Sequence analysis

Identifying \textit{cis}-regulatory modules by combining comparative and compositional analysis of DNA

Nora Pierstorff\textsuperscript{1,2,*}, Casey M. Bergman\textsuperscript{2} and Thomas Wiehe\textsuperscript{1}

\textsuperscript{1}Institute for Genetics, University of Cologne, Zülpicher Strasse 47, 50674 Cologne, Germany and \textsuperscript{2}Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, M13 9PT Manchester, UK

Received on June 2, 2006; revised on September 8, 2006; accepted on September 28, 2006

Advance Access publication October 10, 2006

Associate Editor: Martin Bishop

ABSTRACT

\textbf{Motivation:} Predicting \textit{cis}-regulatory modules (CRMs) in higher eukaryotes is a challenging computational task. Commonly used methods to predict CRMs based on the signal of transcription factor binding sites (TFBS) are limited by prior information about transcription factor specificity. More general methods that bypass the reliance on TFBS models are needed for comprehensive CRM prediction.

\textbf{Results:} We have developed a method to predict CRMs called CisPlusFinder that identifies high density regions of perfect local ungapped sequences (PLUSs) based on multiple species conservation. By assuming that PLUSs contain core TFBS motifs that are locally overrepresented, the method attempts to capture the expected features of CRM structure and evolution. Applied to a benchmark dataset of CRMs involved in early \textit{Drosophila} development, CisPlusFinder predicts more annotated CRMs than all other methods tested. Using the REDfly database, we find that some 'false positive' predictions in the benchmark dataset correspond to recently annotated CRMs. Our work demonstrates that CRM prediction methods that combine comparative genomic data with statistical properties of DNA may achieve reasonable performance when applied genome-wide in the absence of an a priori set of known TFBS motifs.

\textbf{Availability:} The program CisPlusFinder can be downloaded at http://jakob.genetik.uni-koeln.de/bioinformatik/people/nora/nora.html. All software is licensed under the Lesser GNU Public License (LGPL).

\textbf{Contact:} nora.pierstorff@uni-koeln.de.

\textbf{Supplementary information:} Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Identifying \textit{cis}-regulatory modules (CRMs) is a prominent issue in computational biology. In order to have a better understanding of the mechanistic aspects of gene regulation it is necessary to identify systematically the target sites for transcriptional regulatory proteins. CRMs are also important targets for evolutionary change since phenotypic differences between species may be attributable to differences in gene regulation rather than to differences in the gene itself (Wilson, 1975; Tautz, 2000; Khaitovich et al., 2006). In contrast to the gene-finding problem, the problem of finding CRMs is much more difficult—partly because of the lack of a commonly accepted definition of a CRM, owing to their inherent variability in structure. As a working definition we call a CRM of gene \(x\) a stretch of DNA of several hundred bases located within, proximal or distal to gene \(x\), which contains potential target sites of transcription factors and whose presence or absence has a measurable effect on the expression of gene \(x\).

Most programs to predict CRMs, such as Ahab (Rajewsky et al., 2002) or ClusterBuster (Frisch et al., 2003) use information about known TFBSs to identify potential CRMs. The results of these methods can be improved by using information about the conservation of predicted TFBSs, such as done by the programs Stubb (Sinha et al., 2004) and ecis-analyst (Berman et al., 2004). These methods yield reasonably good results but cannot generally be applied to genes for which the upstream transcription factors have not yet been studied experimentally. Thus, they cannot easily be used for a genome-wide CRM annotation since only a minor fraction of all active transcription factors and their binding sites is known.

An alternative approach is to use information contained in the pattern of conservation between related sequences. For example the method of Grad et al. (2004) uses patterns of sequence conservation between two species to predict CRMs with similar properties to a training set. Another approach, generally known as phylogenetic shadowing (Bofelli et al., 2003), uses more than two species to predict functional regions. Methods using conservation signals alone can be limited in their performance because TFBS conservation necessary to maintain regulatory function may not be significantly higher than in non-binding sequence (Emberly et al., 2003) and TFBSs can be gained or lost even when CRM function is conserved (Ludwig et al., 2000). In addition, conservation signals depend on the local mutation rate and selective constraints which vary over the genome.

Here, we show that CRMs can be identified in \textit{Drosophila melanogaster} by combining conservation signals with a second signal based on identifying locally overrepresented sequence motifs, similar to the method described by Papatsenko et al. (2002). Often, functional TFBSs are associated with a cluster of degenerate repeats of their core motifs. Their biological function may be the guidance of a transcription factor to its binding site (Kim et al., 1987; Khory et al., 1990; Coleman and Pugh, 1995). Our method, called CisPlusFinder, locates sequences that are both perfectly conserved in multiple genomes and contain an overrepresented core motif as signal of a TFBS, and subsequently clusters...
sequences with these properties to predict CRMs. We developed and applied our method in the genus Drosophila, a species group that is well suited for a multi-way comparative analysis since draft genome sequences are available for *D.melanogaster* and 11 other species (http://rana.lbl.gov/drosophila/). Our results indicate that combining conservation and compositional signals from a set of multiple closely related species, which sum to a sufficiently high substitution rate can lead to better CRM predictions than has been achieved thus far.

2 METHODS

2.1 Overview over CisPlusFinder

The motivating idea behind CisPlusFinder is to find regions in a target genome, which contain a high density of perfect local ungapped sequences (PLUSs) that are shared among a set of closely related species. Individual PLUSs are selected on the basis that: (1) the length of a PLUS is unlikely relative to random occurrence; (2) at least one core motif contained in a PLUS is locally overrepresented and (3) additional PLUSs fulfilling conditions (1) and (2) occur within the immediate neighborhood. Underlying these requirements is the idea that PLUSs contain core TFBS motifs, and that local clustering of PLUSs represents the comparative genomic signal of clustered TFBS that typify CRMs.

Figure 1 shows a flowchart describing CisPlusFinder. A detailed description of the method can be found in the Supplementary materials. The first section of Figure 1 shows the calculation of the nucleotide homology map, which is based on chained blastn-hits (Altschul et al., 1990). The second part of Figure 1 depicts the stepwise calculation of putative PLUSs, which have a significant length and contain within them a core motif that is overrepresented in a local window centered on the PLUS. In order to find the set of putative PLUSs suffix trees are built using Ukkonen’s algorithm (Gusfield, 1997). The equations to calculate the significance of the length of a PLUS and the overrepresentation of possible core motifs are explained in the Supplementary materials. The last section of the flowchart depicts the selection of PLUSs, which possess at least the minimal regulatory potential, and the clustering of PLUSs into CRMs. We note that our definition of ‘regulatory potential’ is not related to that used by King et al. (2005).

2.2 Measuring prediction accuracy

To evaluate our CRM predictions, we use the same measurements as Chan and Kibler (2005): sensitivity (SN) and positive predicted value (PPV) on the CRM and nucleotide levels and specificity (SP) on the nucleotide level. They are

\[
SN = \frac{TP}{TP + FN}
\]

(1)

\[
PPV = \frac{TP}{TP + FP}
\]

(2)

and

\[
SP = \frac{TN}{TN + FP}
\]

(3)

where TP means true positive, TN means true negative, FP means false positive and FN means false negative nucleotides or CRMs. We believe that the most meaningful measurement is SN on the CRM level, which quantifies the proportion of annotated CRMs discovered by the computational method under evaluation.

These performance measurements are often used to evaluate and compare gene prediction programs and classifiers that partition an input set into two...
groups. Applying the same measurements to CRM prediction the following problems have to be considered. One problem is caused by the fact that the exact boundaries are not known for most CRMs. It is possible that annotated CRMs exceed the boundaries of the functional CRM causing many false negative nucleotides. It is also possible that additional functional regions of a CRM exist outside an annotated CRM giving rise to false positive nucleotides. More false positive nucleotides and predictions are caused by the fact that many CRMs are still unknown and not annotated yet. Thus false positive predictions are expected and high specificity may not necessarily reflect good accuracy of a program. Another problem is that the equation one prediction = one CRM is not true for most of the prediction programs. PPV on CRM level is calculated assuming this fact. PPV also depends strongly on the length of the analysed sequence. If the sequence is short enough the probability of a false positive prediction is limited by space.

Because of these considerations we developed a standardized version of PPV to train our method, denoted PPV', which calculates the length-adjusted probability that a prediction hits an annotated CRM. This measurement on the nucleotide level is defined as

\[
PPV' = \frac{TP}{TP + FP'},
\]

where \( P \) means the number of nucleotides covered by annotated CRMs and \( N \) means the number of nucleotides not covered by annotated CRMs. In addition, in training the parameters of our method (see 3.1.2) we avoided penalizing predictions that overlap annotated CRMs but do not hit the exact boundaries of the annotations. Thus, for a CRM prediction that hit an annotated CRM, the number of false positive nucleotides outside of the boundary of the hit annotation were not used to calculate PPV' or SP. Likewise, nucleotides in annotated CRMs outside the boundary of an overlapping prediction were not counted as false negatives in calculating SN. To evaluate the performance of CisPlusFinder in comparison with other methods or on the test set we used the script by Chan and Kibler (2005) without modification. This script considers a CRM to be discovered if the overlap between it and a predicted CRM exceeds 50 bp and therefore without modification. This script considers a CRM to be discovered if the number of nucleotides not covered by annotated CRMs. In equation one prediction

\[
PPV = \frac{TP}{TP + FP},
\]

where the number of nucleotides covered by annotated CRMs and \( N \) means the number of nucleotides not covered by annotated CRMs. In addition, in training the parameters of our method (see 3.1.2) we avoided penalizing predictions that overlap annotated CRMs but do not hit the exact boundaries of the annotations. Thus, for a CRM prediction that hit an annotated CRM, the number of false positive nucleotides outside of the boundary of the hit annotation were not used to calculate PPV' or SP. Likewise, nucleotides in annotated CRMs outside the boundary of an overlapping prediction were not counted as false negatives in calculating SN. To evaluate the performance of CisPlusFinder in comparison with other methods or on the test set we used the script by Chan and Kibler (2005) without modification. This script considers a CRM to be discovered if the overlap between it and a predicted CRM exceeds 50 bp and therefore excludes CRMs shorter than 50 bp from the analysis. Our analysis follows the same rules. We note that exons were not masked in this study, so as not to exclude prediction of CRMs that are known to overlap exons, and for direct comparison with results from Chan and Kibler (2005).

### 2.3 Datasets

To develop and test our algorithm we used three datasets containing CRMs from genes involved in early development in *D.melanogaster* (Table 1).

To develop the CisPlusFinder method and train its parameters we used the combined datasets from Papatsenko et al. (2002) and Papatsenko and Levine (2005).

To compare results from CisPlusFinder with other CRM detection methods, we used the dataset and published results of Chan and Kibler (2005). We note that HexDiff dataset shares 12 genes in common with the Papatsenko dataset, and therefore this evaluation dataset is not independent from our training dataset, as is the case for many of the PWM-based methods tested in Chan and Kibler (2005). We also note that the CRM annotation for the HexDiff and Papatsenko datasets differ for these 12 genes, with 34 of 44 annotated HexDiff CRMs overlapping 35 of 41 annotated Papatsenko CRMs.

To test our method on a dataset of CRMs that is fully independent from the training data and that covers a wider range of biological processes, we used the HexDiff (Chan and Kibler, 2005) and the REDfly (Papatsenko et al., 2005), many of which are positioned in CRMs from the HexDiff dataset.

### 3 RESULTS

#### 3.1 Training

#### 3.1.1 Choice of informant species

To train the parameters of CisPlusFinder we attempted not to miss CRMs because they are absent from one or more informant sequences. Therefore, we selected manually the informant species from the set of *Drosophila yakuba*, *Drosophila ananassae*, *Drosophila pseudoobscura* and *Drosophila viridis*. Table 1 in the Supplementary materials indicates which informant species were used for which target gene.

#### 3.1.2 Parameters for *Drosophila*

For *D.melanogaster* and the informant species, we systematically varied the following five parameters, and measured the accuracy (PPV') on the training dataset. For other species the parameters can easily be retrained.

We found the following combination to be optimal:

1. (1) length of overrepresented core of a PLUS: \( c = 5 \)
2. (2) threshold for the minimal O/E-ratio: \( t = 2 \)
3. (3) window size for measuring overrepresentation: \( w = 1000 \)
4. (4) window size for measuring regulatory potential: \( r = 500 \)
5. (5) minimal regulatory potential: \( m = 3 \)

#### 3.1.3 Accuracy on the training set

With optimized parameters we found 94 of the known 95 CRMs in the Papatsenko dataset. Only the CRM regulating the gene zen in our prediction, which lacks a third PLUS to reach the required regulatory potential. Sensitivity on CRM level of our method is 0.99 and PPV' is 0.67. Specificity on nucleotide level is 0.60 and sensitivity on nucleotide level is 0.56.

#### 3.2 Comparison with other methods on HexDiff dataset

To compare the performance of CisPlusFinder with other methods, we applied it to a dataset used by Chan and Kibler (2005). In this study the methods Ahab (Rajewsky et al., 2002), ClusterBuster (Frith et al., 2003), MSCAN (Johansson et al., 2003), MCAST (Bailey and Noble, 2003), LWF (Papatsenko et al., 2002) and HexDiff (Chan and Kibler, 2005) were applied to a set of CRMs active in early *Drosophila* development and performance measures of these methods were computed. The results are summarized in Table 2; for details on the parameters used for these method we refer the reader to the paper by Chan and Kibler (2005). We note that the thresholds chosen by Chan and Kibler (2005) were optimized to

<table>
<thead>
<tr>
<th>Table 1. Summary of CRM datasets used in this study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Papatsenko</td>
</tr>
<tr>
<td>Number of target genes</td>
</tr>
<tr>
<td>Number of CRMs</td>
</tr>
</tbody>
</table>
Table 2. Comparison of different CRM prediction methods on the HexDiff dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>SN (CRM)</th>
<th>PPV (CRM)</th>
<th>SP (nuc.)</th>
<th>PPV (nuc.)</th>
<th>Hit CRMs</th>
<th>Missed CRMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>HexDiff</td>
<td>69</td>
<td>34</td>
<td>94</td>
<td>39</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Ahab</td>
<td>44</td>
<td>58</td>
<td>98</td>
<td>22</td>
<td>55</td>
<td>23</td>
</tr>
<tr>
<td>ClusterBuster</td>
<td>60</td>
<td>26</td>
<td>95</td>
<td>34</td>
<td>37</td>
<td>31</td>
</tr>
<tr>
<td>MSCAN</td>
<td>65</td>
<td>19</td>
<td>91</td>
<td>27</td>
<td>21</td>
<td>34</td>
</tr>
<tr>
<td>MCAST</td>
<td>83</td>
<td>11</td>
<td>70</td>
<td>48</td>
<td>13</td>
<td>43</td>
</tr>
<tr>
<td>LWF</td>
<td>52</td>
<td>11</td>
<td>90</td>
<td>13</td>
<td>11</td>
<td>27</td>
</tr>
<tr>
<td>Stubb (default)</td>
<td>67</td>
<td>24</td>
<td>92</td>
<td>29</td>
<td>24</td>
<td>35</td>
</tr>
<tr>
<td>Stubb (vir-PhastCons)</td>
<td>87</td>
<td>11</td>
<td>71</td>
<td>43</td>
<td>12</td>
<td>45</td>
</tr>
<tr>
<td>ecis-analyst</td>
<td>61</td>
<td>65</td>
<td>92</td>
<td>45</td>
<td>33</td>
<td>32</td>
</tr>
<tr>
<td>CisPlusFinder</td>
<td>96</td>
<td>11</td>
<td>54</td>
<td>60</td>
<td>11</td>
<td>50</td>
</tr>
<tr>
<td>PFR</td>
<td>23</td>
<td>36</td>
<td>83</td>
<td>10</td>
<td>5</td>
<td>12</td>
</tr>
</tbody>
</table>

Accuracy of different CRM prediction methods on a set of 52 CRMs active in early Drosophila development. (Results in the top section of the table are taken from hexdiff).

3.3 Analysis of false positive predictions

Table 2 also shows a relatively low specificity achieved by CisPlusFinder. However it is obvious from Figure 2 that there are predictions made by CisPlusFinder that do not overlap with the HexDiff-based annotation (black boxes) but do correspond to annotated CRMs in the REDfly database (grey boxes). To see if the low specificity of CisPlusFinder results from an incompletely annotated evaluation dataset, we extracted all CRMs from the REDfly database that regulate genes in the HexDiff dataset. The REDfly database contains 125 CRMs for the HexDiff regions in contrast to 52 CRMs currently annotated, and also has improved annotations for hairy and oc. A total of 120 of the 125 REDfly CRMs overlap with our predictions, and 5 of the annotated CRMs cannot be found. The number of false positive CisPlusFinder CRM predictions is reduced from 408 to 338. The number of nucleotides in false positive CisPlusFinder CRM predictions is reduced by 15.41% from 296 005 to 250 388.

The current REDfly-based annotation and bed files for upload to the UCSC genome browser for all 16 HexDiff regions containing all the information can be downloaded at the project page.

3.4 Correlation between PLUSs and known TFBS

To test whether PLUSs are caused by TFBSs, we downloaded all annotated TFBSs that are found in the HexDiff regions from the flyreg database (Bergman et al., 2005). This resulted in a dataset of 376 TFBSs regulating 10 target genes bound by 27 different transcription factors. For 67 of the TFBSs, the binding factor is not known. The average length of the TFBS in our subset is 18 bp.

A total of 135 TFBS (36%) overlap with 115 PLUSs found by CisPlusFinder. The average length of a hit TFBS is 19 bp and the average length of a hitting PLUS is 14 bp. In the 602 508 bp analysed sequence 4243 PLUSs are found, covering 8.65% of the sequence. The 376 TFBSs cover 1.07% of the sequence. The hit TFBSs cover 0.39% of the sequence. The probability to have this overlap by chance is 3.12 × 10⁻⁴. These results indicate that while many PLUSs correspond to TFBSs, many TFBSs do not appear in PLUSs because of the degree of substitution and turnover in this species set (Emberly et al., 2003; Ludwig et al., 2000).

3.5 Application to simulated data

To investigate our predictions in unconstrained sequences we applied it to simulated data. We simulated the evolution of the 16 HexDiff sequences as ancestral sequences along the tree shown in the supplementary material using the method CisEvolver by Pollard et al. (2006) (http://rana.lbl.gov/CisEvolver/), which can evolve sequences according to the HKY85-model (Hasegawa et al., 1985) without any selective constraint. We extracted the simulated sequences of D. melanogaster and the informant sequences D. yakuba, D. ananassae, D. pseudoobscura and D. virilis from the generated sequences and applied CisPlusFinder to it. CisPlusFinder did not predict any CRM in the simulated D. melanogaster sequences. Out of 642 508 bp of the simulated D. melanogaster sequence roughly 60% are covered by chained alignments. No significant PLUS can be found in these aligned sequences. We conclude that PLUSs occur preferentially in sequences, which are subject to functional constraint.

3.6 Application to the REDfly test set

3.6.1 Accuracy on the REDfly test set

The results of CisPlusFinder applied to the REDfly CRMs are shown in Table 3. Our method predicts CRMs that overlap 289 of 386 annotated CRMs in the REDfly dataset. We compared our results with the output of Stubb using the same set of PWMs as above, which may limit Stubb to predict CRMs regulated by transcription factors regulating the CRMs in the Chan and Kibler (2005) regions. When D. virilis is used as a reference species Stubb predicts CRMs that overlap 236 annotated REDfly CRMs.
3.6.2 Further analysis of false negatives

A total of 97 CRMs are missed by CisPlusFinder in the REDfly regions. Assuming that the function of a CRM is conserved in all informant species, it can be missed by CisPlusFinder for two reasons. One reason is that the number of PLUSs in this region is lower than the minimal required regulatory potential. The other reason is that the local substitution rate of a region is very high, which means that the probability that a TFBS will cause a PLUS is reduced.

To understand the basis of false negative predictions, we define a missed CRM as those which do not overlap at all with any CisPlusFinder prediction. One missed CRM that overlap with CisPlusFinder predictions for <50 bp and were treated as false negatives above, are treated as true positives in this analysis. For 13 missed CRMs, the informant sequences contain ambiguous or unassembled sequences, i.e. stretches of Ns. The very strong requirement of perfect ungapped alignment makes CisPlusFinder very sensitive to unassembled sequences and Ns in the sequence. For this reason six missed CRMs with a N-content above 60% are excluded from further consideration. After excluding these 16 exceptions, 81 false negative CRMs are left for further analysis.

To find which CRMs are missed because they do not contain enough PLUSs to give rise to the required regulatory potential, we chose the subset of missed CRMs that contain putative PLUSes but are not predicted by CisPlusFinder (named lowRP in the following). lowRP contains 31 missed CRMs from 26 different regions. The remaining 50 missed CRMs contain no PLUS and map to 30 different regions. We speculated that these missed CRMs in no PLUS regions might be caused by a high local substitution rate.

To test this hypothesis, we downloaded the PhastCons scores for multiple alignments of the D. melanogaster genome with eight other
4 DISCUSSION

Transcriptional regulation is a highly complex process that differs for each individual gene and CRM. Therefore, it is difficult to find a necessary and sufficient set of rules that can be used for computational prediction of all CRMs in a genome.

The goal of this project was to develop a method that finds CRMs independent of the information about the transcription factors that regulate a target gene. To do this we have developed a CRM prediction method based solely on sequence conservation and composition. Conservation-based methods limit prediction to conserved CRMs, but this limitation is becoming less restrictive with the increasing number of completely sequenced genomes. The set of informant species can be chosen according to the conservation of regulation for each investigated gene.

Table 2 shows that CisPlusFinder is able to localize Drosophila CRMs better than the other methods applied thus far, which were also trained and developed with CRMs involved in early Drosophila development. The results of CisPlusFinder applied to the REDfly dataset (Table 3) show that CisPlusFinder is not limited to a special group of CRMs, and that our method can predict CRMs that are involved in biological processes different from those in the training data. This generality of CisPlusFinder results from the fact that it does not require any a priori knowledge about the binding transcription factors regulating a target gene.

However, we found that the specificity of our method is the lowest in the comparison to those that condition predictions on TFBS models. To test if CisPlusFinder predicts a high number of false positives or, more interestingly, a high number of undiscovered CRMs, we compared our prediction with an updated annotation of the HexDiff regions (section 3.3). These results clearly indicate that the low specificity of CisPlusFinder is in part because we are predicting unannotated CRMs in the HexDiff regions, underscoring both the importance of high quality CRM annotations as well as the difficulty in evaluating the specificity of CRM prediction methods.

The authors would like to thank R. Nunes da Fonseca for valuable discussion, F. Catania for the name ‘PLUS’ and A. Siepel for sharing the phylogeny. The authors also thank three anonymous reviewers for many helpful comments on the manuscript. N.P. is supported by a Marie Curie visiting fellowship from the European Commission (HPMT-GH-01-00285-13).

Conflict of Interest: none declared.
REFERENCES


