence similarity, number of sequences, and length of the sequences. Confidence levels and P-values are generated based on the probability of generating the alignment by chance, incorporating factors such as the frequency of each nucleotide in the background, or using a Markov model.

*Over-represented words* is a second motif discovery tool. Instead of finding the best local sequence alignment, motif discovery programs such as MobyDick (Bussemaker *et al.*, 2000) and oligo-analysis (van Helden *et al.*, 1998) enumerate all possible ‘words’ in the input sequences, and find the motifs (‘words’) that are most over-represented. This circumvents many of the potential problems of sequence alignment. However, due to the nature of the high conservation that is required in this technique, often only core-conserved motifs are found; positions with variations are often missed.

One method to expand the conservation of this technique is to expand the motif alphabet from ‘ACGT’ to ‘ACGTRYWSMKHBVD’. The additional letters represent the potential combination of letters, which allows for more variation in the motifs. The trade-off is, of course, that many more possible motifs have to be considered, and the computation time is dramatically increased.

Whether one uses local sequence alignment or over-represented words, the process of identifying common sequences that are controlled either in a common biological pathway or by a common transcription factor is similar. With either technique, the DNA binding motif can be determined and provide further understanding for the transcriptional regulatory network.

#### SHARED BETWEEN ORGANISMS OR PHYLOGENETIC FOOTPRINTING

Besides being shared in a common biological process, motif discovery can also be based on sequences that are shared between different organisms. This method, called phylogenetic footprinting, as previously mentioned, is also widely used in motif discovery. There are already many studies aimed in this direction.

Phylogenetic footprinting can be applied to bacteria without much complication, since they have upstream structures that are simple and short. By comparison of the upstream sequences from orthologous genes, conserved regions are identified that are often associated with regulatory elements. In a study on gamma-proteobacteria, it was found that three species were sufficient for motif predictions, with a resulting success

rate of 74% (McCue *et al.*, 2002). However, for more complex genomes, such as yeast, phylogenetic footprinting cannot directly yield functional binding motifs. As previously mentioned, closely related organisms may not have had sufficient time for mutational change and thus will be highly similar in terms of sequence. Therefore, when comparing closely related genomes, it is important to distinguish whether genome similarity is due simply to lack of change or to functional conservation. To distinguish functional and non-functional conserved regions, one usually needs to further apply the traditional motif discovery approaches (see above) to the conserved sequences obtained from phylogenetic footprinting. As one example, Cliften and colleagues were able to discover 2771 conserved short sequences (6- to 30-mers) from multiple sequence alignments of four yeast species' intergenic regions. To identify functional motifs in these n-mers, the sequences from genes with similar functional annotations were further compared. Eighteen n-mers that were identified from this comparison are most likely functional (Cliften *et al.*, 2003).

Additional methods besides the traditional comparisons of conserved regions obtained from phylogenetic footprinting are now being developed. For example, Kellis and co-workers proposed a new approach to functional motif discovery based on the observation that the conservation of functional motifs is higher in: 1) intergenic regions over random motifs; 2) intergenic regions over coding regions; and 3) divergent over convergent intergenic regions (Kellis *et al.*, 2003). They applied these observations to the yeast genome and detected 72 motifs, including 28 of the 33 known motifs.

Thus, similarities between phylogenetically related organisms can be taken advantage of in the determination of DNA binding motifs. This method not only provides an additional source of information, but also takes into consideration conservation rates. Comparative genomics and phylogenetic footprinting continue to be useful tools in target prediction, as well as in DNA binding motif prediction.

#### CONSERVED IN GENE FAMILY USING INFORMATION CONTENT

Besides using shared function or pathway and taking advantage of phylogenetic information, information content in a gene family is another characteristic used to predict DNA binding motifs. This method relies on the fact that not every position in a binding motif requires the same degree of conservation. While some positions show high stringency (i.e. must be a certain nucleotide for high binding affinity), other positions may allow for two, three, or even all four nucleotides. For instance, homeobox transcription factors strongly prefer the core binding motif to be 'ATTA', while considerable flexibility is often allowed in the flanking positions, and the flexibility allowed can vary between different homeodomain proteins. At each position, information content can be calculated (*Figure 3.2d*). The higher the information content, the more conserved the position is. Recently, it has been found that the profile of information content can be used for motif discovery.

For the structurally related families of transcription factors, the information content profile of their binding sites usually show certain similarity. In other words, the position-specific variation in the binding sites of related transcription factors is often similar. Sandelin and colleagues constructed familial binding profiles for well-characterized transcription factor families (Sandelin *et al.*, 2004). To predict the

binding motif of a transcription factor belonging to a particular transcriptional factor family, the familial binding profile can be used and the algorithm can search the specific patterns.

Of note, evolutionary conservation is different from positional conservation. Evolutionary conservation describes how regulatory motifs evolve from species to species. Comparing multiple species, Moses and co-workers observed that the positions in binding motifs with smaller information content (i.e. positions that allow for more variations) usually have a larger rate of evolution (Moses *et al.*, 2003). The correlation between information content and the rate of evolution can help distinguish the true motifs from artifacts of motif discovery programs, since the falsely predicted motifs lack such a correlation.

Taking into consideration different conservation and evolutionary rates in different positions of the DNA motif further helps us understand the interaction between transcription factors and their binding sites. These information content profiles are not only useful for analysis, but provide another tool for prediction and the confirmation of predictions. In a particularly interesting recent study, the occurrence and effects of evolutionary changes in DNA binding elements versus changes in the transcription factors themselves was analysed (Wittkopp *et al.*, 2004).

In summary, there are a variety of methods and tools to predict DNA binding motifs. The three main sources of data mentioned in this section include commonality within biological pathways, between organisms, and shared within a gene family. However, even with the simple paradigm of locating conserved positions between sequences, there is complexity due to different evolutionary rates, information content profiles, and conservation in different positions. Therefore, one must take into consideration the different possible sources of data and be mindful of the different levels of complexity for the accurate prediction of the DNA binding motifs.

### **Relate binding motif and gene expression**

In understanding the transcriptional regulatory network, besides determining the transcription factor targets and the DNA binding motifs, predictions can also be made of the effects on gene expression based on: 1) binding motif sequence; or 2) differential contribution of DNA binding motifs.

#### PREDICTING EXPRESSION PATTERN FROM SEQUENCE

Since transcription factors regulate their targets through interaction with the regulatory elements in the promoter regions of target genes, the presence of regulatory elements is associated with the gene expression levels of the target genes. However, in most cases, the binding sequences in the genome are non-functional. Thus, when we relate the sequence and gene expression, it is important to distinguish functional from non-functional sequences.

One example used to relate binding motifs and gene expression is the 6-mer analysis by Chiang and co-workers. Phylogenetically and spatially conserved 6-mer pairs (regulatory templates) in yeast were identified and found to be associated with gene expression changes in different conditions (Chiang *et al.*, 2003). For instance,

with a given set of genes containing the same 6-mer pairs, if these two 6-mer sequences are related heat shock transcription factor binding sites, this set of genes is often differentially expressed in heat shock conditions. A Bayesian approach was employed to learn the complex combinatorial code underlying gene expression (Beer and Tavazoie, 2004). The genes were clustered into 49 groups based on their gene expression patterns. With the rules learned from the relationship between the gene expression patterns and regulatory motifs, the authors could correctly re-assign expression patterns for 73% of genes based purely on the promoter sequences. This exciting finding means that comparison can be made between predicted and experimental expression patterns to refine our understanding of the mechanisms of transcriptional regulatory networks

#### REGRESSION MODEL

A totally different approach for utilizing both gene expression and binding sequence data is that based on the regression model. Two representative implementations are REDUCE (Bussemaker *et al.*, 2001) and MOTIFREGRESSOR (Conlon *et al.*, 2003). The regression model is based on an assumption that the binding motifs contribute additively to the gene expression level:

$$E_g = \alpha + \sum \beta_m * S_{mg} + e_g \quad (3.1)$$

where  $E_g$  is  $\log_2$  expression level for gene  $g$ ,  $\alpha$  is the baseline expression level, and  $e_g$  is the error term specific to gene  $g$ .  $S_{mg}$  is the match score of motif  $m$  on gene  $g$ .  $\beta_m$  represents the increase or decrease in expression level caused by the presence of motif  $m$  and will be estimated by regression.

From this model, different contributions of the motifs are calculated for the resulting expression pattern. For example, the expression of one gene can be entirely controlled by one transcription factor, or shared between five. The results from this modelling can provide a set of motifs that are active in a given condition. More importantly, it can provide the putative influence of these motifs on a gene at various conditions. This begins to model the fluctuating nature of the influence of transcription factors throughout the life cycle, and provides a more dynamic view (see *Dynamics* below). One such example is the MCB motif, which fluctuates during the cell cycle (Bussemaker *et al.*, 2001).

Thus, in our exploration of transcriptional regulatory networks, we see that not only can the transcriptional targets and DNA binding motifs be predicted, but the expression and differing contributions of different genes can also be predicted and modelled. This provides a glimpse of the potential power of these models in describing the biological processes of gene control and regulation, not only in a specific point in time, but as a dynamic system.

#### Dynamics

Traditionally, gene regulation has been modelled as a fixed relationship between different transcription factors, DNA binding motifs, expression patterns, and target genes. However, there is much more than a simple static relationship. The basic question in this field is 'which transcription factors regulate which genes under

which conditions?’ In many cases, the last phrase is neglected and the question is truncated to ‘which transcription factors regulate which genes?’ The dynamic nature of gene control is often ignored in functional genomics, whereas in developmental biology, precise timing and control of these genes is the very foundation of the field (Davidson *et al.*, 2002). Instead of just cataloguing the regulatory motifs in genomes and the specific function of a gene, one can identify the active regulatory motifs during certain conditions (e.g. heat shock) and determine which genes are functioning at which times (Segal *et al.*, 2003).

Previous studies have focused more on motif finding, and paid less attention to the relationship between these motifs and cellular conditions. Motif regression modelling is an attempt to take on this challenge (Bussemaker *et al.*, 2001; Conlon *et al.*, 2003). Recently, a systematic study on regulatory network dynamics has been carried out on the yeast genome. It shows that distinct sections of the regulatory network are used in different conditions (Luscombe *et al.*, 2004). With increased knowledge about regulatory relationships and more large-scale datasets available (e.g. microarrays under various conditions), one could gain a dynamic picture of gene regulation and hopefully eventually understand the underlying machinery of a living cell.

## Conclusion

In summary, understanding transcriptional regulatory networks presents an exciting and intriguing challenge. Knowing how genes are controlled at the sequence level will enable researchers to tackle other basic genetic processes involving mRNA expression, protein expression, and gene function. Current *in silico* methods have not only made progress in the prediction of gene targets, binding motifs, and gene expression networks, but have also started to provide a dynamic, conditional view of gene control. The understanding of gene regulation on a fundamental level will allow many translational advances, impacting many fields, including drug design, cancer therapeutics, molecular diagnostics, and gene therapy.

## Acknowledgements

J.Q. is supported by a grant from the NIH 1R03EY015684-01 and a generous gift from Mr. and Mrs. Robert and Clarice Smith. D.J.Z. is the Guerrieri Professor of Genetic Engineering and Molecular Ophthalmology, and the recipient of a Research to Prevent Blindness Senior Investigator Award.

## References

- ALTMAN, R.B. AND RAYCHAUDHURI, S. (2001). Whole-genome expression analysis: challenges beyond clustering. *Current Opinion in Structural Biology* **11**, 340–347.
- ALTSCHUL, S.F., MADDEN, T.L., SCHAFFER, A.A. *ET AL.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.
- BAILEY, T.L., BAKER, M.E. AND ELKAN, C.P. (1997). An artificial intelligence approach to motif discovery in protein sequences: application to steroid dehydrogenases. *Journal of Steroid Biochemistry and Molecular Biology* **62**, 29–44.

- BEER, M.A. AND TAVAZOIE, S. (2004). Predicting gene expression from sequence. *Cell* **117**, 185–198.
- BERMAN, B.P., NIBU, Y., PFEIFFER, B.D. *ET AL.* (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 757–762.
- BOFFELLI, D., MCAULIFFE, J., OVCHARENKO, D. *ET AL.* (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394.
- BRUDNO, M., DO, C.B., COOPER, G.M. *ET AL.* (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* **13**, 721–731.
- BUSSEMAKER, H.J., LI, H. AND SIGGIA, E.D. (2000). Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 10096–10100.
- BUSSEMAKER, H.J., LI, H. AND SIGGIA, E.D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics* **27**, 167–171.
- CHIANG, D.Y., MOSES, A.M., KELLIS, M., LANDER, E.S. AND EISEN, M.B. (2003). Phylogenetically and spatially conserved word pairs associated with gene expression changes in yeasts. *Genome Biology* **4**, R43.
- CLIFTEN, P., SUDARSANAM, P., DESIKAN, A.M. *ET AL.* (2003). Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**, 71–76.
- CONLON, E.M., LIU, X.S., LIEB, J.D. AND LIU, J.S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 3339–3344.
- DAVIDSON, E.H., RAST, J.P., OLIVERI, P. *ET AL.* (2002). A genomic regulatory network for development. *Science* **295**, 1669–1678.
- DAVULURI, R.V., GROSSE, I. AND ZHANG, M.Q. (2001). Computational identification of promoters and first exons in the human genome. *Nature Genetics* **29**, 412–417.
- DERISI, J.L., IYER, V.R. AND BROWN, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.
- EISEN, M.B., SPELLMAN, P.T., BROWN, P.O. AND BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.
- FILKOV, V., SKIENA, S. AND ZHI, J. (2002). Analysis techniques for microarray time-series data. *Journal of Computational Biology* **9**, 317–330.
- GERSTEIN, M. AND JANSEN, R. (2000). The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function? *Current Opinion in Structural Biology* **10**, 574–584.
- HELLER, M.J. (2002). DNA microarray technology: devices, systems, and applications. *Annual Review of Biomedical Engineering* **4**, 129–153.
- HERTZ, G.Z. AND STORMO, G.D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577.
- HERTZ, G.Z., HARTZELL, G.W., 3RD AND STORMO, G.D. (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Computer Applications in the Biosciences* **6**, 81–92.
- HORAK, C.E., LUSCOMBE, N.M., QIAN, J. *ET AL.* (2002). Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes and Development* **16**, 3017–3033.
- HUGHES, J.D., ESTEP, P.W., TAVAZOIE, S. AND CHURCH, G.M. (2000a). Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **296**, 1205–1214.
- HUGHES, T.R., MARTON, M.J., JONES, A.R. *ET AL.* (2000b). Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126.
- KANEHISA, M. (2002). The KEGG database. *Novartis Foundation Symposium* **247**, 91–101; discussion 101–103, 119–128, 244–252.
- KELLIS, M., PATTERSON, N., ENDRIZZI, M., BIRREN, B. AND LANDER, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254.



- KENT, W.J. (2002). BLAT – the BLAST-like alignment tool. *Genome Research* **12**, 656–664.
- KIRMIZIS, A. AND FARNHAM, P.J. (2004). Genomic approaches that aid in the identification of transcription factor target genes. *Experimental Biology and Medicine (Maywood)* **229**, 705–721.
- KRIVAN, W. AND WASSERMAN, W.W. (2001). A predictive model for regulatory sequences directing liver-specific transcription. *Genome Research* **11**, 1559–1566.
- KWON, A.T., HOOS, H.H. AND NG, R. (2003). Inference of transcriptional regulation relationships from gene expression data. *Bioinformatics* **19**, 905–912.
- LAWRENCE, C.E., ALTSCHUL, S.F., BOGUSKI, M.S., LIU, J.S., NEUWALD, A.F. AND WOOTTON, J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214.
- LEE, T.I., RINALDI, N.J., ROBERT, F. *ET AL.* (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804.
- LENHARD, B., SANDELIN, A., MENDOZA, L., ENGSTROM, P., JAREBORG, N. AND WASSERMAN, W.W. (2003). Identification of conserved regulatory elements by comparative genome analysis. *Journal of Biology* **2**, 13.
- LIFANOV, A.P., MAKEEV, V.J., NAZINA, A.G. AND PAPATSENKO, D.A. (2003). Homotypic regulatory clusters in *Drosophila*. *Genome Research* **13**, 579–588.
- LIN, J. AND GERSTEIN, M. (2000). Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Research* **10**, 808–818.
- LIVESEY, F.J., FURUKAWA, T., STEFFEN, M.A., CHURCH, G.M. AND CEPKO, C.L. (2000). Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene *Crx*. *Current Biology* **10**, 301–310.
- LUSCOMBE, N.M., BABU, M.M., YU, H., SNYDER, M., TEICHMANN, S.A. AND GERSTEIN, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308–312.
- MARCOTTE, E.M., PELLEGRINI, M., THOMPSON, M.J., YEATES, T.O. AND EISENBERG, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86.
- MCCUE, L.A., THOMPSON, W., CARMACK, C.S. AND LAWRENCE, C.E. (2002). Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Research* **12**, 1523–1532.
- MOSES, A.M., CHIANG, D.Y., KELLIS, M., LANDER, E.S. AND EISEN, M.B. (2003). Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evolutionary Biology* **3**, 19.
- QIAN, J., DOLLED-FILHART, M., LIN, J., YU, H. AND GERSTEIN, M. (2001). Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *Journal of Molecular Biology* **314**, 1053–1066.
- QIAN, J., KLUGER, Y., YU, H. AND GERSTEIN, M. (2003). Identification and correction of spurious spatial correlations in microarray data. *Biotechniques* **35**, 42–44, 46, 48.
- RAJEWSKY, N., SOCCI, N.D., ZAPOTOCKY, M. AND SIGGIA, E.D. (2002). The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Research* **12**, 298–308.
- SANDELIN, A., ALKEMA, W., ENGSTROM, P., WASSERMAN, W.W. AND LENHARD, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research* **32** Database issue, D91–94.
- SCHMID, C.D., PRAZ, V., DELORENZI, M., PERIER, R. AND BUCHER, P. (2004). The eukaryotic promoter database, EPD: the impact of *in silico* primer extension. *Nucleic Acids Research* **32** Database issue, D82–85.
- SEGAL, E., SHAPIRA, M., REGEV, A. *ET AL.* (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* **34**, 166–176.
- SHIRAKI, T., KONDO, S., KATAYAMA, S. *ET AL.* (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.

- Proceedings of the National Academy of Sciences of the United States of America* **100**, 15776–15781.
- STORMO, G.D. AND FIELDS, D.S. (1998). Specificity, free energy and information content in protein–DNA interactions. *Trends in Biochemical Sciences* **23**, 109–113.
- SUZUKI, Y., YAMASHITA, R., NAKAI, K. AND SUGANO, S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Research* **30**, 328–331.
- SUZUKI, Y., YAMASHITA, R., SUGANO, S. AND NAKAI, K. (2004). DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Research* **32** Database issue, D78–81.
- TATUSOV, R.L., FEDOROVA, N.D., JACKSON, J.D. *ET AL.* (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.
- THOMAS, J.W., TOUCHMAN, J.W., BLAKESLEY, R.W. *ET AL.* (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793.
- THOMPSON, J.D., HIGGINS, D.G. AND GIBSON, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.
- VAN HELDEN, J., ANDRE, B. AND COLLADO-VIDES, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology* **281**, 827–842.
- WASSERMAN, W.W. AND FICKETT, J.W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. *Journal of Molecular Biology* **278**, 167–181.
- WASSERMAN, W.W., PALUMBO, M., THOMPSON, W., FICKETT, J.W. AND LAWRENCE, C.E. (2000). Human–mouse genome comparisons to locate regulatory sites. *Nature Genetics* **26**, 225–228.
- WITTKOPP, P.J., HAERUM, B.K. AND CLARK, A.G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature* **430**, 85–88.
- YU, H., LUSCOMBE, N.M., QIAN, J. AND GERSTEIN, M. (2003). Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends in Genetics* **19**, 422–427.