BioProspector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-Expressed Genes


Presentation by Jay Shukla, Alexandre Ismail & Kate Stafford

Mentor: Dr. Hugh Nicholas
Pittsburgh Supercomputing Center
Overview

• Purpose for Paper
• Explanation of Algorithm & Options
• Experiments
• Results
• Discussion
• My Research
Purpose for Developing BioProspector

• “To look for regulatory sequence motifs in the upstream region genes in the same expression group”

• Using this information, it is possible to:
  – See differences in motifs among organisms
  – Learn which motifs control what actions
  – Learn more on developmental/environmental changes which can trigger a change in the activity of the gene
Local Multiple Sequence Alignment

• Today’s programs combine two statistic-based theories
  – Expectation-Maximization [EM] – Sequences are known but patterns are unknown
  – Stochastic Sampling – Reduces amount of data to be searched
Expectation-Maximization

- Belief is that a good motif is easy to recognize
- Good motif will be highly informative

\[ I = \sum_j \sum_i q_{i,j} \left( \log_2 \frac{q_{i,j}}{p_i} \right) \]
Stochastic Sampling

• Randomly selects a segment from each sequence but one
• Compares each segment to the left out sequence to score each segment
• Score is then used to randomly select one of the segments
• Process is repeated until left with the most represented pattern
Stochastic Sampling: Example

Randomly select a sequence segment from every sequence but one to make into a pattern and evaluate the information content of the pattern.

Cyc_Orsya: \[\text{ln[pkkyi]PGTKMvfpglk}\]
Cyc_Tepy: \[\text{hvPGTKMafaglp[adkdr]ad}\]
Cy2_Argtc: \[\text{kkiPGN[KMala]giskpeeldn}\]
Cy2_Rhoru: \[\text{fvleksgdpkAKSKMtfkltkd}\]
C550_Parde: \[\text{pw1[vkmt]dkGATKMtfkm}\]
C550_Nitwi: \[\text{pkakvPGTKMv[fagik]kdsel}\]
Refine the Pattern: Picking the next word

Early scores and selection of the next word to add to the pattern

- Sum the scores to create a scale from zero to the total of the scores
- Pick a random number between zero and the total
- The interval on the scale containing the random number marks the word to be included in the pattern for the next cycle

Late scores and selection of the next word to add to the pattern
Basic Model

• Parameters for algorithm include:
  – $F_{in}$ = Input of ‘n’ DNA sequences
  – $F_{bg}$ = Fraction of the bases in the sequence $[P_i]$
  – $W_1$ & $W_2$ = Width of two motif blocks
  – $G_L$ & $G_M$ = Gap Range

• Additional Inputs
  – Whether the sequence has a copy of the motif
  – Whether the motif can occur in both strands
  – Whether the motif is palindromic [$W_1 = W_2$]
Figure 1. A graphical illustration of the model used for BioProspector. Suppose the input file consists of $N$ DNA sequences, each containing $0$-$n$ copies of a sequence motif. The motif has two binding blocks of width $w_1$ and $w_2$, respectively, which are separated by a gap of variable length ranging from $g_L$ to $g_M$. 
Markov Model

• Utilized by other applications
  – However, 3rd-order just recently fully developed

• When calculating $P_i$, the Markov model will look at the three bases prior to the base being examined, and then calculate
Algorithm Output

• Motif
  – Motif Score
  – Significance Value
  – # of Aligned Segments

• DNA Sequences
  – # of Segments in Each Sequence
  – Starting and Ending Position of Segment
  – Sequence of Segment
Extra Options

• Threshold $[T_H \& T_L]$ used to deal with:
  – Sequences with no copies of the motif
  – Sequences with multiple copies of the motif

• Probability matrix $[\theta_1]$ used to deal with palindromic motifs

• Quality of Motif

Motif Score = $\#seg \times \exp\{\sum_{\text{all positions}} \sum_{\text{all nucleotides}} q_{i,j} \times \log(q_{i,j}/p_j)\} / w$
How well does this work?

• BioProspector tested in three different scenarios.
  – #1: 60 non-coding sequences were searched for site/motif which interacts with Rap1p
  – #2: 136 sequences were searched for TATA-box which is a two-block motif
  – #3: 18 *Ecoli* sequences were searched for CRP-binding sites (site is known to be palindromic)
Results (Markov)

• 0 order provided poor results
• 1$^{st}$ order provided decent results, but many of the high-scoring motifs were incorrect
• 3$^{rd}$ order provided excellent results. The motifs which were the highest scored were always correct
Results (Two-block)

- BioProspector found 70% of the sites (95/136)
- For the 95 sequences, it correctly identified the motif \([\text{TTGACA TATAAT}]\)
- Sites which were missed differed greatly from motif
- In some cases, BioProspector found a site which matched better than the correct site
Above, the sites found by BioProspector were a better fit than the actual correct site.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Correct site</th>
<th>Site found</th>
</tr>
</thead>
<tbody>
<tr>
<td>ald</td>
<td>AAGAAT TACACT</td>
<td>TTTCCA TAAAA</td>
</tr>
<tr>
<td>cspB</td>
<td>TTGTTC TGGAGT</td>
<td>ATTACT TATTTT</td>
</tr>
<tr>
<td>menE</td>
<td>AATACA GATGAT</td>
<td>TTGAGA TCTTTT</td>
</tr>
<tr>
<td>odhA</td>
<td>TTGTGA CAAATT</td>
<td>TTTACT TAGAAT</td>
</tr>
</tbody>
</table>

Above, BioProspector found more than one site, often the second fit better than the correct site.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Correct site</th>
<th>Second site</th>
</tr>
</thead>
<tbody>
<tr>
<td>abrB</td>
<td>TTGACG TAGTCT</td>
<td>CTGACT TACAAT</td>
</tr>
<tr>
<td>veg</td>
<td>TTGACA TACAAT</td>
<td>TTGACA TATAAT</td>
</tr>
<tr>
<td>$\phi105$</td>
<td>TTTACA TACAAT</td>
<td>TTGACG TACAAT</td>
</tr>
</tbody>
</table>
Results (Palindrome)

- Prior experiments found 24 CRP-binding sites
  - Used a one-block motif
  - Few sites were wrong
- In this experiment, motif was described as a two-block motif
- All sites found were correct, and found a more conserved motif
Discussion

- 3rd Order Markov Model vastly improves the output
- Two-block motif works well when transcription is regulated by a combo of proteins
- Gap allows for better detection rather than forcing deletions to align
- Belief is that 30% of RNA polymerase sites were missed because mechanism is heavily dependent on physical-chemical and structural characteristics
My Research

• Apply same ideas to protein sequences rather then DNA sequences

• Attempting to characterize Human APE/Ref-1 protein family
  – Plays a role in DNA repair and transcription regulation
References


The Pittsburgh Supercomputing Workshop Lectures

Acknowledgements

• Dr. Hugh Nicholas
• Kate Stafford & Alexandre Ismail
• Rajan Munshi
• The Pittsburgh Supercomputing Center
• NIH/NSF & all BBSI students
• The Salvation Army
• Jason and The O for catering services
…and what you’ve all been waiting for

The 2004 BBSI Anthem

Written and produced by Mark Connell, Jay Shukla and Jordan Torok

Edited by Mixmaster Mark

Percussion: Coffee Can Backbeat – Jay; Keys on Keychain – Jordan; Cup of Change – Mark

Guitar: Mark

Vocals: Jordan, Mark and Jay