ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles

Thomas Abel1,2, Yvan Saeys1,2, Pierre Rouzé1,3 and Yves Van de Peer1,2,*

1Department of Plant Systems Biology, VIB, 2Department of Molecular Genetics and 3Laboratoire Associé de l’INRA, Ghent University, Technologiepark 927, 9052 Gent, Belgium

1 INTRODUCTION

Currently, the genomic sequence of over 50 eukaryotic organisms is available. Many more sequencing projects are to be finished in the next few years (Liolios et al., 2006). In light of the many genomes that become available, it is important to accurately identify these regions, as better core promoter predictions will improve the genome annotation and allow a better understanding of the transcription initiation process. The core promoter is the region immediately upstream of the TSS, where the transcription-initiation complex assembles (Pedersen et al., 1999; Smale and Kadonaga, 2003). It is located upstream of the coding part of the gene, sometimes up to several thousand base pairs, and is responsible for basal transcription as well as transcriptional regulation of the gene it is linked with (Choi et al., 2004). Genes encoding structural or regulatory RNAs instead of proteins also contain a core promoter, but have no coding part. This is often problematic for the current promoter prediction programs (PPPs) that are often trained to recognize the 5’ end of protein coding genes (Bajic et al., 2004).

Recently, it has been shown that genes usually do not have a single TSS. This finding is well documented in humans as an outcome of genome-wide investigations on gene expression (Carninci et al., 2006; The ENCODE Project Consortium, 2007). Most human genes have multiple promoters and each promoter has multiple TSSs (Frith et al., 2006; Sandelin et al., 2007). For most genes, transcription starts in a cluster of positions, with some positions favored over others. The choice of which alternative promoter to use depends on the conditions in the cell. The use of alternative promoters results in the complexity and diversity of the human transcriptome (Kawai et al., 2006). All these studies imply that core promoter prediction techniques should not try to predict a single TSS but rather a cluster of TSSs. In addition, the use of alternative promoters depends on the promoter type: conserved tissue-specific promoters have a better defined TSS and are TATA-enriched, while more loosely defined CpG-rich promoters have a broader promoter region, often with multiple alternative promoters and many TSSs (Carninci et al., 2006).

Core promoters have distinct features that can be used to distinguish them from other sequences. On a nucleotide level, several motifs have been linked to the core promoter in eukaryotes. The best-known motif, the TATA box, is mainly present in tissue-specific genes. Estimations of the number of TATA-containing promoters in the human genome range from 5–30% (Deng and Roberts, 2005; Florquin et al., 2005; Smale and Kadonaga, 2003). Other motifs that are related to the core promoter are, among others, the initiator element and the TFIIB recognition element (Deng and Roberts, 2005). These motifs have specific consensus sequences that can be used to identify the motif. When scanning all promoters, only a limited number of promoters contain the motif as defined by the consensus sequences. Furthermore, when screening the whole genome for these sequences, many occurrences of the consensus do not correspond to a functional site. Besides motifs, also more
The sequences for the human genome assembly (hg17, May 2004) were since the very beginning (Fickett and Hatzigeorgiou, 1997). Predictions, a problem that has been inherent to promoter prediction. Compared to other PPPs, ProSOM results in significantly fewer false positives which are responsible for the accurate distinction of core promoter regions from the rest of the genome.

ProSOM, that uses an unsupervised self-organizing map (SOM) to distinguish core promoter regions from the rest of the genome, is especially interesting because it points to the global variation of G+C around the TSS, and to the importance of small-scale features that have been described, but then not used in practice. The clustering technique we used is the SOM (Kohonen, 2001), a special classifying technique that takes into account all the properties and selects the relevant ones (Saeys et al., 2007). However, this is not a trivial task and it is not further explored here. Techniques from the field of ensemble learning may prove useful in this context (Polikar, 2006).

In previous research, we have used a single average structural value over a window of 400 bp to predict core promoter regions (Abeel et al., 2008). This simple property was based on a large-scale feature of the core promoter that is present in several eukaryotes (Abeel et al., 2008). Here, we use the small-scale properties of the core promoter that have been described, but then not used for promoter prediction. We will revisit these small-scale features below.

We present a novel promoter prediction technique, called ProSOM, that uses an unsupervised self-organizing map (SOM) to distinguish core promoter regions from the rest of the genome. Compared to other PPPs, ProSOM results in significantly fewer false predictions, a problem that has been inherent to promoter prediction since the very beginning (Fickett and Hatzegeorgiou, 1997). Furthermore, we propose a new validation strategy for promoter prediction, taking into account comparisons with experimentally verified TSS data. This novel validation strategy results in a more realistic evaluation of promoter predictors, and has the potential to become a new evaluation standard for future promoter predictors.

2 MATERIAL AND METHODS

2.1 Data

The sequences for the human genome assembly (hg17, May 2004) were retrieved from the UCSC Genome Bioinformatics Site (http://genome.ucsc.edu/) (Karolchik et al., 2008).

The cap analysis gene expression (CAGE) datasets have been compiled by Carninci et al. (2006) and retrieved from the Fantom3 project (http://fantom.gsc.riken.go.jp/). The dataset was compiled with the CAGE technique (Strakal et al., 2003) and covers the entire human genome. This technique is based on the preparation and sequencing of concatamers of DNA tags derived from the initial 20 nucleotides of 5' end mRNAs. It allows high-throughput analysis of gene expression and the identification of TSSs. The CAGE technique maps the TSS more accurately than other techniques. However, to remove any false hits from the CAGE technique we only retained tag clusters with at least two mapped tags, resulting in 123 400 unique TSSs for humans.

The Ensembl gene annotation was retrieved using the BioMart tool for Ensembl release 37 (Flicek et al., 2008). The ENCODE regions and annotation were retrieved from the ENCODE home page (http://genome.ucsc.edu/ENCODE/) (The ENCODE Project Consortium, 2007).

For the training of the SOM we retrieved promoters, and transcribed and intergenic sequences. The sequences for promoters were extracted from the DRTS8 database (Wakazaki et al., 2008). From the 1250bp-long sequences provided in this database, we extracted the region [-200, 50] around the TSS. Sequences containing ambiguous symbols were discarded. This resulted in 30 964 sequences. The transcribed and intergenic sequences were extracted from the genome assembly. We used the gene coordinates retrieved from Ensembl using the BioMart tool. From the edges of these sequences we stripped 500bp to remove border signals. Finally we selected 30 000 sequences of 250bp on random locations from the set of transcribed and intergenic sequences. The large number of sequences lowers the influence of the few TSSs that may still be present in this set have on the outcome.

In total, we have three datasets, one with promoter sequences, one with intergenic sequences and one with transcribed sequences. Each of these sets contains 30 000 sequences, or slightly more in the case of the promoter set.

2.2 ProSOM implementation

ProSOM was implemented using the Java language version 1.5. The SOM implementation was taken from the Java Machine Learning Library (http://java-ml.sf.net/) and modified to suit our needs. This library is also required to run ProSOM. The program has no size constraints on the number of sequences that can be processed and is designed to run on multiple machines in parallel to process different sequences.

2.3 Structural profiles

The structural profile of a set of DNA sequences is calculated in two steps. First, the nucleotide sequence is converted into a sequence of numbers (i.e. a numerical profile) by replacing each dinucleotide with its energy value, which is obtained from experimentally validated conversion tables. We have used the conversion tables for base-stacking energy from Florquin et al. (2005). Thereafter, for each position, we take the average over all sequences for that position. The resulting numeric vector is called a structural profile. While we take an average of several thousand sequences to calculate the profiles in the SOM, this is not possible for the sequences we extracted from the genome. The sequences for which we want to make predictions, are smoothed using a 3 bp sliding window. This was done to remove some of the noise in the profile of a single sequence.

2.4 Clustering and promoter prediction

The clustering technique we used is the SOM (Kohonen, 2001), a special type of artificial neural network that can be used both for clustering and class prediction. An SOM consists of a rectangular grid of clusters, each of which has a weighted connection to every input node (Fig. 1). In our case, the input nodes represent the different values of a structural profile associated with a potential promoter region. The SOM provides a mapping...
After many cycles, the network ends up associating clusters with groups in the input data set. This is the harmonic mean of the recall (sensitivity) and the precision of correctly predicted promoters divided by the total number of promoters. The recall rate is the number of correctly predicted promoters from other sequences. The recall rate is the number of correctly predicted promoters divided by the total number of promoters.

During the learning phase, weights are updated by iteratively presenting samples to the network. Using the principle of competitive learning, the Euclidean distance to all weight vectors is calculated for each sample, and the cluster with the most similar weight vector is called the best matching unit (BMU). Subsequently, the weight vector of the BMU and its neighboring clusters in the grid are moved slightly towards this sample. The magnitude of the change decreases with distance of the weight vector from the BMU, resulting in a topological mapping of the input data. In a topological mapping, neighboring clusters represent more similar objects than distant clusters. After many cycles, the network ends up associating clusters with groups in the input data set.

Once the cluster structure is learned, the SOM can be used to map new vectors to their corresponding clusters by determining the cluster whose weight vector is nearest to the new vector. To convert this procedure into a PPP, we determined which clusters are clearly associated with promoters (see Fig. 4 and further in the text) and predicted each sequence that is mapped to one of those clusters as a putative promoter.

2.5 Validation

Our technique was validated on two sequence sets; first on the entire human genome assembly (hg17, May 2004) and secondly on the ENCODE regions. For both sets we retrieved a set of experimentally characterized TSSs and a gene annotation. While the training set and validation set overlap to a certain extent, this has little influence on our validation. The training was unsupervised, so the SOM clustering did not use the sequence type information that was provided with the training data. This means that the clustering was not trained to specifically recognize the sequences in the training set, but it clustered all sequences, and clusters with high promoter content emerged. Because the clustering separates promoters from other sequences in an unsupervised way, it can be used for promoter prediction.

An aggregate measure for the performance of a classifier is the F-measure (Van Rijssbergen, 1979). This is the harmonic mean of the recall (sensitivity) and the precision (specificity). The higher this value, the better the classifier is able to separate promoter sequences from other sequences. The recall rate is the number of correctly predicted promoters divided by the total number of promoters. The precision rate is the number of correct predictions divided by the total number of predictions. To assess the number of correctly predicted promoters (true positives, TP), false predictions (false positives, FP) and unpredicted promoters (false negatives, FN), we define the maximum distance that a prediction is allowed to be from the true site. Previous work typically set this number to rather high values of 1000 bp or even 2000 bp. Recently, a more stringent distance of 500 bp was used, and we will also use this more stringent validation for reference to earlier work (Abeel et al., 2006; Goni et al., 2007). For true core promoter prediction validation, predictions should not be further than 50 bp from the actual TSS. Here, we use a validation with a maximum distance of 50 bp to rank the PPPs.

There are two ways to evaluate a PPP. In the classic way to evaluate a PPP (Bajic et al., 2004), one starts from gene annotations, e.g. Ensemble or GENCODE annotation. A TP is then defined as the 5′ end of a gene that has a prediction within 500 bp. A FP is a prediction that lies inside the gene but not within the first 500 bp of the gene. A FN is a 5′ end of a gene that has no associated prediction within 500 bp. All predictions that fall within an intergenic region and are > 500 bp from the 5′ end of a gene, are ignored.

Unfortunately, this technique has some drawbacks and a new scheme has been proposed to evaluate PPPs (Abeel et al., 2008). A first major drawback of the classic scheme is that all intergenic predictions are ignored, while they may well be false predictions. The view of a single TSS at the 5′ end of a gene, on which this technique is based, is no longer viable in light of many recent studies (Carninci et al., 2006; Frith et al., 2008; Kawaij et al., 2006). We included the validation with the classic technique for reference purposes.

We have proposed a more objective way to assess the performance of a PPP based on the genomewide screening for TSSs (Abeel et al., 2008). This technique is based on the CAGE datasets that have been described above. The dataset contains locations where transcription starts. A TP is a known site that has a prediction within 500 bp of a true TSS, a FN is a TSS without a prediction and a FP is a prediction that has no associated TSS in the reference set within 500 bp. However, for core promoter prediction the 500 bp distance is still too wide. The core promoter is only a very small region, and therefore a much smaller distance should be used when validating predictions. Predictions should be within 50 bp of known sites to be counted as TP. Distances larger than 50 bp have little biological significance when talking about core promoters. For this stronger validation we can only use the CAGE dataset because it is the only one containing experimentally characterized TSSs.

Figure 2 shows the new technique that uses CAGE tags as a reference. Two tags are depicted black. Predictions in the region [−500, +500] around each tag (black) are considered to be correct (TP). All other predictions are FPs. In both cases, genes or tags that have no associated prediction are FNs. For the validation to identify true core promoter predictions, all intervals are reduced to [−50, +50].

3 RESULTS

3.1 Sequence profiles

Figure 3 shows the average structural profile of base-stacking energy of the three datasets used for training the SOM. The promoter sequences show a very striking profile with overall lower values than the other two graphs. It has two clear peaks at position −30 (TATA-binding protein location) and position 0 (TSS). The
However, two regions are less stable. The first region, where the TBP is known to bind, is located around the TSS itself, and needs to denature to allow transcription to start. One possible explanation for this stable region is that it provides a contrasting background for the highly unstable peaks, which are essential for transcription initiation. The overall rigid structure combined with the two peaks can be viewed as a guiding mechanism for the transcription apparatus to select the appropriate site to initiate transcription.

### 3.2 Unsupervised clustering identifies promoters

We clustered 30,964 promoters from the DBTSS database (Wakaguri et al., 2008), 30,000 intergenic and 30,000 transcript sequences retrieved from Ensembl (Flicek et al., 2008) using the SOM. Several other clustering techniques, such as k-means, ACO, AQBIC, Cobweb, DB-scan, EM-clustering, farthest first, k-SCS and x-means, have been tested, but did not result in a proper separation of the different sequence types (results not shown). These techniques are thus not suitable for our promoter prediction task and have not been validated as promoter predictor.

Figure 4 shows the result of the SOM clustering of the 30,964 sequences. Each graph in the figure represents a cluster. On these graphs the $X$-axis gives the relative position from the TSS for the promoter set, or the position in the sequence for the other sets. On the $Y$-axis, we see the normalized base-stacking energy. The legend of each graph shows the total number of sequences, the number of promoters (+) and the number of other sequences (−). In the top row, the two left-most graphs contain significantly more promoter than non-promoter sequences. Also the left-most graph on the second row clearly contains more promoter sequences than other ones. Each of these three graphs shows the known core promoter profile with two peaks, one at position −30 associated with the TBP and one on the TSS. Other graphs contain profiles that do not correspond to known core promoter features, as expected, because these clusters contain few promoter sequences.

### 3.3 Parameter tuning

As stated above, two main parameters require tuning: first, the grid size used to do the clustering, and second, the prediction threshold. We have optimized both by applying the approach to the whole genome for each parameter combination. For the grid size we tested $4 \times 4$, $5 \times 5$, $6 \times 6$ and $7 \times 7$. Higher and lower values were not tested as the performance follows a normal distribution with a peak at $6 \times 6$. Each cluster has a different promoter probability. The promoter probability of a cluster is defined as the number of promoter sequences divided by the total number of sequences, e.g. for the top-left cluster in Figure 4 the probability is $0.93$ (3663 promoter sequences divided by 3936). Each cluster has a different promoter probability and we tested each of these probabilities as threshold for promoter classification. All sequences that are mapped to a cluster with a promoter probability higher or equal to the threshold will be predicted to be a putative promoter. So when using the promoter probability from the cluster with the highest promoters rate, only sequences that map to that specific cluster will be predicted to be promoters. When using a lower threshold, all sequences that map to that cluster or to a cluster with a higher promoter probability will be predicted as putative promoters.

Table 1 shows the results for the different grid sizes and the different thresholds. The rows contain the different numbers of clusters to determine the threshold, and the columns the different grid sizes. Even though SOMs are not restricted to square grids, we have opted to check only those to limit the search space for the parameters. Further tuning may improve the promoter prediction slightly. From Table 1 it is clear that a $6 \times 6$ grid combined with the top three clusters performs best.
We used the trained SOM to predict promoter regions. To each cluster we attached a probability that a given sequence assigned to that cluster is a promoter. If the structural profile of a sequence maps to a cluster that has a probability equal to or above the threshold, we predict it as a promoter region.

When designing classification algorithms, researchers often use cross-validation to validate their technique. Cross-validation is usually employed when only a limited amount of data is available. In our case, we have plenty of real-world data that can be used for validation. Once we have trained our approach on the limited set of training samples that is available from DBTSS and Ensembl, we can apply the approach to the entire human genome and compare with experimental screening for TSSs. Since real-world validation is superior to cross-validation, no cross-validation was performed.

To compare programs that result in different recall and precision values different techniques can be utilized. We use the $F$-measure (harmonic mean of precision and recall) to calculate a final score for a program for several reasons. (i) The $F$-measure is a single measure that can compare different programs with different precision and recall scores. (ii) In the case of the validation in a genomewide context it is very hard to estimate the number of TNs correctly because most prediction programs do not classify each site (nucleotide) in the genome as being a promoter or not, but only make predictions of core promoter regions. Recall, precision and $F$-measure do not require knowledge about the number of TNs and are thus suitable to use in this context. (iii) Other comprehensive

### Table 1. Evaluation of the different parameter combinations for ProSOM

<table>
<thead>
<tr>
<th>Grid Size</th>
<th>4 x 4</th>
<th>5 x 5</th>
<th>6 x 6</th>
<th>7 x 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.44</td>
<td>0.40</td>
<td>0.36</td>
<td>0.32</td>
</tr>
<tr>
<td>2</td>
<td>0.46</td>
<td>0.47</td>
<td>0.45</td>
<td>0.41</td>
</tr>
<tr>
<td>3</td>
<td>0.37</td>
<td>0.46</td>
<td>0.48</td>
<td>0.45</td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>0.36</td>
<td>0.44</td>
<td>0.47</td>
</tr>
<tr>
<td>5</td>
<td>0.09</td>
<td>0.29</td>
<td>0.39</td>
<td>0.46</td>
</tr>
</tbody>
</table>

The columns denote the different grid sizes, the rows contain the different thresholds. The first line contains the results when using the threshold of the cluster with the highest promoter probability. The second line contains results when using the promoter probability of the cluster with second highest promoter probability as threshold, and so on. The values in the table are the $F$-measure calculated from the entire human genome when validating against the CAGE dataset. The maximum allowed distance is 500bp. Bold value indicates highest value.

### 3.4 ProSOM validation

We used the trained SOM to predict promoter regions. To each cluster we attached a probability that a given sequence assigned to that cluster is a promoter. If the structural profile of a sequence maps to a cluster that has a probability equal to or above the threshold, we predict it as a promoter region.
Table 2. Evaluation of PPPs using the CAGE and Ensembl datasets with a maximum allowed distance of 500 bp and for the CAGE dataset with a maximum distance of 50 bp.

<table>
<thead>
<tr>
<th>Program</th>
<th>Reference</th>
<th>CAGE (500 bp)</th>
<th>Ensembl (500 bp)</th>
<th>CAGE (50 bp)</th>
<th>Recall</th>
<th>Precision</th>
<th>F</th>
<th>Recall</th>
<th>Precision</th>
<th>F</th>
<th>No. of predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProSOM</td>
<td>Down and Hubbard, 2002</td>
<td>0.38</td>
<td>0.66</td>
<td>0.48</td>
<td>0.48</td>
<td>0.43</td>
<td>0.45</td>
<td>0.17</td>
<td>0.30</td>
<td>0.22</td>
<td>62804</td>
</tr>
<tr>
<td>Eponine</td>
<td>Abeel et al., 2008</td>
<td>0.28</td>
<td>0.75</td>
<td>0.41</td>
<td>0.41</td>
<td>0.56</td>
<td>0.42</td>
<td>0.14</td>
<td>0.29</td>
<td>0.20</td>
<td>60247</td>
</tr>
<tr>
<td>ARTS</td>
<td>Sonnenburg et al., 2006</td>
<td>0.34</td>
<td>0.66</td>
<td>0.45</td>
<td>0.45</td>
<td>0.42</td>
<td>0.46</td>
<td>0.11</td>
<td>0.27</td>
<td>0.16</td>
<td>45765</td>
</tr>
<tr>
<td>FirstEF</td>
<td>Davuluri et al., 2001</td>
<td>0.38</td>
<td>0.74</td>
<td>0.50</td>
<td>0.50</td>
<td>0.53</td>
<td>0.59</td>
<td>0.36</td>
<td>0.27</td>
<td>0.15</td>
<td>47144</td>
</tr>
<tr>
<td>DragonGSSF</td>
<td>Bajic and Bounic, 2003</td>
<td>0.41</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.58</td>
<td>0.58</td>
<td>0.34</td>
<td>0.13</td>
<td>0.14</td>
<td>101985</td>
</tr>
<tr>
<td>PromoterInspector</td>
<td>Scherl et al., 2000</td>
<td>0.32</td>
<td>0.81</td>
<td>0.43</td>
<td>0.43</td>
<td>0.38</td>
<td>0.70</td>
<td>0.49</td>
<td>0.07</td>
<td>0.13</td>
<td>21576</td>
</tr>
<tr>
<td>DragonPF</td>
<td>Bajic et al., 2002</td>
<td>0.51</td>
<td>0.11</td>
<td>0.18</td>
<td>0.18</td>
<td>0.65</td>
<td>0.11</td>
<td>0.19</td>
<td>0.32</td>
<td>0.11</td>
<td>603389</td>
</tr>
<tr>
<td>N-Scan</td>
<td>Gross and Brent, 2006</td>
<td>0.33</td>
<td>0.45</td>
<td>0.38</td>
<td>0.38</td>
<td>0.55</td>
<td>0.51</td>
<td>0.53</td>
<td>0.07</td>
<td>0.09</td>
<td>67748</td>
</tr>
<tr>
<td>CpGProD</td>
<td>Ponger and Mouchoud, 2002</td>
<td>0.34</td>
<td>0.41</td>
<td>0.37</td>
<td>0.37</td>
<td>0.50</td>
<td>0.36</td>
<td>0.42</td>
<td>0.08</td>
<td>0.09</td>
<td>76793</td>
</tr>
<tr>
<td>PromoterExplorer</td>
<td>Xu et al., 2006</td>
<td>0.39</td>
<td>0.30</td>
<td>0.34</td>
<td>0.34</td>
<td>0.55</td>
<td>0.24</td>
<td>0.33</td>
<td>0.09</td>
<td>0.09</td>
<td>132794</td>
</tr>
<tr>
<td>McPromoter</td>
<td>Ohler et al., 2000</td>
<td>0.41</td>
<td>0.17</td>
<td>0.68</td>
<td>0.28</td>
<td>0.24</td>
<td>0.61</td>
<td>0.34</td>
<td>0.05</td>
<td>0.09</td>
<td>20862</td>
</tr>
<tr>
<td>PromFD</td>
<td>Chen et al., 1997</td>
<td>0.44</td>
<td>0.16</td>
<td>0.23</td>
<td>0.23</td>
<td>0.55</td>
<td>0.14</td>
<td>0.22</td>
<td>0.12</td>
<td>0.06</td>
<td>329999</td>
</tr>
<tr>
<td>PromoterScan</td>
<td>Prestridge, 1995</td>
<td>0.16</td>
<td>0.09</td>
<td>0.12</td>
<td>0.12</td>
<td>0.19</td>
<td>0.08</td>
<td>0.11</td>
<td>0.03</td>
<td>0.02</td>
<td>197852</td>
</tr>
<tr>
<td>Promoter2.0</td>
<td>Knudsen, 1999</td>
<td>0.63</td>
<td>0.04</td>
<td>0.08</td>
<td>0.08</td>
<td>0.68</td>
<td>0.03</td>
<td>0.06</td>
<td>0.01</td>
<td>0.01</td>
<td>1919363</td>
</tr>
<tr>
<td>NNPP 2.2</td>
<td>Reese, 2001</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>104746</td>
</tr>
</tbody>
</table>

Bold values indicate ranks for the different programs.
all predictions inside exons as false predictions, which explains the lower than expected precision for the Ensembl dataset. As we have previously shown, the technique to validate a promoter prediction technique with gene annotation is flawed (Abeel et al., 2008), especially since it only takes into account the 5′ TSS and will label any other recognized TSS as a false prediction.

We also analyzed the ENCODE regions of the human genome in more detail. The ENCODE project tries to annotate 1% of the human genome in great detail (The ENCODE Project Consortium, 2007). The regions are partly chosen for their importance and partly random. In total, they cover ∼30 Mbp. For these regions, again are data available for real TSSs, compiled by the Riken Institute using the CAGE technique. There is also a high-quality gene annotation available from the E-GASP/GENCODE project (Grüning et al., 2006). Furthermore, there is a lot of experimental data available that may support TSS predictions. When we apply ProSOM to these regions and compare with the CAGE and GENCODE data and a maximum distance of 500 bp, we get an F-measure of 0.73 and 0.57, respectively. These values are nearly the same as for the EP3 program (0.72 and 0.56). For the maximum distance of 50 bp and the CAGE dataset we get an F-measure of 0.28. These values are higher than those for the validation on the whole genome, which indicates that indeed a lot of the so-called false predictions are, in fact, correct predictions that are missing in the validation set.

To test this hypothesis, we used the Evidence For Transcriptional Activity dataset (Abeel et al., 2008). This dataset is a compendium of experimental indicators for transcription activity. Within this set 98% of our predictions has a hit with a maximum distance of 500 bp and 85% within 50 bp, again indicating that our approach is highly precise.

4 RELATED WORK

Currently, several core PPPs are available that are aimed at predicting TSSs on the whole human genome. Early programs were limited in the amount of data they could process and in their predictive power (Fickett and Hatzigeorgiou, 1997). Only in 2004, promoter prediction tools have been validated on the whole human genome (Bajic et al., 2004). Later, the same authors also reviewed some PPPs on the ENCODE regions of the human genome (Bajic et al., 2006). Gene annotation was used to validate the different PPPs, but in light of new insights gained from the ENCODE project, this is too conservative and even often wrong. The ENCODE project shows that transcription can also start at the 3′ end of a gene or inside exons. Nowadays, it makes more sense to compare real unbiased experimentally characterized TSSs for validation. Recently, we have reviewed several PPPs on a dataset of TSSs that have been determined using the CAGE technique over the whole human genome (Abeel et al., 2008). Wang and co-workers (2007a) focus on alternative promoters and present a novel PPP, MetaProm, based on artificial neural networks. Other recent PPPs include ProStar (Goni et al., 2007), AMOSA-based promoter prediction (Wang et al., 2007b) and EnsemPro (Won et al., 2008). Two of the recent PPPs (MetaProm and EnsemPro) are in fact no real promoter predictors but meta-predictors that aggregate the results of several other programs. These programs probably address the problem of merging the results of several PPPs to some extent, a definitely interesting way to advance the performance of PPPs. Unfortunately, none of these new programs were available for genomewide screening at the time this article was submitted.

5 DISCUSSION AND CONCLUSION

Self-organizing maps provide an intuitive way to cluster DNA sequences. They are unique among unsupervised clustering techniques in their ability to distinguish core promoters from other sequences. When applied to a set of promoter, transcribed and intergenic sequences, they create promoter-rich and promoter-poor clusters. The promoter-rich clusters show the same structural profile that has been observed for core promoter sequences (Abeel et al., 2008). Because the clustering technique is unsupervised we can conclude that the profile that emerges from this clustering technique is indeed one of the hallmarks of the core promoter. Furthermore, the physical description of the core promoter is more general than that based on core promoter motifs. The physical structure of this region is important for transcription initiation as it remains the same, regardless of the presence of motifs.

We packaged this technique as a full-fledged promoter prediction tool, called ProSOM. This technique is made available to academic researchers for free. We used the program to predict core promoters in the human genome and it performs as well as the best existing software packages. Furthermore, it is more balanced regarding the number of retrieved sites and false predictions. The technique is also very precise as 98% of all predictions in the ENCODE region have evidence of transcriptional activity within 500 bp.

ACKNOWLEDGEMENTS

We would like to thank A. De Rycke for the implementation of many of the clustering techniques we tried.

Funding: T.A. is indebted to Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) for a predocitoral fellowship. Y.S. is a postdoctoral fellow of the Research Foundation-Flanders.

Conflict of Interest: none declared.

REFERENCES


