Computational Approaches to Identify Promoters and cis-Regulatory Elements in Plant Genomes

Stephane Rombauts, Kobe Florquin, Magali Lescot, Kathleen Marchal, Pierre Rouze, and Yves Van de Peer

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, B–9000 Gent, Belgium (S.R., K.F., Y.V.d.P.); Laboratoire de Génétique et Physiologie du Développement, Equipe bioinformatique, Centre National de la Recherche Scientifique, Parc Scientifique de Luminy, F-13288 Marseille Cedex 9, France (M.L.); Department of Electrical Engineering (Electronics, Systems, Automatisation and Technology-Signals, Identification, System Theory, and Automation), Katholieke Universiteit Leuven, B–3001 Heverlee, Belgium (K.M.); and Laboratoire Associé de l’Institut National de la Recherche Agronomique (France), Ghent University, K.L. Ledeganckstraat 35, B–9000 Gent, Belgium (P.R.)

The identification of promoters and their regulatory elements is one of the major challenges in bioinformatics and integrates comparative, structural, and functional genomics. Many different approaches have been developed to detect conserved motifs in a set of genes that are either coregulated or orthologous. However, although recent approaches seem promising, in general, unambiguous identification of regulatory elements is not straightforward. The delineation of promoters is even harder, due to its complex nature, and in silico promoter prediction is still in its infancy. Here, we review the different approaches that have been developed for identifying promoters and their regulatory elements. We discuss the detection of cis-acting regulatory elements using word-counting or probabilistic methods (so-called “search by signal” methods) and the delineation of promoters by considering both sequence content and structural features (“search by content” methods). As an example of search by content, we explored in greater detail the association of promoters with CpG islands. However, due to differences in sequence content, the parameters used to detect CpG islands in humans and other vertebrates cannot be used for plants. Therefore, a preliminary attempt was made to define parameters that could possibly define CpG and CpNpG islands in Arabidopsis, by exploring the compositional landscape around the transcriptional start site. To this end, a data set of more than 5,000 gene sequences was built, including the promoter region, the 5′-untranslated region, and the first introns and coding exons. Preliminary analysis shows that promoter location based on the detection of potential CpG/CpNpG islands in the Arabidopsis genome is not straightforward. Nevertheless, because the landscape of CpG/CpNpG islands differs considerably between promoters and introns on the one side and exons (whether coding or not) on the other, more sophisticated approaches can probably be developed for the successful detection of “putative” CpG and CpNpG islands in plants.

Arabidopsis, and probably most plants, encode an exceptionally large number of DNA-binding proteins, potentially acting as transcription factors (TFs). In fact, more than 3,000 genes have been anticipated to be involved in transcription, more than one-half of which were expected to encode TFs (Arabidopsis Genome Initiative, 2000), corresponding to more than 5% of the Arabidopsis genes, and approximately twice the ratio observed for yeast and animal genomes (Riechmann et al., 2000). These TFs bind to the DNA on specific cis-acting regulatory elements (CAREs) and orchestrate the initiation of transcription, which is one of the most important control points in the regulation of gene expression. CAREs are short conserved motifs of five up to 20 nucleotides usually found in the vicinity of the 5′ end of genes in what is called the promoter. The promoter sequence is usually located upstream from the transcription start site (TSS), but regulatory elements can also be located downstream, for example, in the first intron of the gene itself (Zhang et al., 1994; Gidekel et al., 1996; de Boer et al., 1999; Dorsett, 1999). The promoter can roughly be divided in two parts: a proximal part, referred to as the core, and a distal part. The proximal part is believed to be responsible for correctly assembling the RNA polymerase II complex at the right position and for directing a basal level of transcription (Nikolov et al., 1996; Nikolov and Burley, 1997; Berk, 1999). It is mediated by elements, such as TATA and Initiator boxes through the binding of the TATA box-binding protein, and other elements in what is called the proximal part. The distal part is more involved in transcriptional regulation through the binding of specific DNA-binding proteins to specific regulatory sequences in the promoter.
shown as small gray circles. The pre-initiation complex (PIC) situated at the nucleosome-free TSS is shown containing RNA polymerase II (large gray hatched oval), the TATA box-binding protein (gray sphere), and a number of general TFs (white circles). Gene regulatory proteins upstream or downstream of the TSS that stimulate gene-specific transcription and also contribute to the PIC assembly are shown as small gray circles.

Figure 1. Graphical, simplified view of the different elements involved in transcription. The pre-initiation complex (PIC) situated at the nucleosome-free TSS is shown containing RNA polymerase II (large gray hatched oval), the TATA box-binding protein (gray sphere), and a number of general TFs (white circles). Gene regulatory proteins upstream or downstream of the TSS that stimulate gene-specific transcription and also contribute to the PIC assembly are shown as small gray circles.
many false positives, i.e. on average one false posi-
tive per 1 kb. One of the reasons why these programs
did not perform better was that they were focused on
“search by signal” using only one or two given fea-
tures, such as the presence of a TATA box or Initiator
element, but disregarded structural and more general
sequence-based features characteristic for promoter
elements (Fickett and Hatzigeorgiou, 1997). Newer
approaches do take into account more features; they
consider the higher order structure of a promoter
DNA sequence important for transcriptional regula-
tion and are based on the concept that they share
common content features, although polymerase II
promoters are quite different in terms of individual
organization.

On the one hand, sequence-based algorithms aim
at identifying regulatory regions and promoters
based on their sequence composition compared with
that of non-promoters. Among others, Scherf et al.
(2000) and Bajic et al. (2002) have used this approach,
in which the promoter context is described by oligo-
nucleotides (see below; Hutchinson, 1996; Wolfert-
stetter et al., 1996). On the other hand, promoter
regions might be distinguished from non-promoter
regions on the basis of specific structural properties.
These features are either directly or indirectly corre-
lated with the three-dimensional structure a pro-

Figure 2. Flow chart of the computational approaches to detect
promoters and cis-regulatory elements. 1, Promoter prediction
through sequence context and structural features, e.g. CpG islands; 2,
CARE prediction through statistics on overrepresentation, such as word
counting; 3, CARE prediction through comparative genomics (phylo-
genetic footprinting); 4, CARE prediction through analysis of co-
expressed gene clusters, for instance by Gibbs sampling (for details,
see text); 5, Promoter prediction through the identification of CAREs;
and 6, CARE motif prediction through comparative analysis of expres-
sion profiles. These approaches are not described in the text.

its the means to model it. Clear and unequivocal
descriptions of genomic segments that contain all
elements required to activate transcription would be
useful but are still unavailable, although the regula-

tory motifs of some specific genes have been inves-
tigated in detail. Therefore, most in silico research on
promoters is usually restricted to the so-called inter-
genic regions of the genome, i.e. between the coding
regions of two neighboring genes. The most practical
approach is to limit the putative promoter region to
an arbitrary number of base pairs upstream of the
translation start site of the gene of interest, the loca-
tion of the TSS being unknown most of the time.
However, ideally, this number should be chosen in
function of the organism, because the length of inter-
genic regions may differ considerably (Arabidopsis
Genome Initiative, 2000; International Human Ge-

ome Sequencing Consortium, 2001; Aparicio et al.,
2002). In multicellular organisms, regulatory ele-
ments may be found upstream or downstream of the
gene, as well as in introns, and may be spread over
tens or even hundreds of kilobases (Larkin et al.,
1993; Bagga et al., 2000). In such cases, intergenic
sequences will contain only a part of the regulatory
elements necessary to control transcription.

In 1997, Fickett and Hatzigeorgiou (1997) thor-
oughly reviewed the existing promoter prediction
tools. The programs tested could not reliably identify
promoters in a genomic sequence, predicting too
many false positives, i.e. on average one false posi-

Copyright © 2003 American Society of Plant Biologists. All rights reserved.

Downloaded from on August 8, 2017 - Published by www.plantphysiol.org
instead of using only one model, the program constructs three different models that differentiate promoters from exons, from introns, and from 3′-untranslated regions (UTRs). A given sequence will be assigned to the class of “promoters” only when all models are in agreement with the decision. The specificity and significance of this program is highly dependent on the given training sets that build up the different models.

McPromoter (Ohler et al., 1999, 2001; Ohler, 2000) is a content-based probabilistic promoter prediction program that uses an integrative approach combining different structural features, such as bendability (Brukner et al., 1995a, 1995b), propeller twist (El Hassan and Calladine, 1996), and CpG content (Antequera and Bird, 1999; Ioshikhes and Zhang, 2000). Here, a promoter is represented as a sequence of consecutive segments represented by joint likelihood of DNA sequence and profiles of physical properties. A profile for a physical property consists of the corresponding values from a chosen parameter, for example the bendability, set along the given DNA sequence. These parameters usually refer to di- and trinucleotides only, so the profiles are generally very noisy and are, therefore, smoothed with a filter. The program tries to divide a given sequence into one region upstream and one downstream from the TSS. A search by signal is used to distinguish the core promoter from the other parts by looking for a TATA and/or Initiator box separated by a spacer of approximately 15 bp.

Although the prediction tools hitherto developed can produce acceptable results for certain species, none of them have been trained and adapted for plants. For example, McPromoter is trained especially to analyze data of fruitfly (Drosophila melanogaster) and has been used in the Genome Annotation Assessment project (Reese et al., 2000; Ohler, 2000; Ohler et al., 2002); however, when applied to plant genomes, it is not as reliable nor as specific. Here, the same rule applies as with gene prediction: Systems have to be trained and tailored for each species separately (Mathé et al., 2002). For the careful training of systems, large amounts of reliable data are needed. Although the availability of large sets of documented promoter sequences is still problematic, we expect this will improve in the near future. An extensive overview of the available programs for the prediction of promoters is given in Table I.

### Promoters and CpG Islands

A structural feature that has proven useful in the detection of promoters in the human genome are the so-called CpG islands, i.e. regions that are rich in CpGs, which are important because of their strong link with gene regulation. In general, CpG-rich regions are methylated and are associated with inactive DNA often linked to heterochromatin, gene silencing, and pathogen control (Jeddeloh et al., 1998; Kooter et al., 1999; Wolff and Matzke, 1999; Meyer, 2000; Bender, 2001; Vaucheret and Fagard, 2001; Richards and Elgin, 2002; Robertson, 2002). In vertebrate genomes, 60% to 90% of all CpGs are normally methylated. Gene-associated CpG islands are mostly not methylated and are usually linked to transcriptionally active DNA (Panstruga et al., 1998; Razin, 1998; Antequera and Bird, 1999; Jones, 1999; Ng and Bird, 1999; Ashikawa, 2001; Li et al., 2001). Prediction programs have been developed to search for the presence of CpG islands in the 5′ region of genes (Ioshikhes and Zhang, 2000; Ohler et al., 2001; Davuluri et al., 2001; Down and Hubbard, 2002; Ponger and Mouchiroud, 2002). However, so far, application of such prediction programs to CpG islands in plants is very limited. A more detailed analysis on CpG and CpNpG islands in Arabidopsis is given below.

Although the functional significance of methylation appears to be similar in humans and plants (Hershkovitz et al., 1990; Weber et al., 1990; Inamdar et al., 1991; Meyer et al., 1994; Sorensen et al., 1996; Rossi et al., 1997; Meza et al., 2002), in plants, DNA methylation is mainly found on the cytosine of the di- and trinucleotide CpG and CpNpG and on nonsymmetrical trinucleotides (Pradhan et al., 1999; Cao et al., 2000; Finnegan and Kovac, 2000; Lindroth et al., 2001; Cao and Jacobsen, 2002). Many plant genomes contain methylated cytosine in asymmetric sequence contexts (CpHpH with H = A, T, or C). Only symmetrical methylation sites have been shown to be maintained through the propagation of cells and

### Table 1 Promoter prediction programs

<table>
<thead>
<tr>
<th>Programs</th>
<th>Web Sites</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>McPromoter MM</td>
<td><a href="http://genes.mit.edu/McPromoter.html">http://genes.mit.edu/McPromoter.html</a></td>
<td>Ohler et al. (1999, 2001)</td>
</tr>
<tr>
<td>FunSiteP</td>
<td><a href="http://transac.gbf.de/programs/funsite/fsp.html">http://transac.gbf.de/programs/funsite/fsp.html</a></td>
<td>Kondrakhin et al. (1995)</td>
</tr>
</tbody>
</table>
methylation of a promoter CpG island has been proposed to play an important role in gene silencing, genomic imprinting, heterochromatin formation, chromatin modification, vernalization, and parent-dependent effects (Finnegan et al., 1998a, 2000; Jeddeloh et al., 1998; Sturaro and Viotti, 2001). CpNpG and CpG islands can occur together in the same promoter region, but their role might be different (Sorensen et al., 1996).

The Landscape of CpG/CpNpG Islands around the TSS in the Arabidopsis Genome

CpG islands are characterized by a locally increased GC percentage (GC%) compared with local averages and by the presence of CpGs (and CpNpGs in plants). The CpG dinucleotide, usually methylated at the fifth position on the cytosine ring, is counterselected and found much less frequently than expected based on mononucleotide frequencies, for example, 5-fold lower in genomes of vertebrates. This depletion is believed to result from accidental mutations by deamination of 5-methylcytosine to thymine (Sved and Bird, 1990; Duret and Galtier, 2000). In fact, CpG islands are considered evolutionary remnants, because some promoters have somehow been kept free of methylation in the course of evolution, so the deamination process is hampered. Another explanation could be that to function as part of an expression pattern, a selection pressure has to be exerted and, hence, CpG islands stand out in the surrounding regions.

The original pragmatic definition of a CpG island in human sequences considers a GC% higher than 50 and a ratio between observed and expected (o/e) occurrence of CG dinucleotides of 0.6 over a window of 200 bp (Gardiner-Garden and Frommer, 1987). Recently, these parameters have been upscaled to a GC% >55, an o/e CpG >0.65, and a window size of 500 bp, because the previous parameters had been found to overestimate (50-fold) the number of potential CpG islands (Takai and Jones, 2002). In animals, approximately 40% of genes are expected to be associated with CpG islands (Gardiner-Garden and Frommer, 1987; Antequera and Bird, 1999). Actually, this percentage might be too low because a total of 29,000 CpG islands had been estimated after the completion of the human genome sequence (Venter et al., 2001). With the above-mentioned parameters, no CpG islands are discovered in plants (Takai and Jones, 2002; our results). However, DNA methylation occurs in plants, and DNA methylases are even more numerous and diverse in plants than in animals (Finnegan and Kovac, 2000; Cao and Jacobsen, 2002). Therefore, we attempted to define the parameters that could possibly specify CpG and CpNpG islands in Arabidopsis by exploring the compositional landscape around the TSS. To this end, we built a data set of 5,025 gene sequences, designated ARAPROM, by aligning the full-length cDNA sequences generated by Seki et al. (2002) against the genomic sequence (Arabidopsis Genome Initiative, 2000). Generally, these sequences are 2.5 kb long, in which 2 kb represent intergenic sequences upstream from the translation start codon, and 500 bp are taken downstream. Nevertheless, when the upstream neighbor gene lies closer than 2 kb, then only the intergenic sequence is kept, up to the predicted coding boundary of the upstream gene. The genomic sequences in the ARAPROM data set include the promoter region, the 5'-UTR, and the first introns and coding exons of each individual gene.

A program in Perl was written that computes the GC content and the o/e ratios of CpG and CpNpG compared with local characteristics over a certain window size. By applying this program to the ARAPROM data set to extract potential CpG/CpNpG islands, we tested the effect of setting the cut-off values for the GC content and the o/e CpG/CpNpG ratios at different levels (39% to 52% with a stepwise increase of 0.5% for the GC content; 0.6% to 2.0% for the o/e CpG and CpNpG ratios with a stepwise increase of 0.1). The results of this analysis with a window size of 200 bp are shown graphically for CpG and CpNpG islands (Figs. 3 and 4, respectively). The first observation is that no CpG island is detected with the cut-off parameters tuned for humans, except for a few in coding exons. Both parameters, GC% and o/e CpG, appear to influence strongly the number of CpG islands detected and, depending on the position in the genome, to affect differently the number of CpGs found. That number found in the “promoter” region sharply increases while the GC% cut-off de-
Creases (Fig. 3). In contrast, for coding exons, the landscape resembles more a plateau, with many CpG islands found already at much higher GC% values. Only at the lowest GC% values, more CpG islands are predicted in the "promoter" region than in the coding exons. In UTR exons, which show a landscape similar to that of coding exons, fewer CpG islands are found and introns, which show a landscape more similar to that of the "promoter" region, show the lowest number of CpG islands.

Regarding the CpNpG landscape (Fig. 4), the major observation is that the overall number of islands lies well below that of the CpG islands. Only at the lowest GC% values, more CpG islands are predicted in the "promoter" region than in the coding exons. In UTR exons, which show a landscape similar to that of coding exons, fewer CpG islands are found and introns, which show a landscape more similar to that of the "promoter" region, show the lowest number of CpG islands.

Figure 4. CpNpG island landscape exploration of Arabidopsis gene sequences over a range of CG content and CpNpG relative frequency. For the various gene elements, on the z axis, the number of CpNpG islands found in the ARAPROM gene set is plotted against the thresholds defined on the x and y axes, being the CG percentage and the o/e CpNpG ratio, respectively. The window size was 200 bp. Similar landscapes are obtained for other window sizes (100 and 400 bp) and are available at http://www.psb.rug.ac.be/bioinformatics/.

This preliminary in silico analysis shows that prediction of promoter location based on the detection of potential CpG/CpNpG islands in the Arabidopsis genome is not straightforward. Nevertheless, because the landscape of CpG/CpNpG islands differs considerably between promoters and introns on the one side and exons (whether coding or not) on the other, there is some hope that, based on such a classification, more sophisticated approaches can be developed to detect CpG and CpNpG islands in plants.

**SIGNAL-BASED FEATURES**

**Regulatory Elements**

As stated in the introduction, CAREs are short, conserved motifs of approximately 5 to 20 nucleotides. Detection of CAREs in the promoter is not self-evident, because such short motifs are statistically expected to occur at random every few hundred base pairs. Therefore, the main problem lies in discriminating "true" from "false" regulatory elements (Blanchette and Sinha, 2001). It is important to distinguish whether unknown or known motifs are looked for. Compared with the detection of unknown motifs, that of known motifs is fairly straightforward and consists of the scanning of the DNA sequence with a given motif, which can be found in specialized databases such as TRANSFAC (Wingender et al., 1996) and TFD (Ghosh, 2000) and in plant-specific databases such as PLACE (Higo et al., 1999) and PlantCARE (Lescot et al., 2002). An overview of the different databases and motif search programs is given in Tables III and IV, respectively. In contrast, the detection of unknown CAREs in large regions of DNA requires the development and use of novel approaches and algorithms. Specifically, local multiple alignment algorithms that identify regulatory motifs have already been developed, which are merely based on statistical properties. Such algorithms search for DNA patterns that are more frequently present in a set of "related" than "unrelated" sequences. Therefore, the successful identification of

<table>
<thead>
<tr>
<th>CG% Threshold</th>
<th>o/e Threshold</th>
<th>With Island in Promoter</th>
<th>With Island in at Least One Exon</th>
<th>With Island in Promoter AND in Exons</th>
<th>With Island in Promoter ONLY</th>
<th>With Island in Exons ONLY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cpg</td>
<td>39</td>
<td>0.6</td>
<td>85.75</td>
<td>99.42</td>
<td>85.73</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>1.6</td>
<td>0.82</td>
<td>98.95</td>
<td>0.82</td>
<td>0.00</td>
</tr>
<tr>
<td>Cpnpg</td>
<td>39</td>
<td>0.6</td>
<td>64.98</td>
<td>99.38</td>
<td>64.84</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>1.6</td>
<td>0.02</td>
<td>98.73</td>
<td>0.02</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*Refer to exons either in coding regions or in the UTR region.
regulatory DNA patterns depends on the size of the promoter sequence and, to a great extent, on the quality of the set of "related" sequences, i.e. genes that are co-expressed or coregulated and are thus expected to share similar conserved regulatory motifs. Such co-expressed genes are identified based on high-throughput gene expression profiling experiments. Alternatively, instead of coregulated genes, intergenic regions of orthologous sequences can also constitute a valuable data set for motif detection (Duret and Bucher, 1997). When selection pressure tends to conserve DNA patterns in the intergenic regions of homologous genes in related species, such DNA patterns can be expected to be of biological relevance and to reflect a conserved ancestral mode of regulation.

Regulatory Elements in Coregulated Genes

Co-expressed genes can be identified through transcript profiling techniques, such as microarrays (Brown and Botstein, 1999; Lipshutz et al., 1999; Southern, 2001) and cDNA-AFLP (Vos et al., 1995; Breyné et al., 2002). These high-throughput profiling techniques allow the expression level of hundreds or thousands of genes to be monitored simultaneously under the conditions tested. For each gene, an expression profile is obtained that reflects its dynamic behavior during a time-course experiment or its behavior under distinct conditions. Genes with similar expression profiles are considered "co-expressed". To identify sets of co-expressed genes from high-throughput expression data, clustering techniques are required (Heyer et al., 1999; Jensen and Knudsen, 2000). In addition to standard cluster algorithms, such as hierarchical clustering, K-means, and self-organizing maps, more advanced algorithms are also being developed, which are specifically fine-tuned for biological applications (for a review, see Moreau et al., 2002).

Because co-expressed genes tend to behave similarly, they are expected to be coregulated. Under the simplifying assumption that this coregulation occurs at the transcriptional level, co-expressed genes should contain similar cis-regulatory elements in their promoter regions. As a consequence, these yet unknown cis-regulatory elements will be statistically overrepresented in the intergenic regions of the co-expressed genes in comparison with their frequent occurrence in a set of unrelated sequences. This over-representation constitutes the general principle on which motif detection algorithms is based.

Regulatory Elements in Orthologs

Usually, genes are part of more extensive gene families that have originated through both speciation and duplication events. Homologous genes in dis-
distinct species are called orthologs, whereas paralogs refer to homologous genes that are found in the same genome and have been created through gene duplication (Mindell and Meyer, 2001). Regarding promoter analysis and study of regulatory elements, it is important to discriminate between these two types of homologous relationships. True orthologs have usually retained very similar functions in distinct species, whereas this is not necessarily true for paralogs. In many cases, paralogs have only been conserved if they have acquired different or complementary functions. Hughes (1994) and Force et al. (1999) argued that when a gene with multiple functions is duplicated, the duplicates are only redundant for as long as each gene is capable of performing all ancestral roles. When one mutated duplicate is prevented from carrying out one of these ancestral roles, the other duplicate is no longer redundant. According to the "duplication degeneration complementation" model of Force et al. (1999), degenerative mutations preserve rather than destroy duplicated genes, but also change their functions or, at least, restrict them to become more specialized. Duplicated genes can have different expression domains (i.e., the tissue in which both genes are expressed might have changed as well as the time of expression) because of changes in their regulatory elements in the promoter region (Force et al., 1999; Altschmied et al., 2002; Prince and Pickett, 2002). Therefore, promoter regions of true orthologs probably contain similar regulatory motifs, which may no longer be true for paralogs.

Motif Prediction

To conceive a general method that can detect regulatory motifs is a great challenge because of both the complexity and flexibility of the regulatory mechanisms (see the introduction). An important distinction between the different approaches used thus far to detect regulatory motifs lies in the representation of the motif, i.e., the TF-binding site. The simplest description for a motif is a string of characters (A, C, G, and T), extended with the 11 IUPAC characters that represent partly unspecified or ambiguous nucleotides, and is used in the string-based approaches, such as word counting. A more sophisticated description is to represent a given motif by describing it in a probabilistic manner in which a certain likelihood is assessed for each nucleotide at a given position in the motif. An example of a probabilistic representation is the position-weight matrix, where each column corresponds to a position in the aligned binding sites and each row to a nucleotide, as shown in Figure 5. The cells of these matrices contain a number indicating the probability to find a given nucleotide at that particular position. Alternatives to describe motifs in a probabilistic manner are the hidden Markov models (Jarmer et al., 2001) or neural networks (Workman and Stormo, 2000). Software tools for motif prediction are listed in Table V.

String-Based Motif Prediction

Counting all of the possible words that may occur across the different promoter sequences is one of the simplest approaches to find CAREs in a set of promoters. Among word-counting methods, enumerative and suffix-tree approaches can be distinguished, the latter being an optimization of the former. Both methods are string-based: The DNA sequence is considered as text in which oligonucleotides are represented as words or strings. For a given set of promoter sequences, the frequency of each possible word of a defined length is computed (Hutchinson, 1996; Wolpertetster et al., 1996; Brázma et al., 1998; van Helden et al., 1998; Vanet et al., 1999; Busse-maker et al., 2000a, 2000b; Sinha and Tompa, 2000; Hampson et al., 2002). The difference between the two representations is that the enumerative approach will search for each word in the sequence and calculate its frequency, whereas the suffix-tree approach will only look for a certain subset. The suffix tree is used to represent each word together with all of its subwords (or suffixes) so that each word can be reconstructed by going down the tree. For example, when a certain suffix (for instance, ACCT) is not found in the data set, none of the words containing ACCT will be counted anymore (Sagot and Myers, 1998; Marsan and Sagot, 2000). As a consequence, the computational time needed to count words can be highly reduced, which allows the analysis of larger data sets.

Once the frequencies of different words are calculated, the words that are likely to be a "true" regulatory motif have to be differentiated from those that
represented as a weight matrix and altered through a searching process. CONSENSUS was one of the first algorithms that represented a motif by a weight matrix, expectation maximization and its stochastic variant, Gibbs sampling, are often used for optimization strategies. The program CONSENSUS was one of the first algorithms that represented a motif by a weight matrix (Hertz et al., 1990; Hertz and Stormo, 1996, 1999). The algorithm starts with a first sequence from the submitted data set and creates a weight matrix for each possible word of user-specified length. Subsequently, it aligns each possible word from the next sequence with each weight matrix. The obtained alignments are scored for their information content, and those with the highest score are retained for the next iteration. This process is reiterated until all sequences have contributed to the alignment and weight matrix. CONSENSUS was used for example in the identification of CAREs involved in the heat shock response in Caenorhabditis elegans (GuhaThakurta and Stormo, 2001).

### Probabilistic Motif Detection

Probabilistic motif detection aims at constructing a multiple alignment by locally aligning small conserved regions in a set of unaligned sequences. Here, we will focus on the matrix-based approaches to illustrate probabilistic motif detection procedures. All methods start from a random motif model, represented as a weight matrix and altered through a series of iterations by machine-learning algorithms that are aimed at finding the optimal score. The process of optimizing the score for a local alignment already tends to converge toward conserved motifs that occur frequently in the data set. The more advanced algorithms incorporate a background model to compensate for given motifs occurring at high frequencies because of compositions similar to those of the non-conserved parts of the sequence (the “background”). A motif in which the average nucleotide composition differs strongly from the background will be assigned a higher score. Implementations differ from each other in the way the background is represented, in how the score is calculated, and in how the optimization is performed. For motif detection algorithms that describe the motif by a weight matrix, expectation maximization and its stochastic variant, Gibbs sampling, are often used for optimization strategies.

<table>
<thead>
<tr>
<th>Table V. Motif prediction programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programs</td>
</tr>
<tr>
<td>AlignACE</td>
</tr>
<tr>
<td>Improbizer</td>
</tr>
<tr>
<td>BioProspector</td>
</tr>
<tr>
<td>Footprinter 2.0</td>
</tr>
</tbody>
</table>
corresponds to the current motif model (weight matrix). In the maximization step, the parameters that optimize the likelihood are estimated. Once the motif positions are known, the observed frequencies of the nucleotides at each motif position correspond to the maximum-likelihood estimates of the parameters of the motif model. On the basis of the updated probabilities of all motif positions of the previous step, the motif parameters are re-estimated. For motif finding, EM simultaneously computes the alignment positions, the motif weight matrix, and the background model that maximize the likelihood of the sequence.

In the original implementation (Lawrence and Reilly, 1990), the “exactly one occurrence” of the motif in each sequence was assumed. This assumption is a problem because in a cluster of co-expressed genes, sequences might be present without a (or the) motif. Because EM-based motif detection algorithms are deterministic, results for particular queries with similar parameter settings and initializations will be identical. A drawback of the method is that results depend strongly on the initial conditions and often converge into local optima. One of the most widely used EM applications is the program MEME (Bailey and Elkan, 1995).

Gibbs sampling-based strategies have originally been developed to detect protein motifs but have been adapted later on to handle DNA sequences (Neuwald et al., 1995; Roth et al., 1998; Hughes et al., 2000; Liu et al., 2001; Thijs et al., 2002a). Gibbs sampling is a stochastic variant of EM (Lawrence et al., 1993; Neuwald et al., 1995). Because of the stochastic nature of the Gibbs sampling approach, an initially detected motif can be replaced by another one that has a higher score, thus allowing escape from local optima. This feature is the reason why the output of a stochastic motif detection algorithm results in different outputs, even with the same input and parameter settings. However, the more pronounced the optimal solution is in a given data set, the more a motif is overrepresented, and the stronger its conservation, the more frequently it will be retrieved over different runs. Statistics on the outcome of multiple runs of a stochastic implementation can facilitate interpretation of the results.

Adaptative quality-based clustering (De Smet et al., 2002) combined with Motif Sampler based on Gibbs sampling (Thijs et al., 2002a) was applied to the data published by Reymond et al. (2000) in which gene expression was studied in response to mechanical wounding in Arabidopsis leaves. After clustering, the four most populated clusters (>3 genes) of co-expressed genes were selected, and the upstream sequences were analyzed with the Motif Sampler to discover common regulatory elements. To avoid the problem of local optima, each data set was submitted 10 times to the Motif Sampler with the same parameters. The output of these 10 runs was compiled taking into account the individual scores of each motif and the order in which they were found. Subsequently, the consensus of the motifs found were compared with regulatory sites described in the PlantCARE database (Lescot et al., 2002). From all of the high-ranking motifs returned by the Motif Sampler, several were similar to known cis-regulatory elements involved in plant defense (methyl jasmonate-, abscisic acid-, or elicitor-responsive elements) or in light responsiveness. Among these elements, a 12-bp motif was found composed of two sites involved in methyl jasmonate responsiveness. These motifs have been described previously in the upstream sequence of the lipooxygenase isoenzyme 1 gene of barley, where they were separated by 15 bp (Thijs et al., 2002a).

Motif Prediction by Phylogenetic Footprinting

The procedure that identifies regulatory elements based on a set of orthologous sequences is named phylogenetic footprinting (Koop, 1995; Duret and Bucher, 1997; Wasserman et al., 2000). Phylogenetic footprinting has proven its usefulness to detect CAREs in the human genome, based on the pairwise comparison between human and mouse (Hardison, 2000; Wasserman et al., 2000; Krivan and Wasserman, 2001; Dermitzakis and Clark, 2002; Jegga et al., 2002). However, producing a reliable data set for phylogenetic footprinting is not self-evident. When the overall degree of conservation in intergenic sequences between two homologs is too high, conserved motifs will not be detected. At the other extreme, when homologs are compared from species that are too distantly related, the intergenic regions may no longer show any similarity (Tompa, 2001). The ideal composition of a data set can only be derived in retrospect, implying that an algorithm suited for phylogenetic footprinting should ideally identify and discard (or counter-weigh) sequences that are too similar and cope with the presence of sequences that do not contain the conserved motif. Furthermore, the phylogenetic distance between organisms should be taken into account in the weighting schemes of the algorithm. As stated before, closely related sequences are less useful for identifying a motif because of their high overall conservation, complicating the search for functionally conserved regions. Alignment algorithms, such as ClustalW (Thompson et al., 1994) and Bayes-Block Aligner (Zhu et al., 1998), have proven useful for phylogenetic footprinting, but the length of the conserved motif is often too small compared with the length of the non-conserved part of the sequence; therefore, multiple sequence alignment will fail.

A promising novel algorithm has recently been published that identifies the most conserved motifs among the input sequences as measured by a parsimony score on the underlying phylogenetic tree (Blanchette et al., 2002; Blanchette and Tompa, 2002). In general, the algorithm selects motifs that are char-
acterized by a minimal number of mismatches and are conserved over long evolutionary distances. Furthermore, the motifs should not have undergone independent losses in multiple branches. In other words, the motif should be present in the sequences of subsequent taxa along a branch. The algorithm, based on dynamic programming, proceeds from the leaves of the phylogenetic tree to its root and seeks for motifs of a user-defined length with a minimum number of mismatches. Moreover, the algorithm allows a higher number of mismatches for those sequences that span a greater evolutionary distance. Motifs that are lost along a branch of the tree are assigned an additional cost because it is assumed that multiple independent losses are unlikely in evolution. To compensate for spurious hits, statistical significance is calculated based on a random set of sequences in which no motifs occur. Phylogenetic footprinting for the detection of CAREs is steadily gaining importance (Koch et al., 2001, 2002; Quiros et al., 2001; Colinas et al., 2002) and will continue to do so when more plant genomic sequences will become available. To give just one example, using phylogenetic footprinting, Tompa (2001) was able to predict several new binding sites in the 5′-UTR of plant genes coding for the small subunit of ribulose-1,5-bisphosphate carboxylase.

Improvements and Fine Tuning of Motif Detection Algorithms

The most obvious reason why motif detection algorithms fail is because of their sensitivity to noise. All parts of a sequence that do not contain the motif constitute noise in the context of motif detection. Moreover, because sets of related sequences are usually based on other predictive tools, for instance clustering, they are expected to contain sequences without any shared motif. A decreasing signal-to-noise ratio exacerbates the identification of statistically overrepresented motifs and increases the chance of finding false positives. Probabilistic motif detection methods have been improved considerably to cope with a large noise level. Current implementations, such as AlignACE (Hughes et al., 2000) and MEME (Bailey and Elkan, 1995), usually take into account that some sequences lack a shared motif and allow the influence of such sequences to be discarded by estimating the motif model parameters. The more advanced implementations derive the optimal number of motif occurrences in each sequence from the data. Modeling the background with a more complex sequence model contributes also considerably to the robustness of the algorithm in the presence of noise (Liu et al., 2001, 2002; Thijs et al., 2001, 2002a). Besides making more robust algorithms that facilitate discrimination between true and false positives, advanced scoring schemes are being developed that assign a statistical significance to the motifs detected, i.e. that describe the probability of observing a motif with a similar score in a set of unrelated sequences.

Because regulatory motifs, in particular in higher eukaryotes, are concentrated in modules, current research is focusing toward adapting motif detection algorithms to retrieve dyads, i.e. motifs spaced by a fixed or variable gap. Within the enumerative statistical methods, Sinha and Tompa (2000) created an algorithm that searches for motifs with a gap of variable size between them. The algorithm developed by van Helden et al. (2000) enables a search for dyads with a fixed number of base pairs between 3 and 20.

Vanet et al. (2000) and Marsan and Sagot (2000) developed approaches that look for two motifs separated by a fixed number of nucleotides by using the suffix-tree method. Cardon and Stormo (1992) adapted their EM-based algorithm to detect dyads with variable gap size. In their Gibbs sampling-based implementation, Liu et al. (2001) have included an extension that allows searching for dyads, whereas the program Co-bind of GuhaThakurta and Stormo (2001) was specifically created to identify two regulatory sites of gap-separated cooperative TFs.

The need for extensive parameter fine tuning complicates nonexpert use of most of the motif detection approaches described above. Novel implementations of motif detection algorithms tackle this problem by estimating the optimal parameter settings themselves, hence, minimizing the number of user-defined parameters. An example of such a user-defined parameter is the motif length. Because the motif length is generally unknown in advance, it is not obvious to choose the parameter setting that results in the true motif. Some algorithms compute the optimal motif length; for instance, Pattern assembly (van Helden et al., 2000b) groups overlapping motifs to build a longer motif consensus. The implementation of AlignAce determines the optimal motif length from the data (Roth et al., 1998; Hughes et al., 2000). Manually generating a suitable data set (see above) that can be used readily for motif detection can be a tedious job. Therefore, some on-line implementations have been developed, such as the INCLUSIve (Thijs et al., 2002b; Engelen et al., 2003) Web site that offers a pipeline to combine microarray preprocessing, the adaptive quality-based clustering, automatic sequence retrieval, and motif detection based on Gibbs sampling (Motif Sampler). A tool similar to INCLUSIve, called expression profiler (Brázma et al., 1998), is provided by the European Bioinformatics Institute. “Regulatory Sequence Analysis tools” (van Helden et al., 1998) proposes a word counting-based set of tools to analyze a set of intergenic sequences.

CONCLUSIONS

Promoters are very complex structures, defined by many different structural features. The actual regu-
latory elements are usually very short, which highly complicates their unambiguous identification. As a consequence, the in silico prediction of promoters and regulatory motifs is not straightforward. In addition, our knowledge of transcription regulation in general and organism-specific expression regulation in particular, is still very limited. Especially for plants, solid "intrinsic" genomic data are still needed that can be integrated into existing prediction tools. In this respect, we have started with the analysis of CpG and CpNpG islands, known to be often associated with promoters. Although several implementations for the detection of such "islands" in vertebrates have been described (Ioshikhes and Zhang, 2000), parameter settings used to detect these islands in animals cannot be used to find similar islands in the Arabidopsis genome. Software and parameters have to be adapted to the species under investigation. Moreover, even if a reliable tool were available for the detection of CpG and CpNpG islands associated with plant promoters, it remains to be proven whether these islands would be biologically functional and relevant. In addition to a lack of "intrinsic" genomic data, experimental data on promoters are also scarce, because in general, thorough analysis of even one single promoter is very time consuming. Furthermore, the technology is still missing for exhaustive knowledge of gene expression or for understanding the mechanisms behind it. Therefore, for now and despite its many shortcomings, in silico analysis seems to be a privileged alternative to analyze simultaneously a great number of regulatory elements or promoter regions. Experimental testing of these in silico predictions may be a manner to increase knowledge on promoters more quickly and at a lower cost, especially for plants.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for helpful suggestions.

Received November 14, 2002; returned for revision January 10, 2003; accepted March 17, 2003.

LITERATURE CITED


that the O2 B-Zip protein and other endosperm factors can bind to methylated target sequences. J Biol Chem 272: 13758–13765


