USE OF AN INTEGRATED RULE SYSTEM FOR IDENTIFICATION OF THE TRANSCRIPTION FACTOR BINDING SITES FOR MCM1 AND FKH2 IN S. CEREVISIAE

Walker N.J.*1, 2,  Sharrocks A.D.1,  Attwood T.K.1, 2

1 School of Biological Sciences, University of Manchester U.K.; 2 Department of Computer Science, University of Manchester U.K.
* Corresponding author: e-mail: walker@bioinf.man.ac.uk

Keywords: Mcm1, Fkh2, Saccharomyces cerevisiae, transcription factor, binding site, composite elements, comparative genomics

Summary

Motivation: To better understand the co-ordinate activity of two transcription factors, Mcm1 and Fkh2, we wish to study the sequence variation in their composite binding sites. This, we hope, will improve the in silico location of their sites on a genome wide scale. In particular, our approach is to avoid the inherent bias of sparse weight matrices towards strong (less variable) binding sites. We follow a strategy that seeks to take the emphasis off the sequence itself, and instead, produce a set of putative binding sites through the application of rules based on distances between composite motifs, conservation within closely related species, and gene annotation.

Results: A system has been developed that integrates yeast species alignment data, annotation data and genome wide location analysis experiment data. This allows formulation of several rules, giving a set of potential binding sites that include all generally accepted composite Mcm1 and Fkh2 binding sites. With this approach several further potential binding sites are uncovered.

Introduction

The transcription factors Mcm1 and Fkh2 have been implicated in the co-ordinated regulation of genes involved in G2-M phase of the cell cycle. It is interesting to use bioinformatics approaches to further assess the binding location of these transcription factors and test the boundaries of this hypothesis. A recent review by Wasserman and Sandelin (2004) has suggested reasons why a computational search model, based only on functional transcription factor binding sites, leads to many false positives. They note that there are often limited functional binding site sequences available to parameterise the model, and that such sites may include a significant amount of variation, making them difficult to identify. Such is the case with Mcm1 and Fkh2. Core sequences from known Mcm1 binding sites are variable (Shore, Sharrocks, 1995), while partner Fkh2 binding sites are less variable but are shorter.

Methods and Algorithms

Our approach is to try to capture weaker sites, which are not well represented in the known set of transcription factor binding sites. To achieve this, the following steps are employed:

1. Consensus sequences:
Consensus sequences with mismatches are known to be problematic for obtaining good specificity (Stormo, 2000); however we use them here to establish a large amount of variation in the represented binding sites. Composite sites, are retrieved that match the wide consensus pattern with one mismatch.

2. Application of Rules:
   a) Spacing between composite motifs: based on information in Boros et al. (2003). This parameter is set to vary between 3 and 20 base pairs.
b) **Conservation between closely related species**: the average conservation within the site, over the average conservation over 50 base pairs surrounding the site should be greater than 0.9.

c) **Annotation markers**: results can then be enhanced or further specified by indicating key words from annotation associated with a gene. In particular, marking genes that are known to be cell cycle regulated is useful for these particular transcription factors.

3. Comparison with Genome Wide Location Analysis (GWLA) Experiment:

The resulting set of potential sites is analysed against joint p-value scores for Mcm1 and Fkh2 transcription factor binding sites produced by a GWLA experiment (Lee *et al.*, 2002).

**Implementation and Results**

Execution of the method requires a program to produce all matching instances and their positions, and a relational database combining, this data with promoter regions over *S. cerevisiae, S. bayanus, S. paradoxus,* and *S. mikitae* (Kellis *et al.*, 2003) and annotation data gained www.yeastgenome.org, GWLA data.

Based on these rules, the set of matches is greatly reduced (Table 1), making it possible to study the effect of different parameters on known and unknown sites.

Comparison of the joint probability of Mcm1 and Fkh2 binding together on a promoter given by GWLA reveals that 30 of these 38 sites occur in promoters identified to have a joint p-value of less than 0.05 (see Table 2).

<table>
<thead>
<tr>
<th>Step</th>
<th>No binding site matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>MCM1 ~ 30000 FKH2 ~ 50000</td>
</tr>
<tr>
<td>2. Rule a</td>
<td>~ 8000</td>
</tr>
<tr>
<td>2. Rule b</td>
<td>260</td>
</tr>
<tr>
<td>2. Rule c</td>
<td>38</td>
</tr>
</tbody>
</table>

All sites that were strongly identified by Simon *et al.* (2001) are included in these results. Without filtering on genes involved in the cell cycle, at a p-value of 0.0062, (the highest p-value of a known independently verified gene in Table 2), the GWLA of Lee *et al.* (2002) returns 51 sites. Without any rules applied to potential sequences in the promoter, 441 promoter regions are returned at this p-value, clearly involving many false positives. At 0.02, the lowest p-value in Table 2, the figures are respectively, 136 and 689 sites. This shows that Mcm1 and Fkh2 do bind cooperatively to weak sites, and underscores the problem of identifying such weak sites on a genome wide scale.

The system can be easily extended to employ further rules, such as conditional rules on nucleotides within the sequence, and rules based on alignment gaps within a sequence. However, after application of rules a to c, they lead only to trivial reduction in the resulting data set.

**Discussion**

We show that by emphasising non-sequence characteristics of binding sites using simple rules, it is possible both to capture strong sites, and to suggest putative weaker binding sites. Clearly, this method relies on well chosen initial parameters to achieve reductions in matches with out loss of known functional binding sites. This underscores the fact that detection of weak binding sites is problematic for both computational and wet-lab techniques when applied at a genome-wide scale.

We are currently using this rule based system to study variation in binding sites given such rules, ultimately with a view to automating parameter setting. Our analyses will also produce test-sets for discrimination algorithms. Several limiting factors in this approach include the reliance on good multiple alignments of related genomes, and the assumption that most binding sites, whether weak or strong, are conserved in closely related species.
Table 2. 30 putative composite Mcm1, Fkh2 binding sites identified using rules a to c. Underlined ORF codes indicate that the gene or site has previously been implicated in binding by the two transcription factors

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Alias</th>
<th>ORF code</th>
<th>MCM1 motif</th>
<th>FKH2 motif</th>
<th>Joint p-value for MCM1 and FKH2 binding to promoter given by GWLA study</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC20</td>
<td>PAC5</td>
<td>YGL116W</td>
<td>gccgaaaggg</td>
<td>gtaaata</td>
<td>2.4E-18</td>
</tr>
<tr>
<td>SPO12</td>
<td></td>
<td>YHR152W</td>
<td>tcttaatttgg</td>
<td>gtaaaca</td>
<td>1.56E-16</td>
</tr>
<tr>
<td>BUD4</td>
<td></td>
<td>YJR092W</td>
<td>acccgatttgg</td>
<td>gtaaaca</td>
<td>1.078E-15</td>
</tr>
<tr>
<td>SWI5</td>
<td></td>
<td>YDR146C</td>
<td>acctgtttagg</td>
<td>gtaaaca</td>
<td>2.016E-15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YJR051W</td>
<td>tctttttttgg</td>
<td>gtaaaca</td>
<td>5.27E-15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YJR051W</td>
<td>tctttttttgg</td>
<td>gtaaaca</td>
<td>5.27E-15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YLR190W</td>
<td>tcccaacggg</td>
<td>ataaaaa</td>
<td>1.558E-13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YLR190W</td>
<td>accttttttgg</td>
<td>gtaaaca</td>
<td>1.558E-13</td>
</tr>
<tr>
<td>CLB2</td>
<td></td>
<td>YPR119W</td>
<td>acgggaatcgg</td>
<td>gtaaaca</td>
<td>8.84E-13</td>
</tr>
<tr>
<td>SUR7</td>
<td></td>
<td>YML052W</td>
<td>acceccatcgg</td>
<td>gtaaaca</td>
<td>9.96E-13</td>
</tr>
<tr>
<td>SUR7</td>
<td></td>
<td>YML052W</td>
<td>ccccaatcggg</td>
<td>gtaaaca</td>
<td>9.96E-13</td>
</tr>
<tr>
<td>UTH1</td>
<td></td>
<td>YKR042W</td>
<td>tcctagttcg</td>
<td>gtaaata</td>
<td>9.36E-12</td>
</tr>
<tr>
<td>KIN3</td>
<td>FUN52/NPK1</td>
<td>YAR018C</td>
<td>tcttaattg</td>
<td>gtaaata</td>
<td>8.4E-10</td>
</tr>
<tr>
<td>ALK1</td>
<td></td>
<td>YGL021W</td>
<td>gecctttttag</td>
<td>gtaaaca</td>
<td>1.218E-08</td>
</tr>
<tr>
<td>DBF2</td>
<td></td>
<td>YGR092W</td>
<td>gctcttttagg</td>
<td>gtaaata</td>
<td>6.12E-07</td>
</tr>
<tr>
<td>DBF2</td>
<td></td>
<td>YGR092W</td>
<td>gecctttttag</td>
<td>gtaaata</td>
<td>6.12E-07</td>
</tr>
<tr>
<td>NCE102</td>
<td>NCE2</td>
<td>YPR149W</td>
<td>t--ctttttata</td>
<td>gtaaata</td>
<td>0.00000273</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YNL058C</td>
<td>tcttaaagg</td>
<td>gtaaata</td>
<td>4.361E-06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YNL058C</td>
<td>acccttaata</td>
<td>ggaata</td>
<td>4.361E-06</td>
</tr>
<tr>
<td>PMA1</td>
<td></td>
<td>YGL008C</td>
<td>acctttctagg</td>
<td>gaaaca</td>
<td>0.00000492</td>
</tr>
<tr>
<td>PMA1</td>
<td></td>
<td>YGL008C</td>
<td>gctcttaaag</td>
<td>gtaa--atc</td>
<td>0.00000492</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YMRS25C3</td>
<td>tcttaactagg</td>
<td>gtaaaaa</td>
<td>0.0000976</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YMRS25C3</td>
<td>tcttaactag</td>
<td>gtaaaaa</td>
<td>0.0000976</td>
</tr>
<tr>
<td>BNS1</td>
<td></td>
<td>YGR230W</td>
<td>tcttaaaagg</td>
<td>gtaatta</td>
<td>0.000414</td>
</tr>
<tr>
<td>BNS1</td>
<td></td>
<td>YGR230W</td>
<td>atcttaaaag</td>
<td>gtaatta</td>
<td>0.000414</td>
</tr>
<tr>
<td>MYO1</td>
<td></td>
<td>YHR023W</td>
<td>gecctttttagg</td>
<td>gtaaaca</td>
<td>0.0013</td>
</tr>
<tr>
<td>SUT1</td>
<td></td>
<td>YGL162W</td>
<td>t--cttaacgg</td>
<td>gtaaaca</td>
<td>0.00224</td>
</tr>
<tr>
<td>HOF1</td>
<td>CYK2</td>
<td>YMRS32W</td>
<td>tctcttttgaa</td>
<td>gtaaaca</td>
<td>0.0062</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YPR013C</td>
<td>gctcttttggg</td>
<td>gtaaaca</td>
<td>0.02553</td>
</tr>
</tbody>
</table>

Acknowledgements

We are grateful to the Wellcome Trust for individual support (NW).

References


