Chapter 11

Functional *in silico* analysis of gene regulatory polymorphism

Chaolin Zhang[1,2], Xiaoyue Zhao[1], Michael Q. Zhang[1]

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724 USA

[2]Department of Biomedical Engineering, State University of New York at Stony Brook, NY 11794 USA

Abstract

Regulatory polymorphisms play important roles in determining phenotypes and disease susceptibilities through perturbing gene expression. After genetic markers with LD have been mapped, an important task is to evaluate their functional importance based on the identification of regulatory regions and specific regulatory elements. Here we introduce the leading bioinformatics tools and invaluable resources to facilitate the functional analysis *in silico*. We focus on promoter prediction, modeling, prediction of transcription factor binding sites and identification of sequence elements important for splicing regulation. These tools are becoming increasingly accurate and powerful to guide experimental investigations and to transform correlated alleles into mechanistic understandings.

Keywords

Promoter prediction, CpG island, motif finding, TFBS prediction, splicing enhancer and repressor, ESE, ESS, splicing regulation, gene expression regulation

11.1 Introduction

Gene expression refers to the cellular processes that lead to functional products (primarily proteins) from the genetic information stored in the genomic sequences. Tightly regulated gene expression for specific cell types and developmental stages in response to different physiological conditions is driven by the orchestration of complex and multi-layered gene regulatory networks (GRNs) (Maniatis and Reed, 2002). Inferring GRNs is of fundamental importance and a great challenge for molecular biologists and geneticists.

Mutations, including point mutations, insertions and deletions, translocations, and duplications, play critical roles in determining biological phenotypes and disease susceptibilities by perturbing the GRNs. Among them, single nucleotide polymorphisms (SNPs) generated by point mutations occur approximately one per 1,000 bases and are the predominant variations among humans. The interplay between the adaptive benefits caused by mutations and natural selection shapes the genome into unique patterns of genetic variations in different regions. Therefore, investigating the functional roles of these genetic variations provides a great opportunity towards the understanding of complex common diseases, such as cancer. The compilation of human and other metazoan genome sequences (see Chapter 4 and 5) and the availability of genome-wide high resolution genotyping data (see Chapter 3) have provided an extraordinary resources for this purpose.

By either family-based linkage analysis or population-based association studies, an increasing number of genes and genomic loci have been associated with disease traits, or more recently, gene expression quantitative traits (see Chapter 17). However, because of the linkage disequilibrium (LD), it remains extremely difficult to distinguish the real causative pathogenic loci from correlated markers, which is the key step to transform genetic findings into mechanistic understanding of GRNs and effective prevention, diagnosis and treatment of diseases. Nevertheless, this provides an important starting point to identify functional polymorphisms. Functional polymorphisms can be classified into two categories: *cis*-acting regulatory polymorphisms, which disrupt or create regulatory elements in DNA or RNA sequences, and *trans*-acting polymorphisms, which alter protein structures and activities, and potentially affect many target loci (Buckland, 2006). Methods to predict the impact of coding polymorphisms on protein structures will be discussed in Chapter 12. Here, we introduce the rapidly emerging and improved bioinformatics tools which can help the analysis of regulatory polymorphisms. We emphasize the principles underlying the leading algorithms in the field to help geneticists

understand their advantages and disadvantages. Hopefully, this chapter will be a practical guide for geneticists to choose available tools and resources to facilitate their experimental studies.

Gene expression regulation can take place at any step during the path of expression, including transcription, mRNA splicing and processing, export and subcellular localization, translation, post-translational modifications, etc. These steps are often coupled with each other (Maniatis and Reed, 2002). Currently, it is still too early to build comprehensive and accurate dynamic models for truly realistic GRNs. The majority of computational methods attempt to detect *cis*-*trans* relationships, the basic building blocks of GRNs, by modern statistical or machine learning approaches. In this chapter, we will focus on finding *cis*-regulatory elements or modules (multiple collaborative elements) at the transcriptional level (DNA) and the splicing (RNA) level, with an emphasis in mammalian species. We choose these two fields because of the extensive research efforts in recent years and their representativeness. We first introduce methods for identifying the regulatory regions, such as CpG islands and promoters. Then we describe tools to pinpoint specific regulatory elements, using the analysis of transcription factor binding sites (TFBSs) as examples. Almost all of these methods can be employed to identify regulatory elements important for other regulation steps. Specific methods and resources for studying splicing regulatory elements are then given. Finally, we summarize steps of combining the genomic variation data and the prediction of regulatory elements and give examples how this approach can help infer from associated alleles to causative alleles. A selection of tools and resources are given in Table 11.1.

11.2 Predicting regulatory regions

The first step of studying regulatory polymorphisms is to determine whether the polymorphisms are located in a regulatory region or in a coding region. Different steps of regulation involve very different regulatory regions. For example, the promoter is the most important regulatory region that controls and regulates the very first step of gene expression: mRNA transcription. The signal for splicing lies in splice sites at the boundaries of exons, as well as exonic and intronic sequences flanking the splice sites. The mRNA stability and localization are usually controlled by regulatory elements in 5’ UTRs and/or 3’UTRs. For organisms like human or yeast, whose gene annotations are relatively complete, genome browsers are very useful tools to identify gene structures and other related annotations (Hinrichs et al., 2006, Birney et al., 2006). These genome browsers include both transcript supported genes and computationally predicted genes. Many other resources, including promoter databases and computational methods for promoter predictions, are also available to characterize promoters more

accurately.

11.2.1 An operational definition of promoter

Promoter is commonly referred to as the DNA region that is required for controlling and regulating the transcription initiation of the gene immediately downstream. For a typical eukaryotic (Pol II or protein-coding) gene, it contains a core promoter about 100 bp centered at the transcriptional start site (TSS) and a proximal promoter about 500 bp immediately upstream of the core promoter. For most purposes, people use the region (-500, +100) with respect to a TSS as a specific definition.

The pre-initiation complex (PIC), which comprises many general transcription factors (GTFs), assembles onto the core promoter by interacting with several core promoter elements, such as TATA-box, Inr, DPE, BRE, and DCE. The core promoter can direct transcription mediated by purified GTFs and Pol II *in vitro* at the basal level. The functional form of the PIC *in vivo* must also contain coactivators/mediators and its interactions with other TFs, which recruit the complex to the core promoter and allow for response of the polymerase to the regulatory signals. During the development, genes are turned on and off in a pre-programmed fashion, a process orchestrated by TFs, whose binding sites aggregate in the promoters near their controlled genes. A combinatorial control is achieved via different combinations of ubiquitous and cell-specific regulatory factors. Moreover, genes can initiate transcription at multiple loci (alternative promoters), which creates RNA isoforms with different 5' regions. Alternative promoters are potentially important for gene expression regulation or generating different protein products. Complex regulation *in vivo* can also involve many more features, such as enhancers, locus control regions (LCRs), and/or scaffold/matrix attachment regions (S/MARs). Enhancers are also referred as the distal promoter elements, which can be either upstream of, downstream of or within a gene and can be in any orientation. It should be noticed that there is no real distinction between proximal and distal (enhancer) regulatory elements, as they often involve the same set of TF binding sites. The cooperative binding of some TFs to enhancers and proximal promoters can lead to the assembly of nucleoprotein structures termed "enhanceosomes". For a comprehensive review on the related biology, one is referred to the excellent book by (Carey and Smale, 2000).

11.2.1 CpG islands

CpG island is an important signature of 5' regions of more than 70% mammalian genes, often overlapping with, or within 1,000 bases downstream of the promoter (Ioshikhes and Zhang, 2000) . Vertebrate genomic DNA is known to be generally depleted of the dinucleotide CpG. In the human genome, for example, the occurrence of CpG dinucleotides is five times less than statistically predicted

from the nucleotide composition (Bird, 1980). CpG depletion is believed to result from methylation of Cs at 80% CpG dinucleotides which leads to the mutation of the methylated C to T, and thus the conversion of the CpG dinucleotides to TpG. There are, however, genomic regions of high GC content, termed CpG islands, where the level of methylation is significantly lower than the overall genome. In these regions, the occurrence of CpGs is significantly higher, close to the expected frequency. As defined by Gardiner-Garden & Frommer, CpG islands are greater than 200bp in length, have more than 50% in GC content, and have a ratio of the CpG frequency to the product of the C and G frequencies above 0.6 (Gardiner-Garden and Frommer, 1987). The CpGPlot program in the EMBOSS package can be used to map CpG islands according to this definition (Larsen et al., 1992). This information is also included in the UCSC genome browser (Hinrichs et al., 2006).

11.2.2 Promoter databases and resources

One promoter resource with the best quality is the Eukaryotic Promoter Database (EPD), where transcription start sites were determined experimentally (Schmid et al., 2006). With high-throughput technologies such as 5' SAGE (Hashimoto et al., 2004) or CAGE (Carninci et al., 2005) emerging for mapping TSS, EPD starts collecting TSS from these databases with a built-in quality evaluation procedure. Currently, EPD (release 86) contains 4,809 promoters, including 2,540 vertebrate promoters and 1,871 human promoters. Database of Transcriptional Start Sites (DBTSS) is another useful source which is based on full-length oligo-capped cDNA sequences and provides alternative promoter annotations (Suzuki et al., 2004). The current release of DBTSS (5.2.0) contains 30,964 human promoters and 425,117 corresponding TSS. By clustering TSSs, they found that 8,308 human genes and 4,276 mouse genes have alternative promoters.

The Cold Spring Harbor Laboratory mammalian promoter database (CSHLmpd) (Xuan et al., 2005) is a comprehensive promoter database for human, mouse and rat. It used all known as well as predicted transcripts to construct gene sets. The corresponding promoter information was collected from multiple resources including EPD, DBTSS, GenBank and also computational predictions. They are integrated with an internal quality quantitation and control system. It enables users to extract the sequences of their specified regions around TSS, with a specified quality. Promoters of orthologous genes can be compared to detect sequence conservations in those regions.

Recent advances in ChIP-chip technology (see Section 11.3.4) provide the opportunity to study the genome-wide map of active promoters in specific cell types. Using this technology, Kim et al. experimentally located the sites of PIC binding throughout the genome in human fibroblast cells (Kim et al., 2005). Databases based on 5’SAGE, CAGE (fantom3) and oligo-capping (DBTSS) technology

also start to provide tissue information of each TSS. The accumulated tissue-specific mapping will be very useful for studying how genes are differentially expressed in different tissues.

11.2.3 Computational promoter prediction

Despite the availability of experimentally validated promoter resources, computational prediction algorithms are still of great importance in the identification and characterization of novel genes, as well as large scale annotations of many other species after genome sequencing. There have been extensive efforts to improve promoter predictions computationally (see (Werner, 2003) and references therein). The primary goal of these programs is to identify TSS and/or core promoter elements for all (protein coding) genes in a genome, in contrast to the programs identifying specific transcription factor binding sites (TFBSs) that are shared by a particular set of co-regulated genes (see Section 11.3). The underlying principle of these programs is that promoter regions have some distinctive and characteristic features different from non-promoters. A classifier is trained on experimentally validated promoters/TSSs (obtained from databases such as EPD or DBTSS), and then used to scan novel genomic sequences. Different programs differ in the features and classification algorithms used.

Features important for computational promoter prediction programs include GC content, CpG ratio, TFBS density, word compositions and core promoter elements. These have been modeled by many programs. For example, PROMOTERSCAN (Prestridge, 1995) and AUTOGENE (Kondrakhin et al., 1995) are two of the earliest programs which utilizes different densities of TFBSs in promoters and non-promoter sequences, together with a TATA-matrix score. Due to the very limited number of TFs with known binding motifs (see Section 11.3.2), short sequences (words) more abundant in promoters compared to non-promoter regions have been employed for prediction. This idea, with some variations, has been implemented in PromFind (Hutchinson, 1996), CorePromoter (Zhang, 1998) and PromoterInspector (Scherf et al., 2000). CpG_Promoter is an effective algorithm discriminating the promoter-associated CpG islands from the non promoter-associated ones (Ioshikhes and Zhang, 2000). It uses three features to train a quadratic discriminant classifier: length, GC content and the CpGratio (observed / expected). This algorithm only aims for a promoter region, not the exact locations of TSSs. Other methods, especially those developed in recent years, try to integrate as much information as possible to improve the accuracy of promoter prediction. Here we introduce a few representative ones.

11.2.3.1 More comprehensive modeling of TFBSs

TSSG and TSSW (Solovyev and Salamov, 1997) both use LDA (linear discriminant analysis) to combine (a) a TATA score, (b) triplet preferences around TSS, (c) hexamer score in three non-

overlapping windows of 100 bp upstream TSS, and (d) putative binding site scores. A program called Eponine (Down and Hubbard, 2002) models the preferential spacing between binding sites of each TF and TSS as well as the over-representation of the binding sites. Over-represented binding sites with conserved spacing receive high scores and are recovered de novo using a relevant vector machine. They found that TATA box and the flanking region with GC enrichment are the most important signals. A linear combination of binding site scores is then used for prediction.

11.2.3.2 Physical properties

Regulatory regions often exhibit distinct physical properties such as DNA flexibility and GC content in their sequences. McPromoter integrates such structural features into a neural network in conjunction with the Markov modeling of the sequence information from different segments (upstream, core promoter and downstream) and is able to reduce false positives(Ohler et al., 2001).

11.2.3.3 Cross-species conservation

It was observed that some major promoter components such as TSS, TATA and regulatory motifs are significantly more conserved than the sequences around them. A recent program PromH (Solovyev and Shahmuradov, 2003) uses linear discriminant functions that take into account the conservation features and nucleotide sequences of promoter regions in pairs of orthologous genes. To use PromH, othologous sequences must be provided. It should be noted that they are not always available due to the difficulty to align orthologous sequences, especially for distal species.

11.2.3.4 CpG related vs Non-CpG related

It is computationally useful and also biologically meaningful to treat CpG related promoters and non-CpG related ones separately. Non-CpG related promoters are more heterogeneous and therefore more difficult for computational prediction. Two programs explicitly build different promoter models for these two classes. In FirstEF (Davuluri et al., 2001), CpG related promoters and non-CpG related ones are modeled separately, each using three quadratic discriminant functions to recognize structural and compositional features of promoter regions, first exons and first splice-donor sites in conjunction. All these functions are then incorporated into a decision tree. The predictions of the first exons and promoter regions in the human genome are available in the UCSC genome browser. Another program called Dragon Promoter Finder (Dragon PF) (Bajic et al., 2002) uses sensors for three functional regions: promoters, exons and introns, and then combines them via artificial neural networks (ANNs) for GC-rich and GC-poor sequences, respectively. Each sensor is based on the frequencies of

pentamers at each position. Dragon Gene Start Finder (Dragon GSF) combines Dragon PF and the prediction of presence of CpG islands using ANN(Bajic and Seah, 2003).

11.2.3.5 Performance evaluation

Promoter prediction has been a difficult problem in gene finding and characterization. Choosing appropriate programs is very important since the types of information built into the different models are not completely the same. This is further complicated by the lack of benchmark data for training and evaluation during original publication. A most recent review compared eight programs for whole human genome predictions (Bajic et al., 2004). According to their comparison, Dragon GSF and FirstEF might be the good choices to start with for general promoter predictions. Approximately, they can predict more than half of promoters correctly at the cost of one or a few false predictions for each correct one, at the resolution of several hundred to 2kb. They are quite successful in locating the transcription start sites for CpG related promoters, but the performance for non-CpG related ones is less satisfactory due to the diverse nature of vertebrate promoter sequences. Although improving, current programs are still insufficient to pinpoint TSSs, therefore difficult to distinguish alternative promoters.

11.3 Modeling and predicting transcription factor binding sites

Promoter prediction and TFBS identification are not closely related. While promoter prediction is to locate the beginning and *cis*-regulatory regions of a gene, the focus of computational methods modeling and predicting TFBSs is to understand *cis-trans* interactions for transcription regulation. TFBSs are short (about 6-20 bp in length), usually degenerate, and often found in promoters. Both experimental and computational methods have been developed to identify TFBSs with different throughputs and at different resolutions. There are two general problems in the computational studies of TFBSs. First, with a set of sequences (e.g. promoters) believed to be co-regulated, statistical methods are used to identify the pattern of the binding sites (motif) for regulators. Second, given the motif of a specific TF, computational methods are used to scan for putative binding sites of that factor. Note that almost all methods described here are applicable to other types of protein-DNA/RNA interaction, such as the regulation of mRNA splicing and stability.

11.3.1 Motif representation: consensus or matrix

The binding sequence of a TF allows a certain degree of variations, which create a spectrum of binding affinity. The variations of binding sites can be collected from known target genes, mutagensis studies

(Hallikas et al., 2006), phylogenetic shadowing (orthologous binding sites in different species), (Ostrin et al., 2006), and *in vitro* SELEX experiments(Liu and Stormo, 2005). Several recent technologies, such as SELEX-SAGE (Roulet et al., 2002) and protein-binding microarray (PBM) (Mukherjee et al., 2004) allow for the determination of binding specificity in a high throughput manner.

The profile or motif of binding sites can then be described with a consensus sequence. By aligning the sites, the base(s) with the largest affinity (or the most frequent base among known binding sites) at each position is chosen as a representative. For example, the consensus of E. coli TATA-box can be written as TATAAT or TATRNT using the IUPAC code allowing degeneracy. This representation is straightforward and useful when the motif is relatively long and conserved. However, TF motifs of higher eukaryotes are generally degenerate. Consensus can not quantitatively reflect the binding affinities of sites and thus is not optimal for predicting the occurrence of new sites. In most applications, a position weight matrix (PWM) is a better choice.

To maintain the binding affinity, point mutations inside binding sites must be constrained. This requirement, which connects energetic constraints and base frequencies, forms the foundation for statistical mechanic motif modeling (Berg and von Hippel, 1987). When the non-random base composition of the background or a control set (e.g. the whole genome) is taken into account, the relation can be expressed as follows

$$s_{B,j} = \ln(p_{B,j}/p_{B,0}) \tag{11.1}$$

where $s_{B,j}$ is the score (PWM element) for a base B at position j (j=1,2,…,J, J is the length of the motif), $p_{B,j}$ is the frequency of the base B at position j and $p_{B,0}$ indicates the background base frequency which does not depend on position. Choosing $p_{B,0}$ appropriately (representing the correct background contrast) can be very important for searching new binding sites in a genome. Under the assumption that the binding affinity of each position is independent and additive, the total score of the site is the sum of individual position scores in all J positions

$$s = \sum_{j=1}^{J} s_{B_j,j}\,, \tag{11.2}$$

where B_j is the base at position j. A straightforward interpretation of the score is the log likelihood ratio of being a site to being a non-site. Similarly, the log likelihood ratio of observing K sites can be represented by

$$S = \sum_k s_k = K \sum_{j=1}^{J} \sum_{B=A}^{T} p_{B,j} \log\left(p_{B,j} / p_{B,0} \right). \tag{11.3}$$

Here $I = \sum_{j=1}^{J} \sum_{B=A}^{T} p_{B,j} \log\left(p_{B,j} / p_{B,0} \right)$ is the information content motif and represent the level of degeneracy.

The PWM motif model can be visualized by a "pictogram" or motif logo (Crooks et al., 2004). In these visualizations, each position is a stack of letters, reflecting the frequency of observing each nucleotide. The total height of each position can be scaled according to the information content of that position.

The PWM representation can be generalized to more complex models, such as high order Markov models (Roulet et al., 2002, Zhang and Marr, 1993), the maximum entropy model (Yeo and Burge, 2004), Bayesian networks (Barash et al., 2003) and generalizations of Bayesian networks (Ben-Gal et al., 2005, Zhao et al., 2005), when more binding sites are available. However, in most cases, the simpler PWM model is sufficient.

11.3.2 De novo motif finding

There are approximately two thousand of TFs in human and likely in other mammals (Messina et al., 2004, Kummerfeld and Teichmann, 2006). TF motifs determined experimentally have been collected into databases such as SCPD (Zhu and Zhang, 1999), TRANSFAC (Matys et al., 2003) and JASPAR(Sandelin et al., 2004), which however contain limited data. For example, there are currently around 600 vertebrate motifs in TRANSFAC, among which many are derived from only a few known sites. In order to discover novel motifs, one has to resort to de novo motif finding algorithms. Given a set of related sequences as described in Section 11.3.4, these algorithms attempt to find the most over-represented patterns of short sequences in a reasonable time. Numerous algorithms have been proposed in the past decade. These algorithms differ in motif representation, objective function and the procedure for optimization. More importantly, they incorporate different data or prior knowledge in the modeling and therefore can fit for various situations. In the following we introduce computational motif finding algorithms according to these considerations, rather than technical details. Interested users are also referred to a recent review comparing 13 popular methods (Tompa et al., 2005).

11.3.2.1 Finding the most over-represented motifs

Most of earlier approaches attempt to identify motifs with the most over-represented binding sites. These algorithms achieve the goal by optimizing the local sequence alignment using Equation 11.3 or its variation as an objective function. Since neither the motif nor the binding sites are known, this optimization is a combinatorial problem, which needs heuristic searching strategies to get a reasonably good solution in a feasible time. Representative heuristic strategies include greedy search, implemented in CONSENSUS (Hertz and Stormo, 1999), expectation maximization (EM), implemented in MEME (Bailey and Elkan, 1994), and Gibbs sampler (Lawrence et al., 1993). In the latter two approaches, the

motif model and site locations are optimized iteratively by pretending either the motif or the sites are known at the beginning. After the best motif is recovered, the sites are erased to identify the second best motifs and so on. A few other programs, such as MDScan (Liu et al., 2002) and Weeder (Pavesi et al., 2001), start from searching for over-represented consensus (allowing degeneracy), rather than from the matrix search directly.

Additional features included in these programs, as well as in their variations such as AlignACE (Roth et al., 1998), Bioprospector (Liu et al., 2001) and the Improbizer (http://www.cse.ucsc.edu/~kent/improbizer/), make them smarter. For example, MEME allow user to specify whether every sequence has one or multiple binding sites. MEME, CONSENSUM and Weeder is able to optimize motif length automatically. MEME and Bioprospector can limit the search for only two block motifs or palindromic motifs. Higher order markov models have been used in Biospector and MDScan to characterize background sequences more accurately, by which a significant improvement has been observed.

Note that repeat sequences should be masked before motif finding. Also, it might always be worth trying multiple (similar) programs to see whether a motif is detected consistently. It is reported that some programs are complementary to each other, and are thus able to improve specificity when used in combination (Tompa et al., 2005). The identified matrices can be compared with matrices of known TFs to identify putative regulators (Schones et al., 2005).

11.3.2.2 Finding discriminative motifs

In contrast to most motif finding algorithms, discriminative motif finding attempts to identify the best motifs which discriminate two sets of sequences. This is extremely useful in studying gene expression regulation, e.g. to explain different responses of two groups of genes after a stimuli or to distinguish genes expressed in different tissues. The most discriminative motif is not necessarily the most abundant. Master regulators, such as *E2F* family members and *myc*, have thousands of binding sites over the human genome (Cawley et al., 2004), whereas some TFs may regulate only a handful of targets. Therefore, discriminative but subtle signals might be overwhelmed by common binding sites without careful modeling. The word-counting algorithms compare the relative enrichment of each word (may allow degeneracy) between the foreground sequences and background sequences. This approach is very effective in identifying short and less degenerate motifs, such as many typical TF sites in yeast (Zhang, 1999). Other programs of this category, with slightly different scoring functions, include WORDUP (Pesole et al., 1992), DMOTIFS (Sinha, 2003) and DWE (Sumazin et al., 2005). To detect more degenerate motifs, one has to use the matrix model. The statistical framework used to find over-

represented motif (Liu et al., 1995) can be easily extended to model the relative over-representation, as shown in a recent work (Smith et al., 2005). In addition, the implemented program called discriminative matrix enumerator (DME) "exhaustively" searches discrete spaces of matrices followed by a local optimization step. This method is very successful in identifying tissue specific motifs which can be highly degenerate (Smith et al., 2006, Smith et al., 2005).

11.3.2.3 Finding conserved motifs

Sequence conservation across different species is an important indicator of functionality. Phylogenetic footprinting is referred to as the identification of functional regions by comparing orthologous genomic sequences between species (Fickett and Wasserman, 2000). With more sequenced genomes available, comparative analysis of noncoding regions has become an important approach for detecting promoters or regulatory regions in general (Bejerano et al., 2004, Siepel et al., 2005). Several earlier methods for detecting conserved blocks from a multiple alignment have been evaluated by (Stojanovic et al., 1999). Programs designed for very long alignments of syntenic regions have also become available (see (Blanchette et al., 2004) and references therein). More general information of comparative genomics is available in Chapter 6.

With the alignment of multiple orthologous sequences, it is possible to detect short TF motifs which are significantly more conserved than random. This idea has been applied to screen regulatory elements conserved in multiple yeast species (Kellis et al., 2003) and recently in four mammalian species (Xie et al., 2005). The motifs identified included many known ones as well as novel ones. Given a set of presumably co-regulated sequences and their orthologs, it is also possible to incorporate both over-representation and conservation into motif finding algorithms. A straightforward strategy is to use a two-step procedure: find conserved regions and then search for over-represented motifs only in those regions (Wasserman et al., 2000). The two steps can be applied in the opposite order: first over-represented motifs are identified separately in each species or in the pooled data; then motifs without significant conservation are eliminated (GuhaThakurta et al., 2002, Pritsker et al., 2004, Li et al., 2005). It was argued that these methods are somewhat ad hoc and may miss over-represented but divergent motifs or conserved motifs not very over-represented. Therefore, the two criteria can also be integrated to a single statistical framework for optimization (Li and Wong, 2005, Prakash et al., 2004). However, since more parameters need to be estimated from an often small dataset, these methods may also identify noisy motifs (Li et al., 2005).

It should also be noted that the conservation of regulatory regions may vary widely. In principle, the regulatory programs that control early development in metazoan systems tend to be extremely

complex, almost always involving distal enhancers and/or complicated locus control regions (LCRs). Subtle change of these programs can lead to dramatic effects. Therefore, lineage developmental master TFs and their binding sites are often more conserved. In contrast, in the terminally differentiated tissues, the regulatory program is often relatively simple, *cis*-regulatory regions in the promoters tend to be closer to TSS and many TFBS are less conserved among distant species.

11.3.2.4 Constructing cis-regulatory modules

Since genes are always regulated by multiple TFs and composite binding sites (*cis*-regulatory modules or CRMs), simultaneous detection of CRMs rather than individual sites may provide better specificity. A module can be composed of multiple sites of the same type (homotypic) or different types (heterotypic). Palindromic motifs can be regarded as a special type of CRMs and are common for TFs. However, most CRMs studies require that individual motifs are known. De novo CRM discovery is a much more difficult problem, which usually needs larger data set (Zhou and Wong, 2004, Gupta and Liu, 2005, Bussemaker et al., 2001). The CisModule algorithm has been applied to identify CRMs important for muscle-specific expression in Ciona savignyi (Johnson et al., 2005). REDUCE (Bussemaker et al., 2001), MotifRegressor (Conlon et al., 2003), MARSMotif (Das et al., 2004), and more recently MatrixREDUCE (Foat et al., 2005) and MARSMotif-M (Das et al., 2006) are regression-based algorithms which can maximize the explained variation of gene expression by a limited number of motifs in combination.

11.3.3 Predicting novel binding sites

Given a motif determined experimentally or computationally, an important task is to search new sequences for novel binding sites using consensus matching or matrix scoring. However, one must first assess the quality of the motif and determine a threshold before using it for searching new sites. One way to do this is to perform a standard classification test by which both the threshold score and the motif length may be optimized by minimizing the classification (Bayesian) error. The MATCH program (Kel et al., 2003) included in the TRANSFAC database uses pre-calculated thresholds with different stringencies. The storm program in the CREAD package is tailored to search a set of sequences for multiple motifs very quickly(Smith et al., 2006).

For most TF motifs, due to the short length and degeneracy, the signal to noise ratio is quite low. Therefore, predicting novel functional binding sites of a known motif is by no means easier than de novo motif finding (Hu et al., 2005). It was estimated that for a typical motif, without any other information except the matrix, the specificity of genomic search can be as low as 0.001, which means

one functional sites among 1000 predictions (Wasserman and Sandelin, 2004). These false predictions, which might bind TFs with high affinity *in vitro*, are never used *in vivo*, suggesting that important signals also reside outside the cognate binding sites to distinguish from decoy sites. These include CRMs, chromatin structure, and DNA stability and flexibility, which are important for determining the affinity and accessibility of the binding sites. Conservation information in other species is not accessible for the cellular machinery, but is effective for eliminating false positives by one order of magnitude (Wasserman and Sandelin, 2004).

One way for predicting CRMs is via evaluating the significance of the co-occurrence of TFBSs within a certain distance. This approach requires the least prior knowledge. Claverie and Sauvaget published one of the earliest methods for detecting two sites in a fixed distance and the same orientation in the heat-shock promoters (Claverie and Sauvaget, 1985). Alternatively, more subtle rules can be learned from known functional sites co-occurring in the same regions. Although still limited, several databases such as COMPEL and TRRD have started collecting experimentally validated CRMs (Heinemeyer et al., 1998), which will greatly facilitate the advances in this field. An interesting example was given by the identification of regulatory modules that confer muscle-specific gene expression (Wasserman and Fickett, 1998), where logistic regression was used to combine matrix scores for multiple sites. The authors reported that focusing on CRMs rather than individual binding sites can reduce false positives by two orders of magnitude while retain more than half of the true sites. In a recent study, not only co-occurrences, but also geometric constraints, were modeled quantitatively from known examples. The implemented program called EEL identified vertebrate enhancers successfully (Hallikas et al., 2006). Therefore, the analysis of CRMs is able to improve prediction accuracy to a level which makes follow-up experimental investigations feasible.

Other functional annotations and co-localization information are also helpful. Computational approaches incorporating DNA mechanical properties and nucleosome structures are still rare, but represent important directions in the future.

11.3.4 Experimental approaches for identifying co-regulated targets

Two types of high throughput technologies, namely, microarrays and genomic occupancy assays, are highly effective to identify co-regulated genes, thereby narrowing down the putative regions of functional binding sites dramatically. These technologies become routinely used for gene expression regulation studies.

There have been many studies based on expression microarrays. For the purpose of regulation studies, it becomes more powerful when data are collected under multiple conditions or at multiple

time-points after the perturbation of the upstream TF (e.g. TF knockout, mutation in DNA binding domain or knock-down). The underlying assumption is that genes with similar expression profiles (co-expression) are likely to be regulated by the same factor(s) (co-regulation). Classical approaches are based on the clustering analysis to identify genes with correlated expression patterns, from which one could try to identify *cis*-elements enriched in their promoters (Spellman et al., 1998, Hughes et al., 2000). Some algorithms are specifically designed for motif finding by looking for "tight clusters" of expression profiles (Tseng and Wong, 2005).

However, co-expression is not equal to co-regulation. When the perturbation is on some master regulators, it can also activate/repress many downstream TFs, as in the case of heat shock or other stress responses. Because multiple TFs are involved, responsive target genes would be a mixture of direct targets for different TFs. ChIP-chip(Cawley et al., 2004, Carroll et al., 2005, Odom et al., 2004, Lee et al., 2002) and ChIP-tag technologies (Sabo et al., 2004, Impey et al., 2004, Ng et al., 2005, Wei et al., 2006) allow for more direct detections of genomic regions occupied by endogenous transcription factors. ChIP-chip cross-links bound proteins to chromatins *in vivo*. Immunoprecipitated DNA fragments are then hybridized to genomic DNA microarrays or sequenced using the SAGE-tag technology. The power of these approaches has been demonstrated in many applications. Despite non-specific binding and cross-linking, it has been shown that highly enriched chip-regions are very accurate in predicting bona fide targets. In a study of ER binding sites in chromosome 21 and 22 using chip-chip data, all 57 predictions were validated to be real (Carroll et al., 2005). Due to the current resolution of 500-2000 bp, computational analysis for motif finding and binding site prediction is indispensable. Almost all the motif finding and TFBS prediction tools described in Section 11.3.2 and 11.3.3 can be applied to chromatin occupancy data.

11.4 Predicting regulatory elements for splicing regulation

The next level of gene expression regulation is RNA processing, including capping, splicing, polyadenylation, editing, stability and transport. Many of these (in particular, the first three) steps are co-transcriptional and hence coupled to transcriptional regulation (Maniatis and Reed, 2002). In this section, we will only focus on mRNA splicing, especially alternative splicing (AS) that is responsible for generating diverse protein isoforms from a single gene locus. Recent estimates of alternatively spliced genes are more than 60% in human and likely other mammals (Lander et al., 2001, Modrek and Lee, 2002, Johnson et al., 2003). Alternative splicing plays critical roles in many regulatory pathways in metazoans, including those controlling cell growth, cell death, differentiation and development

(Black, 2003). Aberrant splicing has been implicated in a large number of human diseases (Faustino and Cooper, 2003).

The boundaries of introns and exons are marked by splice sites. The canonical splice sites are composed of GU dinucleotide in the exon/intron boundary (5'ss or donor site), and AG dinucleotide in the intron/exon boundary (3'ss or acceptor site). Each dinucleotide is flanked by a larger, less conserved sequence. The branch site and polypyrimidine tract close to the 3'ss in the intron are also critical for splicing. Minor types of splice sites (e.g. AU/AC introns), although less than 0.1%, also exist (Burset et al., 2000). Although the key biochemical steps of splicing have been worked out, far less is known about the mechanism of accurate splicing regulation. In mammal, the signal carried by the splice sites is insufficient to drive specific exon recognition. The tight regulation of alternative splice site selection in response to different physiological conditions is mediated though the interactions of numerous *cis*-elements outside the splice sites, such as enhancers and repressors, and a very large protein/snRNA complex, called splicesome, which is composed of hundreds of proteins (Rappsilber et al., 2002, Zhou et al., 2002) and five critical snRNAs. Furthermore, the splicing studies rely heavily on *in vitro* systems (mini-genes). It is difficult to generate mini-gene construct, reproducing the same splicing pattern *in vivo*. Previous computational analyses mainly focused on the detection of AS events and evolutionary properties based on cDNA/EST data. Methods for facilitating the understanding of the splicing regulation are emerging in recent years.

11.4.1 Statistical modeling and prediction of splice sites

The motif of splice sites can also be represented by consensus sequences or PWMs, as described in Section 11.3.1. However, this is perhaps the most appropriate place to test complex models because a very large number of known splice sites are available by transcript-genome alignment. These methods include higher order Markov model (Zhang and Marr, 1993), maximum entropy model (Yeo and Burge, 2004) and Bayesian networks (Chen et al., 2005), all of which attempted to model the dependencies among different positions. It was claimed that integrating correlations between nucleotides helps discriminate authentic splice sites from pseudo splice sites.

11.4.2 Identification of splicing enhancers and silencers

Cis-elements for splicing regulation can be in exons or introns, and can be enhancers or repressors. They are important for both constitutive splicing and alternative splicing (Smith and Valcarcel, 2000). The best-characterized enhancers and silencers are recognized by one of two SF classes: hnRNPs (heterogeneous nuclear ribonucleoprotein) and SR (serine/arginine rich) proteins. They are usually

identified in tissue-specific exons or disease mutants with aberrant splicing. Many ESEs are purine rich. A well-studied example is the 73 nt exonic splicing enhancer (ESE) in the alternative exon M2 of the mouse IgM gene. This ESE can stimulate the splicing when inserted into a heterologous regulated intron of Drosophila doublesex (dsx) gene (see references in (Liu et al., 1998)). In another example, a single nucleotide C/T silent transition causes the skipping of exon 7 in the human *SMN2* gene, which mediates the severity of spinal muscular astropy (SMA) in the absence of the wild-type *SMN1* gene, a paralog of *SMN2* (Lorson et al., 1999). It was demonstrated that the transition disrupts an SF2/ASF dependent ESE and creates an ESS bound by hnRNP A1 (Cartegni and Krainer, 2002, Kashima and Manley, 2003). Comprehensive lists of exonic and intronic splicing regulatory sequences reported in literature have been compiled (Ladd and Cooper, 2002, Zheng, 2004). Since elements recognized by a splicing factor seem to be very degenerate, it is difficult to derive a motif from these known examples.

11.4.2.1 SR-protein binding sites and ESEfinder

The SR proteins are a family of highly conserved serine/arginine-rich RNA-binding proteins. They are essential splicing factors with overlapping functions, involved in early steps of spliceosome assembly They can regulate the selection of alternative splice sites in a concentration-dependent manner, in part by antagonizing the activity of hnRNP A1 (see references in (Liu et al., 1998)). The binding sites of four SR proteins, including SF2/ASF, SC35, SRp40 and SRp55, have been determined by the Krainer lab using a functional SELEX assay (Liu et al., 2000, Liu et al., 1998). The ESE matrix for each SR protein was then derived from the winner sequences as described in Section 11.3.1. These matrices have been included in a web-based resource called ESEfinder (Cartegni et al., 2003) which can be used for predicting and visualizing novel ESEs of these SR proteins.

11.4.2.2 Exonic splicing silencers

A systematic screening for exonic splicing silencers has been performed by the Burge lab (Wang et al., 2004). The principle of the system shares similarity with that of SELEX in spirit except that (i) the screening is GFP-based and (ii) random sequences are not selected for a specific splicing factor, but can be any element with repressive activities. This screening identified 141 ESS decamers. Most of these are likely repressive when introduced into heterogeneous gene contexts. These decamers can be clustered according to sequence similarity to identify consensus motifs. Although some of them resemble the binding sites of known SFs, such as hnRNP A1 and HnRNP H, generally it is not clear which SFs can specifically recognize these sequences.

11.4.2.3 Enhancers and silencers derived in silico

Pure computational screening for enhancers and silencers has also been performed (Fairbrother et al., 2002, Zhang and Chasin, 2004, Sironi et al., 2004, Yeo et al., 2004). Due to the difficulty to obtain a list of co-regulated AS events, these studies do not focus on specific SFs. Instead, these studies assume different densities of regulatory elements in different genic regions and attempt to identify general enhancing or repressive elements. In particular, in the RESCUE approach, Fairbrother et al assumed that for an exon to be constitutively included, weak splice sites have to be complemented by a higher density of ESEs nearby. They also assumed that the density of ESEs in exonic regions is higher than that in intronic regions. All hexamers were scored using these two criteria of relative overrepresentation. As a result, 238 hexamers were identified as potential human ESEs (called RESCUE-ESEs), and then were clustered into 10 motifs based on their sequence similarity. Some of these motifs resemble the binding sites of known SFs. When inserted into the test exon of a minigene construct, these hexamers can indeed enhance exon inclusion. This approach has been extended to mouse, zebrafish and fugu (Yeo et al., 2004) for predicting intronic splicing enhancers (ISEs). Among hundreds of the predicted ISEs hexamers, the GGG motif is the most prevalent and contained in majority of ISEs hexamers. The predicted ESEs were included in the RESCUE-ESE server and can be used to scan new sequences (Fairbrother et al., 2004).

Since ESEs are largely imposed on coding constraints, Zhang and Chasin argued that the codon usage bias might complicate the ESE identification (Zhang and Chasin, 2004). Therefore, in their study, only constitutively spliced, internal, non-coding exons were used. They assumed that ESEs should have a relative enrichment in these spliced non-coding exons compared to 5' UTRs of intronless genes and pseudoexons (intronic regions flanked by splice-site like sequences with a similar length as real exons). Using this approach, they identified 2069 octomers as putative ESEs (PESEs) and 974 octomers as putative ESSs (PESSs). An online tool is also available for identifying the occurrences of these PESEs and PESSs.

It should be noted that the *trans*-acting factors interacting with these enhancers and silencers predicted *in silico* are not obvious, although their splicing role has been demonstrated in *in vitro* splicing assays and endogenous genes (Zhang et al., 2005). Since these elements were originally identified using constitutively spliced exons, although they are likely important for alternative splicing as well, it is not very clear how much bias might be introduced. The constitutive and alternative splicing may result from the different balances of the same positive and negative splicing signals. Moreover, at least for several known SFs, their specific elements are only associated with tissue-specifically spliced exons. These elements are likely missed by these *in silico* predictions.

The enhancers and silencers derived by different approaches have also been compared for their ability to predict splicing alterations caused by point mutations. Interestingly, although these methods (SELEX, RESCUE ESEs, PESEs) are comparable in predictive power, the overlap is moderate (Zhang et al., 2005, Wang et al., 2005). This might imply that a number of ESEs, as well as ESSs, have not been identified.

11.4.3 Splicing microarrays

A catalogue of regulatory elements described above is only the first step towards the understanding of splicing regulation. A more challenging step is to understand how the interaction of these regulatory elements and splicing factors generates highly regulated splicing patterns in different tissues types or under different conditions in response to stimuli. The combinatorial interaction of multiple factors may contribute greatly to the subtle regulation. The variation of expression of splicing factors may also add another layer of complexity. Currently, the detailed mechanistic studies of splicing regulation are limited to only a few model systems using minigene constructs. One concern is whether the rules inferred from these models can reflect the regulation *in vivo* and whether they are general enough to extend to other genes. Like in the study of transcriptional regulation, the high throughput technologies measuring splicing activities and protein-RNA interactions under specific conditions can provide invaluable information.

The feasibility of using microarrays to study the regulation of RNA splicing was first demonstrated in yeast (Clark et al., 2002). These splicing microarrays are designed to be capable of distinguishing splicing variants by probes in exon bodies and exon junctions(Modrek and Lee, 2002). It was demonstrated that the loss of key mRNA processing factors leads to dramatic splicing defects, which can be measured by microarrays. Since 40-60% of the mammalian genes have introns (a typical gene has about 8 introns) comparing to 3.8% intron-containing genes in yeast, detecting AS in a mammalian system with microarrays has only become possible very recently (Sugnet et al., 2006, Johnson et al., 2003, Pan et al., 2004, Li et al., 2006). The largest study so far measured AS in more than 50 human tissues and cell lines, using 36nt oligonucleotide probes tiled on every consecutive exon junctions of Refseq genes (Johnson et al., 2003). Since probes are included even for “constitutive” exon junctions, where there is no cDNA/EST evidence of alternative splicing, this platform can identify novel AS events in a more unbiased manner, compared to the EST sequencing approach. However, it should be noted that these data are noisy since each AS event is represented by only one or two probes (in affymetrix arrays, more than 10 probes are used to summarize the mRNA abundance level). As there are very limited choices for probe positions in exon junctions to optimize hybridization efficiency

and specificity, a few number of probes might behave poorly and thus do not reflect the correct abundance of the splicing junctions. In the recent affymetrix exon microarrays, multiple probes are tiled on each exon to get more reliable signals of exon inclusion. However, there are no junction probes and it is difficult to infer splicing patterns. In several other studies, arrays are designed for the AS events with transcript evidence(Pan et al., 2004, Sugnet et al., 2006, Ule et al., 2005). In these designs, each AS event is represented by multiple probes in exons, introns or exon junctions. Therefore, they can measure AS more accurately and are suitable to infer rules of splicing regulation. For example, motifs have been discovered from the flanking intronic regions of brain/muscle specific AS exons (Sugnet et al., 2006). Other microarray-based assays are also developed. For example, in DASL, a high specificity is achieved by the ligation of a pair of oligos across the splice junction. The primer extension step before ligation makes the choice of probes more flexible (Fan et al., 2004). This approach has been applied to screen a panel of prostate cancer tissues and normal tissues to identify signature splicing events (Li et al., 2006) and compared with conventional microarrays (Zhang et al., 2006). The combination of splicing microarrays and knock-down experiments of splicing factors have been demonstrated as a powerful tool to dissect important pathways regulated through tissue specific splicing (Ule et al., 2005). Obviously, splicing microarrays will be routinely used to study GRNs in the coming years.

Several technical difficulties should be noted. As transcription and splicing are intrinsically coupled, it is very difficult to separate their individual contributions to the steady state levels of the transcripts despite several attempts(Shai et al., 2006, Li et al., 2006, Johnson et al., 2003, Cline et al., 2005). Also, the alterations detected by microarrays are individual AS events, which are local. Usually, the complete isoforms and protein products cannot be tracked unambiguously (Wang et al., 2003). Another challenge, even combined with knock-out experiments, is to distinguish direct effects and indirect effects as the direct measurement of protein-RNA interaction in a large scale is still in its infancy(Ule et al., 2003).

11.5 Evaluating the functional importance of regulatory polymorphisms

The functional importance of polymorphisms is determined by how they can affect the regulation of gene expression regulation. In the most common scenario for geneticists, a polymorphism, e.g. a SNP, is linked to a disease or gene expression trait and one has to ask whether it is a causative allele or just an allele with LD. The first step is to determine the regulatory region it falls in, the basis to choose appropriate tools. For polymorphisms in promoters, it would be a strong indication of disrupting

transcription if they overlap with a transcription factor binding sites. Putative binding sites can be predicted as described in Section 11.3. Although the false positive rate of such *in silico* prediction for a single binding site is usually high, accuracy can be improved by incorporating information from different resources.

- If a binding site overlaps with the SNP, do the different alleles have different binding affinities (motif scores)? In a study of 127 SNPs in the promoters of cell-cycle check point genes, a majority of them potentially affect binding affinity, which were validated at a high success rate by Gel shift *in vitro* (Belanger et al., 2005). A further step is to ask whether the identified binding site co-occur with others, which can potentially form a CRM.
- Is the SNP in the core promoter? According to the report assays which evaluated the effect of 674 haplotypes in 247 promoters for promoter activity, there is a strong bias that functional polymorphisms are close to the TSS (Buckland et al., 2005). This is likely due to the fact that the density of regulatory elements is higher near the TSS. Also, subtle changes in mechanical or geometric properties of DNA may also alter the efficiency of transcription (Buckland, 2006) even the SNP does not change a binding site directly.
- Is the overlapped region conserved? UCSC genome browser is an excellent resource for interactive analysis of cross species conservation. See (Bejerano et al., 2005) for a step-by-step tutorial on how to screen conserved regions for functional elements.
- Is there any *in vivo* binding evidence? Besides the known binding sites collected in different databases, it is also worth checking whether a genome-wide assay for chromatin occupancy has been performed for the TF under study.
- Is there any functional annotation (such as gene ontology, tissue specific expression, protein interaction, etc) for the associated gene? Is it involved in the pathway implicated in the disease?

The *in silico* analyses are economical and fast, often can provide good evidence about the potential importance. The candidates which pass these filtering steps may be validated with a reporter assay before further investigation. If the SNP is located in the exonic region or flanking intronic region, it might alter mRNA splicing.

- Disruption of splice sites is very suggestive of aberrant splicing.
- Otherwise, one needs to examine whether it disrupts or creates splicing enhancers or silencers. Successful examples have been demonstrated to identify point mutations affecting ESEs using ESEfinder (Cartegni and Krainer, 2002, Liu et al., 2001). It is relatively easy to predict erroneous exon skipping than other splicing errors (e.g. cryptic splicing) as shown in these examples.

- Can splicing alteration lead to dramatic change of protein product? Skipping of an internal coding exon whose length is not a multiple of three or a non-sense mutation generally induces mRNA non-sense mediated decay (NMD) or a large truncation at 3' part of the protein. If the exon overlaps with important domain, skipping of the exon can also have dramatic effects. Different isoforms can be virtually translated into different proteins, which can then be analyzed for their protein structures (see Chapter 12).

Several tools have been developed to partially automate these analyses (Conde et al., 2004, Xu et al., 2005).

Acknowledgement

MQZ lab is supported by grants from NIH, NSF and CSHL Associations.

References

BAILEY, T. & ELKAN, C. (1994) Fitting A Mixture Model By Expectation Maximization To Discover Motifs In Biopolymers *the Second International Conference on Intelligent Systems for Molecular Biology.* Menlo Park, California, AAAI Press.

BAILEY, T. L. & GRIBSKOV, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics,* 14**,** 48-54.

BAJIC, V. B. & SEAH, S. H. (2003) Dragon Gene Start Finder: An Advanced System for Finding Approximate Locations of the Start of Gene Transcriptional Units. *Genome Res.,* 13**,** 1923-1929.

BAJIC, V. B., SEAH, S. H., CHONG, A., ZHANG, G., KOH, J. L. Y. & BRUSIC, V. (2002) Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. *Bioinformatics,* 18**,** 198-199.

BAJIC, V. B., TAN, S. L., SUZUKI, Y. & SUGANO, S. (2004) Promoter prediction analysis on the whole human genome. *Nat Biotech,* 22**,** 1467-1473.

BARASH, Y., KAPLAN, T., FRIEDMAN, N. & ELIDAN, G. (2003) Modeling dependencies in protein-DNA binding sites. *Proceedings of the 7th International Conference on Research in Computational Molecular Biology (RECOMB)***,** 28-37.

BEJERANO, G., PHEASANT, M., MAKUNIN, I., STEPHEN, S., KENT, W. J., MATTICK, J. S., et al. (2004) Ultraconserved Elements in the Human Genome. *Science,* 304**,** 1321-1325.

BEJERANO, G., SIEPEL, A. C., KENT, W. J. & HAUSSLER, D. (2005) Computational screening of conserved genomic DNA in search of functional noncoding elements. *Nat Meth,* 2**,** 535-545.

BELANGER, H., BEAULIEU, P., MOREAU, C., LABUDA, D., HUDSON, T. J. & SINNETT, D. (2005) Functional promoter SNPs in cell cycle checkpoint genes. *Hum. Mol. Genet.,* 14**,** 2641-2648.

BEN-GAL, I., SHANI, A., GOHR, A., GRAU, J., ARVIV, S., SHMILOVICI, A., et al. (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics,* 21**,** 2657-2666.

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**,** 723-743.

BIRD, A. P. (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.,* 8**,** 1499-1504.

BIRNEY, E., ANDREWS, D., CACCAMO, M., CHEN, Y., CLARKE, L., COATES, G., et al. (2006) Ensembl 2006. *Nucl. Acids Res.,* 34**,** D556-561.

BLACK, D. L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.,* 72**,** 291-336.

BLANCHETTE, M., KENT, W. J., RIEMER, C., ELNITSKI, L., SMIT, A. F. A., ROSKIN, K. M., et al. (2004) Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Res.,* 14**,** 708-715.

BUCKLAND, P. R. (2006) The importance and identification of regulatory polymorphisms and their mechanisms of action. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease,* 1762**,** 17-28.

BUCKLAND, P. R., HOOGENDOORN, B., COLEMAN, S. L., GUY, C. A., SMITH, S. K. & O'DONOVAN, M. C. (2005) Strong bias in the location of functional promoter polymorphisms. *Human Mutation,* 26**,** 214-223.

BURSET, M., SELEDTSOV, I. A. & SOLOVYEV, V. V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucl. Acids Res.,* 28**,** 4364-4375.

BUSSEMAKER, H. J., LI, H. & SIGGIA, E. D. (2001) Regulatory element detection using correlation with expression. *Nat Genet,* 27**,** 167-174.

CAREY, M. & SMALE, S. (2000) *Transcriptional regulation in eukaryotes: concepts, strategies, and techniques,* Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press.
CARNINCI, P., KASUKAWA, T., KATAYAMA, S., GOUGH, J., FRITH, M. C., MAEDA, N., et al. (2005) The Transcriptional Landscape of the Mammalian Genome. *Science,* 309**,** 1559-1563.
CARROLL, J. S., LIU, X. S., BRODSKY, A. S., LI, W., MEYER, C. A., SZARY, A. J., et al. (2005) Chromosome-Wide Mapping of Estrogen Receptor Binding Reveals Long-Range Regulation Requiring the Forkhead Protein FoxA1. *Cell,* 122**,** 33-43.
CARTEGNI, L. & KRAINER, A. R. (2002) Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. 30**,** 377-384.
CARTEGNI, L., WANG, J., ZHU, Z., ZHANG, M. Q. & KRAINER, A. R. (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucl. Acids Res.,* 31**,** 3568-3571.
CAWLEY, S., BEKIRANOV, S., NG, H. H., KAPRANOV, P., SEKINGER, E. A., KAMPA, D., et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell,* 116**,** 499-509.
CHEN, T.-M., LU, C.-C. & LI, W.-H. (2005) Prediction of splice sites with dependency graphs and their expanded bayesian networks. *Bioinformatics,* 21**,** 471-482.
CLARK, T. A., SUGNET, C. W. & ARES, M., JR. (2002) Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays. *Science,* 296**,** 907-910.
CLAVERIE, J.-M. & SAUVAGET, I. (1985) Assessing the biological significance of primary structure consensus patterns using sequence databanks. I. Heat-shock and glucocorticoid control elements in eukaryotic promoters. *Comput. Appl. Biosci.,* 1**,** 95-104.
CLINE, M. S., BLUME, J., CAWLEY, S., CLARK, T. A., HU, J.-S., LU, G., et al. (2005) ANOSVA: a statistical method for detecting splice variation from expression data. *Bioinformatics,* 21**,** i107-115.
CONDE, L., VAQUERIZAS, J. M., SANTOYO, J., AL-SHAHROUR, F., RUIZ-LLORENTE, S., ROBLEDO, M., et al. (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucl. Acids Res.,* 32**,** W242-248.
CONLON, E. M., LIU, X. S., LIEB, J. D. & LIU, J. S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *PNAS,* 100**,** 3339-3344.
CROOKS, G. E., HON, G., CHANDONIA, J.-M. & BRENNER, S. E. (2004) WebLogo: A Sequence Logo Generator. *Genome Res.,* 14**,** 1188-1190.
DAS, D., BANERJEE, N. & ZHANG, M. Q. (2004) Interacting models of cooperative gene regulation. *PNAS,* 101**,** 16234-16239.
DAS, D., NAHLE, Z. & ZHANG, M. (2006) Adaptively Inferring Human Transcriptional Subnetworks. *Mol. System Biol.,* In press.
DAVULURI, R. V., GROSSE, I. & ZHANG, M. Q. (2001) Computational identification of promoters and first exons in the human genome. *Nat Genet,* 29**,** 412-417.
DOWN, T. A. & HUBBARD, T. J. P. (2002) Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA. *Genome Res.,* 12**,** 458-461.
FAIRBROTHER, W. G., YEH, R.-F., SHARP, P. A. & BURGE, C. B. (2002) Predictive Identification of Exonic Splicing Enhancers in Human Genes. *Science,* 297**,** 1007-1013.
FAIRBROTHER, W. G., YEO, G. W., YEH, R., GOLDSTEIN, P., MAWSON, M., SHARP, P. A., et al. (2004) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucl. Acids Res.,* 32**,** W187-190.
FAN, J.-B., YEAKLEY, J. M., BIBIKOVA, M., CHUDIN, E., WICKHAM, E., CHEN, J., et al. (2004) A Versatile Assay for High-Throughput Gene Expression Profiling on Universal Array Matrices. *Genome Res.,* 14**,** 878-885.
FAUSTINO, N. A. & COOPER, T. A. (2003) Pre-mRNA splicing and human disease. *Genes Dev.,* 17**,** 419-437.

FICKETT, J. W. & WASSERMAN, W. W. (2000) Discovery and modeling of transcriptional regulatory regions. *Current Opinion in Biotechnology,* 11**,** 19-24.

FOAT, B. C., HOUSHMANDI, S. S., OLIVAS, W. M. & BUSSEMAKER, H. J. (2005) Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *PNAS,* 102**,** 17675-17680.

GARDINER-GARDEN, M. & FROMMER, M. (1987) CpG Islands in vertebrate genomes. *Journal of Molecular Biology,* 196**,** 261-282.

GUHATHAKURTA, D., PALOMAR, L., STORMO, G. D., TEDESCO, P., JOHNSON, T. E., WALKER, D. W., et al. (2002) Identification of a Novel cis-Regulatory Element Involved in the Heat Shock Response in Caenorhabditis elegans Using Microarray Gene Expression and Computational Methods. *Genome Res.,* 12**,** 701-712.

GUPTA, M. & LIU, J. S. (2005) De novo cis-regulatory module elicitation for eukaryotic genomes. *PNAS,* 102**,** 7079-7084.

HALLIKAS, O., PALIN, K., SINJUSHINA, N., RAUTIAINEN, R., PARTANEN, J., UKKONEN, E., et al. (2006) Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity. *Cell,* 124**,** 47-59.

HASHIMOTO, S.-I., SUZUKI, Y., KASAI, Y., MOROHOSHI, K., YAMADA, T., SESE, J., et al. (2004) 5[prime]-end SAGE for the analysis of transcriptional start sites. *Nat Biotech,* 22**,** 1146-1149.

HEINEMEYER, T., WINGENDER, E., REUTER, I., HERMJAKOB, H., KEL, A. E., KEL, O. V., et al. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucl. Acids Res.,* 26**,** 362-367.

HERTZ, G. Z. & STORMO, G. D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics,* 15**,** 563-577.

HINRICHS, A. S., KAROLCHIK, D., BAERTSCH, R., BARBER, G. P., BEJERANO, G., CLAWSON, H., et al. (2006) The UCSC Genome Browser Database: update 2006. *Nucl. Acids Res.,* 34**,** D590-598.

HU, J., LI, B. & KIHARA, D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucl. Acids Res.,* 33**,** 4899-4913.

HUGHES, J. D., ESTEP, P. W., TAVAZOIE, S. & CHURCH, G. M. (2000) Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology,* 296**,** 1205-1214.

HUTCHINSON, G. B. (1996) The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput. Appl. Biosci.,* 12**,** 391-398.

IMPEY, S., MCCORKLE, S. R., CHA-MOLSTAD, H., DWYER, J. M., YOCHUM, G. S., BOSS, J. M., et al. (2004) Defining the CREB regulon: A genome-wide analysis of transcription factor regulatory regions. *Cell,* 119**,** 1041-1054.

IOSHIKHES, I. P. & ZHANG, M. Q. (2000) Large-scale human promoter mapping using CpG islands. *Nat Genet,* 26**,** 61-63.

JOHNSON, D. S., ZHOU, Q., YAGI, K., SATOH, N., WONG, W. & SIDOW, A. (2005) De novo discovery of a tissue-specific gene regulatory module in a chordate. *Genome Res.,* 15**,** 1315-1324.

JOHNSON, J. M., CASTLE, J., GARRETT-ENGELE, P., KAN, Z., LOERCH, P. M., ARMOUR, C. D., et al. (2003) Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. *Science,* 302**,** 2141-2144.

KASHIMA, T. & MANLEY, J. L. (2003) A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. 34**,** 460-463.

KEL, A. E., GOSSLING, E., REUTER, I., CHEREMUSHKIN, E., KEL-MARGOULIS, O. V. & WINGENDER, E. (2003) MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucl. Acids Res.,* 31**,** 3576-3579.

KELLIS, M., PATTERSON, N., ENDRIZZI, M., BIRREN, B. & LANDER, E. S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature,* 423**,** 241-254.

KIM, T. H., BARRERA, L. O., ZHENG, M., QU, C., SINGER, M. A., RICHMOND, T. A., et al. (2005) A high-resolution map of active promoters in the human genome. *Nature,* 436**,** 876-880.

KONDRAKHIN, Y. V., KEL, A. E., KOLCHANOV, N. A., ROMASHCHENKO, A. G. & MILANESI, L. (1995) Eukaryotic promoter recognition by binding sites for transcription factors. *Comput. Appl. Biosci.,* 11**,** 477-488.

KUMMERFELD, S. K. & TEICHMANN, S. A. (2006) DBD: a transcription factor prediction database. *Nucl. Acids Res.,* 34**,** D74-81.

LADD, A. & COOPER, T. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biology,* 3**,** reviews0008.1 - reviews0008.16.

LANDER, E., LINTON, L., BIRREN, B., NUSBAUM, C., ZODY, M., BALDWIN, J., et al. (2001) Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LARSEN, F., GUNDERSEN, G., LOPEZ, R. & PRYDZ, H. (1992) CpG islands as gene markers in the human genome. *Genomics,* 13**,** 1095-1107.

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.

LEE, T. I., RINALDI, N. J., ROBERT, F., ODOM, D. T., BAR-JOSEPH, Z., GERBER, G. K., et al. (2002) Transcriptional Regulatory Networks in Saccharomyces cerevisiae. *Science,* 298**,** 799-804.

LI, H.-R., WANG-RODRIGUEZ, J., NAIR, T. M., YEAKLEY, J. M., KWON, Y.-S., BIBIKOVA, M., et al. (2006) Two-Dimensional Transcriptome Profiling: Identification of Messenger RNA Isoform Signatures in Prostate Cancer from Archived Paraffin-Embedded Cancer Specimens. *Cancer Res,* 66**,** 4079-4088.

LI, X. & WONG, W. H. (2005) Sampling motifs on phylogenetic trees. *PNAS,* 102**,** 9481-9486.

LI, X., ZHONG, S. & WONG, W. H. (2005) Reliable prediction of transcription factor binding sites by phylogenetic verification. *PNAS,* 102**,** 16945-16950.

LIU, H.-X., CHEW, S. L., CARTEGNI, L., ZHANG, M. Q. & KRAINER, A. R. (2000) Exonic Splicing Enhancer Motif Recognized by Human SC35 under Splicing Conditions. *Mol. Cell. Biol.,* 20**,** 1063-1071.

LIU, H.-X., ZHANG, M. & KRAINER, A. R. (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev.,* 12**,** 1998-2012.

LIU, J. & STORMO, G. D. (2005) Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions. *Nucl. Acids Res.,* 33**,** e141-.

LIU, J. S., 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.

LIU, X., BRUTLAG, D. & LIU, J. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput***,** 127-138.

LIU, X. S., BRUTLAG, D. L. & LIU, J. S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotech,* 20**,** 835-839.

LORSON, C. L., HAHNEN, E., ANDROPHY, E. J. & WIRTH, B. (1999) A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *PNAS,* 96**,** 6307-6311.

MAGLOTT, D., OSTELL, J., PRUITT, K. D. & TATUSOVA, T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucl. Acids Res.,* 33**,** D54-58.

MANIATIS, T. & REED, R. (2002) An extensive network of coupling among gene expression machines. *Nature,* 416**,** 499-506.

MATYS, V., FRICKE, E., GEFFERS, R., GOSSLING, E., HAUBROCK, M., HEHL, R., et al. (2003) TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucl. Acids Res.,* 31**,** 374-378.

MESSINA, D. N., GLASSCOCK, J., GISH, W. & LOVETT, M. (2004) An ORFeome-based Analysis of Human Transcription Factor Genes and the Construction of a Microarray to Interrogate Their Expression. *Genome Res.,* 14**,** 2041-2047.

MODREK, B. & LEE, C. (2002) A genomic view of alternative splicing. *Nat Genet,* 30**,** 13-19.

MUKHERJEE, S., BERGER, M. F., JONA, G., WANG, X. S., MUZZEY, D., SNYDER, M., et al. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet,* 36**,** 1331-1339.

NG, P., WEI, C.-L., SUNG, W.-K., CHIU, K. P., LIPOVICH, L., ANG, C. C., et al. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Meth,* 2**,** 105-111.

ODOM, D. T., ZIZLSPERGER, N., GORDON, D. B., BELL, G. W., RINALDI, N. J., MURRAY, H. L., et al. (2004) Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. *Science,* 303**,** 1378-1381.

OHLER, U., NIEMANN, H., LIAO, G.-C. & RUBIN, G. M. (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics,* 17**,** S199-206.

OSTRIN, E. J., LI, Y., HOFFMAN, K., LIU, J., WANG, K., ZHANG, L., et al. (2006) Genome-wide identification of direct targets of the Drosophila retinal determination protein Eyeless. *Genome Res.,* 16**,** 466-476.

PAN, Q., SHAI, O., MISQUITTA, C., ZHANG, W., SALTZMAN, A. L., MOHAMMAD, N., et al. (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Molecular Cell,* 16**,** 929-941.

PAVESI, G., MAURI, G. & PESOLE, G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics,* 17**,** S207-214.

PESOLE, G., PRUNELLA, N., LIUNI, S., ATTIMONELLI, M. & SACCONE, C. (1992) WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. *Nucl. Acids Res.,* 20**,** 2871-2875.

PRAKASH, A., BLANCHETTE, M., SINHA, S. & TOMPA, M. (2004) Motif Discovery in Heterogeneous Sequence Data *Pacific Symposium on Biocomputing,* 9.

PRESTRIDGE, D. S. (1995) Predicting Pol II Promoter Sequences using Transcription Factor Binding Sites. *Journal of Molecular Biology,* 249**,** 923-932.

PRITSKER, M., LIU, Y.-C., BEER, M. A. & TAVAZOIE, S. (2004) Whole-Genome Discovery of Transcription Factor Binding Sites by Network-Level Conservation. *Genome Res.,* 14**,** 99-108.

RAPPSILBER, J., RYDER, U., LAMOND, A. I. & MANN, M. (2002) Large-Scale Proteomic Analysis of the Human Spliceosome. *Genome Res.,* 12**,** 1231-1245.

REESE, M., EECKMAN, F., KULP, D. & HAUSSLER, D. (1997) Improved splice site detection in Genie. *J Comput Biol,* 4**,** 311-323.

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. *Nat Biotech,* 16**,** 939-945.

ROULET, E., BUSSO, S., CAMARGO, A. A., SIMPSON, A. J. G., MERMOD, N. & BUCHER, P. (2002) High-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotech,* 20**,** 831-835.

SABO, P. J., HAWRYLYCZ, M., WALLACE, J. C., HUMBERT, R., YU, M., SHAFER, A., et al. (2004) Discovery of functional noncoding elements by digital analysis of chromatin structure. *PNAS,* 101**,** 16837-16842.

SANDELIN, A., ALKEMA, W., ENGSTROM, P., WASSERMAN, W. W. & LENHARD, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucl. Acids Res.,* 32**,** D91-94.

SCHERF, M., KLINGENHOFF, A. & WERNER, T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *Journal of Molecular Biology,* 297**,** 599-606.

SCHMID, C. D., PERIER, R., PRAZ, V. & BUCHER, P. (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucl. Acids Res.,* 34**,** D82-85.

SCHONES, D. E., SUMAZIN, P. & ZHANG, M. Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics,* 21**,** 307-313.

SHAI, O., MORRIS, Q. D., BLENCOWE, B. J. & FREY, B. J. (2006) Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics***,** btk028.

SIEPEL, A., BEJERANO, G., PEDERSEN, J. S., HINRICHS, A. S., HOU, M., ROSENBLOOM, K., et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.,* 15**,** 1034-1050.

SINHA, S. (2003) Discriminative Motifs. *Journal of Computational Biology,* 10**,** 599-615.

SIRONI, M., MENOZZI, G., RIVA, L., CAGLIANI, R., COMI, G. P., BRESOLIN, N., et al. (2004) Silencer elements as possible inhibitors of pseudoexon splicing. *Nucl. Acids Res.,* 32**,** 1783-1791.

SMITH, A. D., SUMAZIN, P., XUAN, Z. & ZHANG, M. Q. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *PNAS,* 103**,** 6275-6280.

SMITH, A. D., SUMAZIN, P. & ZHANG, M. Q. (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc. Natl. Acad. Sci. U.S.A.,* 102**,** 1560-1565.

SMITH, C. W. J. & VALCARCEL, J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends in Biochemical Sciences,* 25**,** 381-388.

SOLOVYEV, V. & SALAMOV, A. (1997) The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Proc Int Conf Intell Syst Mol Biol,* 5**,** 294-302.

SOLOVYEV, V. & SHAHMURADOV, I. (2003) PromH: promoters identification using orthologous genomic sequences. *Nucl. Acids Res.,* 31**,** 3540-3545.

SPELLMAN, P. T., SHERLOCK, G., ZHANG, M. Q., IYER, V. R., ANDERS, K., EISEN, M. B., et al. (1998) Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol. Biol. Cell,* 9**,** 3273-3297.

STOJANOVIC, N., FLOREA, L., RIEMER, C., GUMUCIO, D., SLIGHTOM, J., GOODMAN, M., et al. (1999) Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucl. Acids Res.,* 27**,** 3899-3910.

SUGNET, C. W., SRINIVASAN, K., CLARK, T. A., BRIEN, G., CLINE, M. S., WANG, H., et al. (2006) Unusual Intron Conservation near Tissue-Regulated Exons Found by Splicing Microarrays. *PLoS Computational Biology,* 2**,** e4.

SUMAZIN, P., CHEN, G., HATA, N., SMITH, A. D., ZHANG, T. & ZHANG, M. Q. (2005) DWE: Discriminating Word Enumerator. *Bioinformatics,* 21**,** 31-38.
SUZUKI, Y., YAMASHITA, R., SUGANO, S. & NAKAI, K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucl. Acids Res.,* 32**,** D78-81.
THANARAJ, T. A., STAMM, S., CLARK, F., RIETHOVEN, J.-J., LE TEXIER, V. & MUILU, J. (2004) ASD: the Alternative Splicing Database. *Nucl. Acids Res.,* 32**,** D64-69.
TOMPA, M., LI, N., BAILEY, T. L., CHURCH, G. M., DE MOOR, B., ESKIN, E., et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotech,* 23**,** 137-144.
TSENG, G. C. & WONG, W. H. (2005) Tight Clustering: A Resampling-based Approach for Identifying Stable and Tight Patterns in Data. *Biometrics,* 61**,** 10-16.
ULE, J., JENSEN, K. B., RUGGIU, M., MELE, A., ULE, A. & DARNELL, R. B. (2003) CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science,* 302**,** 1212-1215.
ULE, J., ULE, A., SPENCER, J., WILLIAMS, A., HU, J.-S., CLINE, M., et al. (2005) Nova regulates brain-specific splicing to shape the synapse. *Nat Genet,* 37**,** 844-852.
WANG, H., HUBBELL, E., HU, J.-S., MEI, G., CLINE, M., LU, G., et al. (2003) Gene structure-based splice variant deconvolution using a microarry platform. *Bioinformatics,* 19**,** i315-322.
WANG, J., SMITH, P. J., KRAINER, A. R. & ZHANG, M. Q. (2005) Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucl. Acids Res.,* 33**,** 5053-5062.
WANG, Z. F., ROLISH, M. E., YEO, G., TUNG, V., MAWSON, M. & BURGE, C. B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell,* 119**,** 831-845.
WASSERMAN, W. W. & 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. & LAWRENCE, C. E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat Genet,* 26**,** 225-228.
WASSERMAN, W. W. & SANDELIN, A. (2004) APPLIED BIOINFORMATICS FOR THE IDENTIFICATION OF REGULATORY ELEMENTS. *Nature Reviews Genetics,* 5**,** 276-287.
WEI, C.-L., WU, Q., VEGA, V. B., CHIU, K. P., NG, P., ZHANG, T., et al. (2006) A Global Map of p53 Transcription-Factor Binding Sites in the Human Genome. *Cell,* 124**,** 207-219.
WERNER, T. (2003) The state of the art of mammalian promoter recognition. *Brief Bioinform,* 4**,** 22-30.
XIE, X., LU, J., KULBOKAS, E. J., GOLUB, T. R., MOOTHA, V., LINDBLAD-TOH, K., et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3[prime] UTRs by comparison of several mammals. *Nature,* 434**,** 338-345.
XU, H., GREGORY, S. G., HAUSER, E. R., STENGER, J. E., PERICAK-VANCE, M. A., VANCE, J. M., et al. (2005) SNPselector: a web tool for selecting SNPs for genetic association studies. *Bioinformatics,* 21**,** 4181-4186.
XUAN, Z., ZHAO, F., WANG, J., CHEN, G. & ZHANG, M. (2005) Genome-wide promoter extraction and analysis in human, mouse, and rat. *Genome Biol.,* 6**,** R72.
YEO, G. & BURGE, C. B. (2004) Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *Journal of Computational Biology,* 11**,** 377-394.
YEO, G., HOON, S., VENKATESH, B. & BURGE, C. B. (2004) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *PNAS,* 101**,** 15700-15705.
ZHANG, C., LI, H.-R., FAN, J.-B., WANG-RODRIGUEZ, J., DOWNS, T., FU, X.-D., et al. (2006) Profiling alternatively spliced mRNA isoforms for prostate cancer classification. *BMC Bioinformatics,* In press.
ZHANG, M. (1999) Promoter analysis of co-regulated genes in the yeast genome. *Comp. & Chem.,* 23**,** 233-250.

ZHANG, M. Q. (1998) Identification of Human Gene Core Promoters in silico. *Genome Res.,* 8**,** 319-326.
ZHANG, M. Q. & MARR, T. G. (1993) A weight array method for splicing signal analysis. *CAABIOS.*
ZHANG, X. H.-F. & CHASIN, L. A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.,* 18**,** 1241-1250.
ZHANG, X. H.-F., KANGSAMAKSIN, T., CHAO, M. S. P., BANERJEE, J. K. & CHASIN, L. A. (2005) Exon Inclusion Is Dependent on Predictable Exonic Splicing Enhancers. *Mol. Cell. Biol.,* 25**,** 7323-7332.
ZHAO, X., HUANG, H. & SPEED, T. P. (2005) Finding Short DNA Motifs Using Permuted Markov Models. *Journal of Computational Biology,* 12**,** 894-906.
ZHENG, Z.-M. (2004) Regulation of Alternative RNA Splicing by Exon Definition and Exon Sequences in Viral and Mammalian Gene Expression. *J Biomed Sci,* 11**,** 278-294.
ZHOU, Q. & WONG, W. H. (2004) CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. *PNAS,* 101**,** 12114-12119.
ZHOU, Z., LICKLIDER, L. J., GYGI, S. P. & REED, R. (2002) Comprehensive proteomic analysis of the human spliceosome. *Nature,* 419**,** 182-185.
ZHU, J. & ZHANG, M. Q. (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics,* 15**,** 607-611.

Tables

Table 11.1 Resources related to the analysis of gene regulatory sequences

Resource Name	URL	References
Genome browsers and gene structure analysis		
UCSC genome browser	http://genome.ucsc.edu	(Hinrichs et al., 2006)
Ensembl	http://www.ensembl.org	(Birney et al., 2006)
Entrez gene	http://www.ncbi.nlm.nih.gov	(Maglott et al., 2005)
Promoter databases and resources		
EPD	http://www.epd.isb-sib.ch	(Schmid et al., 2006)
DBTSS	http://dbtss.hgc.jp	(Suzuki et al., 2004)
CSHLmpd	http://rulai.cshl.edu/CSHLmpd2	(Xuan et al., 2005)
CpGPlot	http://www.sanger.ac.uk/Software/EMBOSS	Larsen et al. 1992
Epionine	http://www.sanger.ac.uk/Users/td2/eponine	(Down and Hubbard, 2002)
McPromoter	http://genes.mit.edu/McPromoter.html	(Ohler et al., 2001)
Dragon PF & GSF	http://research.i2r.a-star.edu.sg/promoter	(Bajic et al., 2002)
FirstEF	http://rulai.cshl.edu/tools/FirstEF	(Davuluri et al., 2001)
Transcription factor binding site databases		
TRANSFAC®	http://www.biobase.de/pages/index.php?id=111	(Matys et al., 2003)
JASPAR	http://jaspar.cgb.ki.se	(Sandelin et al., 2004)
De novo motif finding		
CONSENSUS	http://bifrost.wustl.edu/consensus	(Hertz and Stormo, 1999)
MEME	http://meme.sdsc.edu/meme	(Bailey and Elkan, 1994)
AlignACE	http://atlas.med.harvard.edu	(Roth et al., 1998)
MDScan	http://ai.stanford.edu/~xsliu/MDscan	(Liu et al., 2002)
DWE	http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=analysisMotifDWEForm	(Sumazin et al., 2005)
DME	http://rulai.cshl.edu/software/index1.htm	(Smith et al., 2005)
CisModule	http://www.people.fas.harvard.edu/~qingzhou/CisModScan/index.html	(Zhou and Wong, 2004)
Novel TFBS prediction		
MATCH™	http://www.biobase.de/pages/index.php?id=291	(Kel et al., 2003)
Storm (in CREAD)	http://rulai.cshl.edu/cread	(Smith et al., 2006)
MAST	http://meme.sdsc.edu/meme/mast.html	(Bailey and Gribskov, 1998)
CisModuleScan	http://www.people.fas.harvard.edu/~qingzhou/CisModScan/index.html	(Zhou and Wong, 2004)
Splice site prediction		
ASD	http://www.ebi.ac.uk/asd-srv/wb.cgi	(Thanaraj et al., 2004)
MaxEntScan	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html	(Yeo and Burge, 2004)
Splice Site Prediction by Neural Network	http://www.fruitfly.org/seq_tools/splice.html	(Reese et al., 1997)
Enhancer and repressor prediction		
ESEfinder	http://rulai.cshl.edu/tools/ESE	(Cartegni et al., 2003)
RESCUE-ESE	http://genes.mit.edu/burgelab/rescue-ese	(Fairbrother et al., 2004)
PESX	http://cubweb.biology.columbia.edu/pesx	(Zhang and Chasin, 2004)
Regulatory SNP analysis		
PupaSNP Finder	http://pupasuite.bioinfo.cipf.es	(Conde et al., 2004)
SNPselector	http://primer.duhs.duke.edu	(Xu et al., 2005)