CHAPTER NINE

Promoter and Regulatory Element Prediction

An issue related to gene prediction is promoter prediction. Promoters are DNA elements located in the vicinity of gene start sites (which should not be confused with the translation start sites) and serve as binding sites for the gene transcription machinery, consisting of RNA polymerases and transcription factors. Therefore, these DNA elements directly regulate gene expression. Promoters and regulatory elements are traditionally determined by experimental analysis. The process is extremely time consuming and laborious. Computational prediction of promoters and regulatory elements is especially promising because it has the potential to replace a great deal of extensive experimental analysis.

However, computational identification of promoters and regulatory elements is also a very difficult task, for several reasons. First, promoters and regulatory elements are not clearly defined and are highly diverse. Each gene seems to have a unique combination of sets of regulatory motifs that determine its unique temporal and spatial expression. There is currently a lack of sufficient understanding of all the necessary regulatory elements for transcription. Second, the promoters and regulatory elements cannot be translated into protein sequences to increase the sensitivity for their detection. Third, promoter and regulatory sites to be predicted are normally short (six to eight nucleotides) and can be found in essentially any sequence by random chance, thus resulting in high rates of false positives associated with theoretical predictions.

Current solutions for providing preliminary identification of these elements are to combine a multitude of features and use sophisticated algorithms that give either ab initio–based predictions or predictions based on evolutionary information or experimental data. These computational approaches are described in detail in this chapter following a brief introduction to the structures of promoters and regulatory elements in both prokaryotes and eukaryotes.

PROMOTER AND REGULATORY ELEMENTS IN PROKARYOTES

In bacteria, transcription is initiated by RNA polymerase, which is a multi-subunit enzyme. The σ subunit (e.g., σ^{70}) of the RNA polymerase is the protein that recognizes specific sequences upstream of a gene and allows the rest of the enzyme complex to bind. The upstream sequence where the σ protein binds constitutes the promoter sequence. This includes the sequence segments located 35 and 10 base pairs (bp) upstream from the transcription start site. They are also referred to as the -35 and -10 boxes. For the σ^{70} subunit in *Escherichia coli*, for example, the -35 box

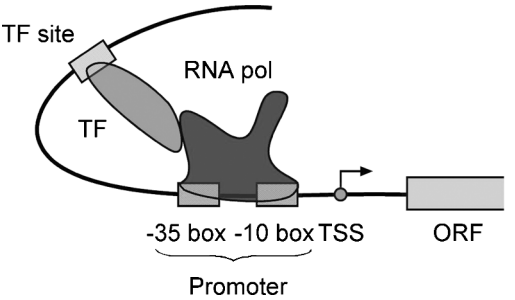

Figure 9.1: Schematic representation of elements involved in bacterial transcription initiation. RNA polymerase binds to the promoter region, which initiates transcription through interaction with transcription factors binding at different sites. *Abbreviations:* TSS, transcription start site; ORF, reading frame; pol, polymerase; TF, transcription factor (see color plate section).

has a consensus sequence of TTGACA. The –10 box has a consensus of TATAAT. The promoter sequence may determine the expression of one gene or a number of linked genes downstream. In the latter case, the linked genes form an operon, which is controlled by the promoter.

In addition to the RNA polymerase, there are also a number of DNA-binding proteins that facilitate the process of transcription. These proteins are called *transcription factors*. They bind to specific DNA sequences to either enhance or inhibit the function of the RNA polymerase. The specific DNA sequences to which the transcription factors bind are referred to as *regulatory elements*. The regulatory elements may bind in the vicinity of the promoter or bind to a site several hundred bases away from the promoter. The reason that the regulatory proteins binding at long distance can still exert their effect is because of the flexible structure of DNA, which is able to bend and and exert its effect by bringing the transcription factors in close contact with the RNA polymerase complex (Fig. 9.1).

PROMOTER AND REGULATORY ELEMENTS IN EUKARYOTES

In eukaryotes, gene expression is also regulated by a protein complex formed between transcription factors and RNA polymerase. However, eukaryotic transcription has an added layer of complexity in that there are three different types of RNA polymerase complexes, namely RNA polymerases I, II, and III. Each polymerase transcribes different sets of genes. RNA polymerases I and III are responsible for the transcription of ribosomal RNAs and tRNAs, respectively. RNA polymerase II is exclusively responsible for transcribing protein-encoding genes (or synthesis of mRNAs).

Unlike in prokaryotes, where genes often form an operon with a shared promoter, each eukaryotic gene has its own promoter. The eukaryotic transcription machinery also requires many more transcription factors than its prokaryotic counterpart to help initiate transcription. Furthermore, eukaryotic RNA polymerase II does not directly bind to the promoter, but relies on a dozen or more transcription factors to recognize and bind to the promoter in a specific order before its own binding around the promoter.

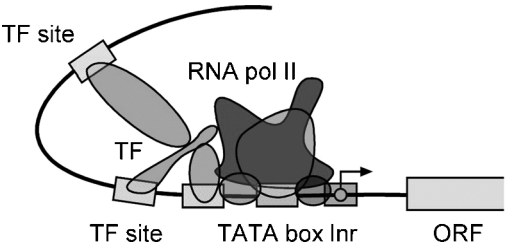

Figure 9.2: Schematic diagram of an eukaryotic promoter with transcription factors and RNA polymerase bound to the promoter. *Abbreviations:* Inr, initiator sequence; ORF, reading frame; pol, polymerase; TF, transcription factor (see color plate section).

The core of many eukaryotic promoters is a so-called TATA box, located 30 bps upstream from the transcription start site, having a consensus motif TATA(A/T)A(A/T) (Fig. 9.2.). However, not all eukaryotic promoters contain the TATA box. Many genes such as housekeeping genes do not have the TATA box in their promoters. Still, the TATA box is often used as an indicator of the presence of a promoter. In addition, many genes have a unique initiator sequence (Inr), which is a pyrimidine-rich sequence with a consensus (C/T)(C/T)CA(C/T)(C/T). This site coincides with the transcription start site. Most of the transcription factor binding sites are located within 500 bp upstream of the transcription start site. Some regulatory sites can be found tens of thousands base pairs away from the gene start site. Occasionally, regulatory elements are located downstream instead of upstream of the transcription start site. Often, a cluster of transcription factor binding sites spread within a wide range to work synergistically to enhance transcription initiation.

PREDICTION ALGORITHMS

Current algorithms for predicting promoters and regulatory elements can be categorized as either ab initio based, which make de novo predictions by scanning individual sequences; or similarity based, which make predictions based on alignment of homologous sequences; or expression profile based using profiles constructed from a number of coexpressed gene sequences from the same organism. The similarity type of prediction is also called phylogenetic footprinting. As mentioned, because RNA polymerase II transcribes the eukaryotic mRNA genes, most algorithms are thus focused on prediction of the RNA polymerase II promoter and associated regulatory elements. Each of the categories is discussed in detail next.

Ab Initio–Based Algorithms

This type of algorithm predicts prokaryotic and eukaryotic promoters and regulatory elements based on characteristic sequences patterns for promoters and regulatory elements. Some ab initio programs are signal based, relying on characteristic promoter sequences such as the TATA box, whereas others rely on content information such as

hexamer frequencies. The advantage of the ab initio method is that the sequence can be applied as such without having to obtain experimental information. The limitation is the need for training, which makes the prediction programs species specific. In addition, this type of method has a difficulty in discovering new, unknown motifs.

The conventional approach to detecting a promoter or regulatory site is through matching a consensus sequence pattern represented by regular expressions (see Chapter 7) or matching a position-specific scoring matrix (PSSM; see Chapter 6) constructed from well-characterized binding sites. In either case, the consensus sequences or the matrices are relatively short, covering 6 to 10 bases. As described in Chapter 7, to determine whether a query sequence matches a weight matrix, the sequence is scanned through the matrix. Scores of matches and mismatches at all matrix positions are summed up to give a log odds score, which is then evaluated for statistical significance. This simple approach, however, often has difficulty differentiating true promoters from random sequence matches and generates high rates of false positives as a result.

To better discriminate true motifs from background noise, a new generation of algorithms has been developed that take into account the higher order correlation of multiple subtle features by using discriminant functions, neural networks, or hidden Markov models (HMMs) that are capable of incorporating more neighboring sequence information. To further improve the specificity of prediction, some algorithms selectively exclude coding regions and focus on the upstream regions (0.5 to 2.0 kb) only, which are most likely to contain promoters. In that sense, promoter prediction and gene prediction are coupled.

Prediction for Prokaryotes

One of the unique aspects in prokaryotic promoter prediction is the determination of operon structures, because genes within an operon share a common promoter located upstream of the first gene of the operon. Thus, operon prediction is the key in prokaryotic promoter prediction. Once an operon structure is known, only the first gene is predicted for the presence of a promoter and regulatory elements, whereas other genes in the operon do not possess such DNA elements.

There are a number of methods available for prokaryotic operon prediction. The most accurate is a set of simple rules developed by Wang et al. (2004). This method relies on two kinds of information: gene orientation and intergenic distances of a pair of genes of interest and conserved linkage of the genes based on comparative genomic analysis. More about gene linkage patterns across genomes is introduced in Chapters 16 and 18. A scoring scheme is developed to assign operons with different levels of confidence (Fig. 9.3). This method is claimed to produce accurate identification of an operon structure, which in turn facilitates the promoter prediction.

This newly developed scoring approach is, however, not yet available as a computer program. The prediction can be done manually using the rules, however. The few dedicated programs for prokaryotic promoter prediction do not apply the Wang et al. rule for historical reasons. The most frequently used program is BPROM.

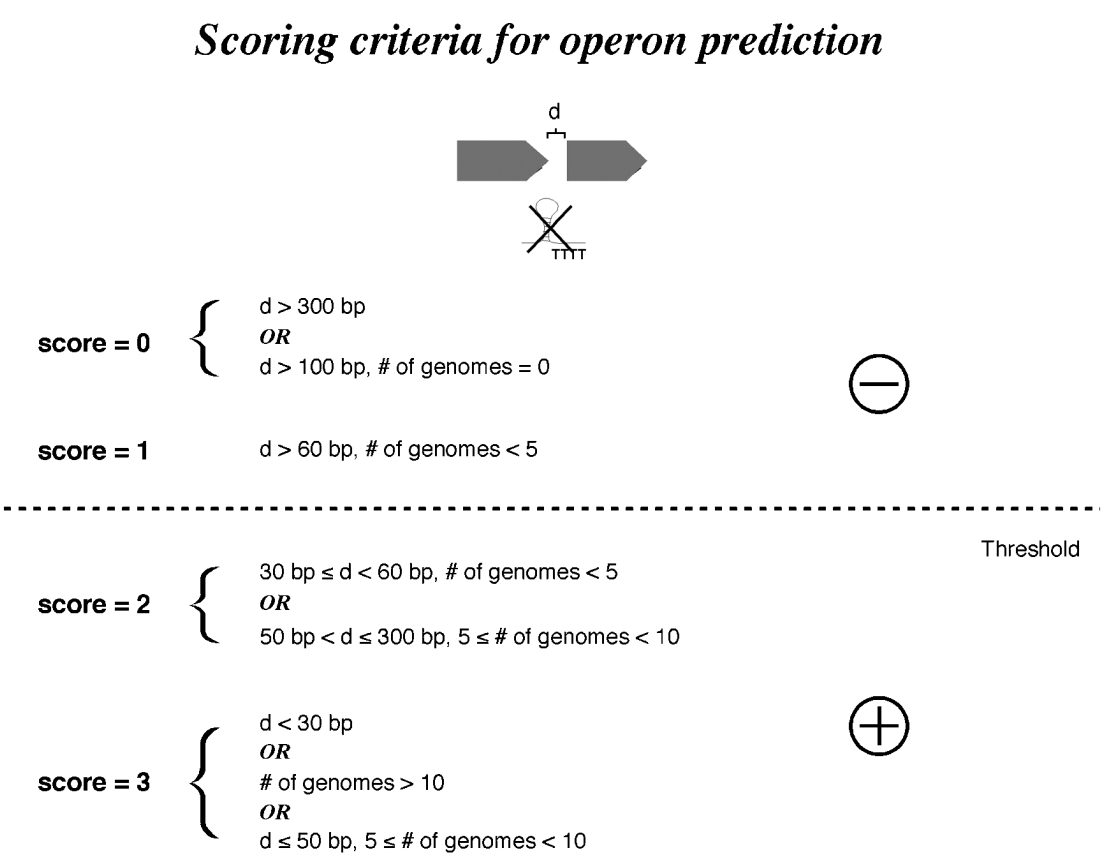

Figure 9.3: Prediction of operons in prokaryotes based on a scoring scheme developed by Wang et al. (2004). This method states that, for two adjacent genes transcribed in the same orientation and without a ρ-independent transcription termination signal in between, the score is assigned 0 if the intergenic distance is larger than 300 bp regardless of the gene linkage pattern or if the distance is larger than 100 bp with the linkage not observed in other genomes. The score is assigned 1 if the intergenic distance is larger than 60 bp with the linkage shared in less than five genomes. The score is assigned 2 if the distance of the two genes is between 30 and 60 bp with the linkage shared in less than five genomes or if the distance is between 50 and 300 bp with the linkage shared in between five to ten genomes. The score is assigned 3 if the intergenic distance is less than 30 bp regardless of the conserved linkage pattern or if the linkage is conserved in more than ten genomes regardless of the intergenic distance or if the distance is less than 50 bp with the linkage shared in between five to ten genomes. A minimum score of 2 is considered the threshold for assigning the two genes in one operon.

BPROM (www.softberry.com/berry.phtml?topic=bprom&group=programs &subgroup=gfindb) is a web-based program for prediction of bacterial promoters. It uses a linear discriminant function (see Chapter 8) combined with signal and content information such as consensus promoter sequence and oligonucleotide composition of the promoter sites. This program first predicts a given sequence for bacterial operon structures by using an intergenic distance of 100 bp as basis for distinguishing genes to be in an operon. This rule is more arbitrary than the Wang et al. rule, leading to high rates of false positives. Once the operons are assigned, the program is able to predict putative promoter sequences. Because most bacterial promoters are located within 200 bp of the protein coding region, the program is most effectively used when about

200 bp of upstream sequence of the first gene of an operon is supplied as input to increase specificity.

FindTerm (http://sun1.softberry.com/berry.phtml?topic=findterm&group=programs&subgroup=gfindb) is a program for searching bacterial ρ-independent termination signals located at the end of operons. It is available from the same site as FGENES and BPROM. The predictions are made based on matching of known profiles of the termination signals combined with energy calculations for the derived RNA secondary structures for the putative hairpin-loop structure (see Chapter 16). The sequence region that scores best in features and energy terms is chosen as the prediction. The information can sometimes be useful in defining an operon.

Prediction for Eukaryotes

The ab initio method for predicting eukaryotic promoters and regulatory elements also relies on searching the input sequences for matching of consensus patterns of known promoters and regulatory elements. The consensus patterns are derived from experimentally determined DNA binding sites which are compiled into profiles and stored in a database for scanning an unknown sequence to find similar conserved patterns. However, this approach tends to generate very high rate of false positives owing to nonspecific matches with the short sequence patterns. Furthermore, because of the high variability of transcription factor binding sites, the simple sequence matching often misses true promoter sites, creating false negatives.

To increase the specificity of prediction, a unique feature of eukaryotic promoter is employed, which is the presence of CpG islands. It is known that many vertebrate genes are characterized by a high density of CG dinucleotides near the promoter region overlapping the transcription start site (see Chapter 8). By identifying the CpG islands, promoters can be traced on the immediate upstream region from the islands. By combining CpG islands and other promoter signals, the accuracy of prediction can be improved. Several programs have been developed based on the combined features to predict the transcription start sites in particular.

As discussed, the eukaryotic transcription initiation requires cooperation of a large number of transcription factors. *Cooperativity* means that the promoter regions tend to contain a high density of protein-binding sites. Thus, finding a cluster of transcription factor binding sites often enhances the probability of individual binding site prediction.

A number of representatives of ab initio promoter prediction algorithms that incorporate the unique properties of eukaryotic promoters are introduced next.

CpGProD (http://pbil.univ-lyon1.fr/software/cpgprod.html) is a web-based program that predicts promoters containing a high density of CpG islands in mammalian genomic sequences. It calculates moving averages of GC% and CpG ratios (observed/expected) over a window of a certain size (usually 200 bp). When the values are above a certain threshold, the region is identified as a CpG island.

Eponine (http://servlet.sanger.ac.uk:8080/eponine/) is a web-based program that predicts transcription start sites based on a series of preconstructed PSSMs of several regulatory sites, such as the TATA box, the CCAAT box, and CpG islands. The query sequence from a mammalian source is scanned through the PSSMs. The sequence stretches with high-score matching to all the PSSMs, as well as matching of the spacing between the elements, are declared transcription start sites. A Bayesian method is also used in decision making.

Cluster-Buster (http://zlab.bu.edu/cluster-buster/cbust.html) is an HMM-based, web-based program designed to find clusters of regulatory binding sites. It works by detecting a region of high concentration of known transcription factor binding sites and regulatory motifs. A query sequence is scanned with a window size of 1 kb for putative regulatory motifs using motif HMMs. If multiple motifs are detected within a window, a positive score is assigned to each motif found. The total score of the window is the sum of each motif score subtracting a gap penalty, which is proportional to the distances between motifs. If the score of a certain region is above a certain threshold, it is predicted to contain a regulatory cluster.

FirstEF (First Exon Finder; http://rulai.cshl.org/tools/FirstEF/) is a web-based program that predicts promoters for human DNA. It integrates gene prediction with promoter prediction. It uses quadratic discriminant functions (see Chapter 8) to calculate the probabilities of the first exon of a gene and its boundary sites. A segment of DNA (15 kb) upstream of the first exon is subsequently extracted for promoter prediction on the basis of scores for CpG islands.

McPromoter (http://genes.mit.edu/McPromoter.html) is a web-based program that uses a neural network to make promoter predictions. It has a unique promoter model containing six scoring segments. The program scans a window of 300 bases for the likelihoods of being in each of the coding, noncoding, and promoter regions. The input for the neural network includes parameters for sequence physical properties, such as DNA bendability, plus signals such as the TATA box, initiator box, and CpG islands. The hidden layer combines all the features to derive an overall likelihood for a site being a promoter. Another unique feature is that McPromoter does not require that certain patterns must be present, but instead the combination of all features is important. For instance, even if the TATA box score is very low, a promoter prediction can still be made if the other features score highly. The program is currently trained for *Drosophila* and human sequences.

TSSW (www.softberry.com/berry.phtml?topic=promoter) is a web program that distinguishes promoter sequences from non-promoter sequences based on a combination of unique content information such as hexamer/trimer frequencies and signal information such the TATA box in the promoter region. The values are fed to a linear discriminant function (see Chapter 8) to separate true motifs from background noise.

CONPRO (http://stl.bioinformatics.med.umich.edu/conpro) is a web-based program that uses a consensus method to identify promoter elements for human DNA.

To use the program, a user supplies the transcript sequence of a gene (cDNA). The program uses the information to search the human genome database for the position of the gene. It then uses the GENSCAN program to predict 5′ untranslated exons in the upstream region. Once the 5′-most exon is located, a further upstream region (1.5 kb) is used for promoter prediction, which relies on a combination of five promoter prediction programs, TSSG, TSSW, NNPP, PROSCAN, and PromFD. For each program, the highest score prediction is taken as the promoter in the region. If three predictions fall within a 100-bp region, this is considered a consensus prediction. If no three-way consensus is achieved, TSSG and PromFD predictions are taken. Because no coding sequence is used in prediction, specificity is improved relative to each individual program.

Phylogenetic Footprinting–Based Method

It has been observed that promoter and regulatory elements from closely related organisms such as human and mouse are highly conserved. The conservation is both at the sequence level and at the level of organization of the elements. Therefore, it is possible to obtain such promoter sequences for a particular gene through comparative analysis. The identification of conserved noncoding DNA elements that serve crucial functional roles is referred to as *phylogenetic footprinting;* the elements are called *phylogenetic footprints*. This type of method can apply to both prokaryotic and eukaryotic sequences.

The selection of organisms for comparison is an important consideration in this type of analysis. If the pair of organisms selected are too closely related, such as human and chimpanzee, the sequence difference between them may not be sufficient to filter out functional elements. On the other hand, if the organisms' evolutionary distances are too long, such as between human and fish, long evolutionary divergence may render promoter and other elements undetectable. One example of appropriate selection of species is the use of human and mouse sequences, which often yields informative results.

Another caveat of phylogenetic footprinting is to extract noncoding sequences upstream of corresponding genes and focus the comparison to this region only, which helps to prevent false positives. The predictive value of this method also depends on the quality of the subsequent sequence alignments. The advanced alignment programs introduced in Chapter 5 can be used. Even more sophisticated expectation maximization (EM) and Gibbs sampling algorithms can be used in detecting weakly conserved motifs.

There are software programs specifically designed to take advantage of the presence of phylogenetic footprints to make comparisons among a number of related species to identify putative transcription factor binding sites. The advantage in implementing the algorithms is that no training of the probabilistic models is required; hence, it is more broadly applicable. There is also a potential to discover new regulatory

motifs shared among organisms. The obvious limitation is the constraint on the evolutionary distances among the orthologous sequences.

ConSite (http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite) is a web server that finds putative promoter elements by comparing two orthologous sequences. The user provides two individual sequences which are aligned by ConSite using a global alignment algorithm. Alternatively, the program accepts precomputed alignment. Conserved regions are identified by calculating identity scores, which are then used to compare against a motif database of regulatory sites (TRANSFAC). High-scoring sequence segments upstream of genes are returned as putative regulatory elements.

rVISTA (http://rvista.dcode.org/) is a similar cross-species comparison tool for promoter recognition. The program uses two orthologous sequences as input and first identifies all putative regulatory motifs based on TRANSFAC matches. It then aligns the two sequences using a local alignment strategy. The motifs that have the highest percent identity in the pairwise comparison are presented graphically as regulatory elements.

PromH(W) (www.softberry.com/berry.phtml?topic=promhw&group=programs&subgroup=promoter) is a web-based program that predicts regulatory sites by pairwise sequence comparison. The user supplies two orthologous sequences, which are aligned by the program to identify conserved regions. These regions are subsequently predicted for RNA polymerase II promoter motifs in both sequences using the TSSW program. Only the conserved regions having high scored promoter motifs are returned as results.

Bayes aligner (www.bioinfo.rpi.edu/applications/bayesian/bayes/bayes_align12.pl) is a web-based footprinting program. It aligns two sequences using a Bayesian algorithm which is a unique sequence alignment method. Instead of returning a single best alignment, the method generates a distribution of a large number of alignments using a full range of scoring matrices and gap penalties. Posterior probability values, which are considered estimates of the true alignment, are calculated for each alignment. By studying the distribution, the alignment that has the highest likelihood score, which is in the extreme margin of the distribution, is chosen. Based on this unique alignment searching algorithm, weakly conserved motifs can be identified with high probability scores.

FootPrinter (http://abstract.cs.washington.edu/~blanchem/FootPrinterWeb/FootPrinterInput2.pl) is a web-based program for phylogenetic footprinting using multiple input sequences. The user also needs to provide a phylogenetic tree that defines the evolutionary relationship of the input sequences. (One may obtain the tree information from the "Tree of Life" web site [http://tolweb.org/tree/phylogeny.html], which archives known phylogenetic trees using ribosomal RNAs as gene markers.) The program performs multiple alignment of the input sequences to identify conserved motifs. The motifs from organisms spanning over the widest evolutionary distances are identified as promoter or regulatory motifs. In other words, it identifies unusually well-conserved motifs across a set of orthologous sequences.

Expression Profiling–Based Method

Recent advances in high throughput transcription profiling analysis, such as DNA microarray analysis (see Chapter 18) have allowed simultaneous monitoring of expression of hundreds or thousands of genes. Genes with similar expression profiles are considered coexpressed, which can be identified through a clustering approach (see Chapter 18). The basis for coexpression is thought to be due to common promoters and regulatory elements. If this assumption is valid, the upstream sequences of the coexpressed genes can be aligned together to reveal the common regulatory elements recognizable by specific transcription factors.

This approach is essentially experimentally based and appears to be robust for finding transcription factor binding sites. The problem is that the regulatory elements of coexpressed genes are usually short and weak. Their patterns are difficult to discern using simple multiple sequence alignment approaches. Therefore, an advanced alignment-independent profile construction method such as EM and Gibbs motif sampling (see Chapter 7) is often used in finding the subtle sequence motifs. As a reminder, EM is a motif extraction algorithm that finds motifs by repeatedly optimizing a PSSM through comparison with single sequences. Gibbs sampling uses a similar matrix optimization approach but samples motifs with a more flexible strategy and may have a higher likelihood of finding the optimal pattern. Through matrix optimization, subtly conserved motifs can be detected from the background noise.

One of the drawbacks of this approach is that determination of the set of coexpressed genes depends on the clustering approaches, which are known to be error prone. That means that the quality of the input data may be questionable when functionally unrelated genes are often clustered together. In addition, the assumption that coexpressed genes have common regulatory elements is not always valid. Many coexpressed genes have been found to belong to parallel signaling pathways that are under the control of distinct regulatory mechanisms. Therefore, caution should always be exercised when using this method.

The following lists a small selection of motif finders using the EM or Gibbs sampling approach.

MEME (http://meme.sdsc.edu/meme/website/meme-intro.html) is the EM-based program introduced in Chapter 7 for protein motif discovery but can also be used in DNA motif finding. The use is similar to that for protein sequences.

AlignACE (http://atlas.med.harvard.edu/cgi-bin/alignace.pl) is a web-based program using the Gibbs sampling algorithm to find common motifs. The program is optimized for DNA sequence motif extraction. It automatically determines the optimal number and lengths of motifs from the input sequences.

Melina (Motif Elucidator In Nucleotide sequence Assembly; http://melina.hgc.jp/) is a web-based program that runs four individual motif-finding algorithms – MEME, GIBBS sampling, CONSENSUS, and Coresearch – simultaneously. The user compares the results to determine the consensus of motifs predicted by all four prediction methods.

INCLUSive (www.esat.kuleuven.ac.be/~dna/BioI/Software.html) is a suite of web-based tools designed to streamline the process of microarray data collection and sequence motif detection. The pipeline processes microarray data, automatically clusters genes according expression patterns, retrieves upstream sequences of coregulated genes and detects motifs using a Gibbs sampling approach (Motif Sampler). To further avoid the problem of getting stuck in a local optimum (see Chapter 7), each sequence dataset is submitted to Motif Sampler ten times. The results may vary in each run. The results from the ten runs are compiled to derive consensus motifs.

PhyloCon (Phylogenetic Consensus; http://ural.wustl.edu/~twang/PhyloCon/) is a UNIX program that combines phylogenetic footprinting with gene expression profiling analysis to identify regulatory motifs. This approach takes advantage of conservation among orthologous genes as well as conservation among coregulated genes. For each individual gene in a set of coregulated genes, multiple sequence homologs are aligned to derive profiles. Based on the gene expression data, profiles between coregulated genes are further compared to identify functionally conserved motifs among evolutionary conserved motifs. In other words, regulatory motifs are defined from both sets of analysis. This approach integrates the "single gene–multiple species" and "single species–multiple genes" methods and has been found to reduce false positives compared to either phylogenetic footprinting or simple motif extraction approaches alone.

SUMMARY

Identification of promoter and regulatory elements remains a great bioinformatic challenge. The existing algorithms can be classified as ab initio based, phylogenetic footprinting based, and expression profiling based. The true accuracy of the ab initio programs is still difficult to assess because of the lack of common benchmarks. The reported overall sensitivity and specificity levels are currently below 0.5 for most programs. For a prediction method to be acceptable, both accuracy indicators have to be consistently above 0.9 to be reliable enough for routine prediction purposes. That means that the algorithmic development in this field still has a long road ahead. To achieve better results, combining multiple prediction programs seems to be helpful in some circumstances. The comparative approach using phylogenetic footprinting is able to take a completely different approach in identifying promoter elements. The resulting prediction can be used to check against the ab initio prediction. Finally, the experimental based approach using gene expression data offers another route to finding regulatory motifs. Because the DNA motifs are often subtle, EM and Gibbs motif sampling algorithms are necessary for this purpose. Alternatively, the EM and Gibbs sampling programs can be used for phylogenetic footprinting if the input sequences are from different organisms. In essence, all three approaches are interrelated. The results from all three types of methods can be combined to further increase the reliability of the predictions.

FURTHER READING

Dubchak, I., and Pachter, L. 2002. The computational challenges of applying comparative-based computational methods to whole genomes. *Brief. Bioinform*. 3:18–22.

Hannenhalli, S., and Levy, S. 2001. Promoter prediction in the human genome. *Bioinformatics* 17(Suppl):S90–6.

Hehl, R., and Wingender, E. 2001. Database-assisted promoter analysis. *Trends Plant Sci*. 6:251–5.

Ohler, U., and Niemann, H. 2001. Identification and analysis of eukaryotic promoters: Recent computational approaches. *Trends Genet.* 17:56–60.

Ovcharenko, I., and Loots, G. G. 2003. Finding the needle in the haystack: Computational strategies for discovering regulatory sequences in genomes. *Curr. Genomics* 4:557–68.

Qiu, P. 2003. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem. Biophys. Res. Commun*. 309:495–501.

Rombauts S., Florquin K., Lescot M., Marchal K., Rouze P., and van de Peer Y. 2003. Computational approaches to identify promoters and *cis*-regulatory elements in plant genomes. *Plant Physiol*. 132:1162–76.

Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., and Collado-Vides, J. 2000. Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc. Natl. Acad. Sci. U S A* 97:6652–7.

Wang, L., Trawick, J. D., Yamamoto, R., and Zamudio, C. 2004. Genome-wide operon prediction in *Staphylococcus aureus*. *Nucleic Acids Res.* 32:3689–702.

Werner, T. 2003. The state of the art of mammalian promoter recognition. *Brief. Bioinform*. 4:22–30.