LOGOS: a modular Bayesian model for 
de novo motif detection

Eric Xing$^1$
Wei Wu$^2$    Michael Jordan$^1$    Richard Karp$^1$

$^1$University of California, Berkeley
{epxing, jordan, karp}@cs.berkeley.edu

$^2$Lawrance Berkeley National Lab
wwu@lbl.gov

13 Aug 2003
The *cis*-regulatory machinery of eukaryotic organisms
A DNA motif (Gcn4)

Motif features:

- PSMD: position-specific multinomial distribution
- PWM: position weight matrix
Meta-sequence features

- **Shape bias**: typical patterns of site-conservativeness due to binding surface constraints.

- **Spatial organization of motifs**: dependencies between motif instances.
  
  — Believed to be crucial in distinguishing biologically meaningful motifs from random background or trivial recurring patterns.

Our goal: Build statistical models for motif-bearing sequences that:

- exploit the meta-sequence features and provide improved sensitivity to biologically plausible motifs

- can generalize easily to various motifs and handle complex \emph{cis}-regulatory structures

- can make use of biologically identified motifs for \emph{de novo} detection
Introduction

- Using a hierarchical Bayesian Markovian model to posit hidden dependency structures of motifs

- Exploit this model to predict locations of novel motifs

- Talk outline
  - Modeling generation of motif instances with a hidden Markov Dirichlet-multinomial model
  - Modeling motif instance distribution with a hidden Markov grammar
  - The LOGOS motif model and a variational EM algorithm for motif detection
  - Experiments
  - Future work
The Local Alignment Model

— for multiple-aligned motif instances
Signature internal structures of motifs

Shape bias:

- **gal4:**
  ```
  CGGCTC
  
  3′-end
  
  5′-end
  ``

- **gal4-permute:**
  ```
  CCGTGC
  
  3′-end
  
  5′-end
  ```

- **pho4:**
  ```
  cACGTG
  
  3′-end
  
  5′-end
  ``

- **pho4-permute:**
  ```
  GAGTCC
  
  3′-end
  
  5′-end
  ```

Two yeast motifs and their column-permuted versions.

- The conservativeness of sites in the motif may be spatially dependent.
  - The conserved sites are more likely to occur consecutively and possibly followed (or preceded) by heterogeneous sites that are also consecutive.
  - The "shape" is only associated with the conservativeness pattern of a motif PWM, but **not** with any specific consensus sequence of the motif.
A standard motif alignment model

The Product Multinomial (PM) Model:

\[ A = \{ A_1, A_2, \ldots, A_M \} \]

- The nucleotide distributions at different sites within the motif are assumed to be independent.
- Limitation:
  - Unable to capture site dependence and shape preference in motifs
The PSMD simplex and the Dirichlet Distribution

Each possible position-specific nucleotide distribution corresponds to a point in the simplex.
A Markov model for PSMD-simplex sequence

The PSMD-simplex sequence defines a structural prior for the PWM.
Parameters (for all motifs):
- \( I \) sets of 4-dimensional Dirichlet parameter \( \alpha = \{\alpha_1, \ldots, \alpha_I\} \).
- Initial probability \( \pi \) and transition probability \( B \) of an HMM.

Random variables (for each motif site):
- The \textit{site prototype indicator} indexing different distributions on the PSMD simplex is represented by a multinomial r.v. \( s_I \).
- The \textit{PSMD} associated with the site is represented by a Dirichlet r.v. \( \theta_I \).
- The \textit{nucleotide content} of the site is represented as a multinomial r.v. \( A_{m,I} \).
The Global Distribution Model

— for patterns of motif occurrences in genomic sequences
Global organization of motif instances

cis-regulatory modules (Davidson [2001]):

- There exist instance (resp. cluster)-level dependencies between motifs (resp. modules).
The *uniform and independent* (UI) start-position model:

- The locations of motif instances are assumed to be independently sampled from a uniform distribution over possible positions.

- Limitations:
  - instances can overlap
  - does not capture inter-instance, inter-motif dependences
A hidden Markov model for motif occurrences

The state-transition grammar of the global HMM:

- **Latent states:**
  - all possible sites within motifs on the forward and reverse complementary strand
  - intra-cluster backgrounds
  - inter-cluster (global) background

- **Observed output:** unannotated sequence of nucleotides.

- **Parameters:** HMM parameters \( \{\pi_g, B_g\} \); motif PWMs \( \theta \); background frequencies \( \theta_b \).
A modular motif model for de novo motif detection

- The integrated LOcal and GLObal motif Sequence model
  - The occurrences of motifs in a DNA sequence, as indicated by $x$, are governed by a global distribution model $p_g(x|\Theta_g)$.
  - For each type of motif, the nucleotide sequence pattern shared by all its instances admits a local alignment model $p_l(A(x,y)|x,\Theta_l)$.
  - Assume that the background non-motif sequences are modeled by a simple conditional model, $p_b(y - A(y,x)|x,\theta_b)$.

- The likelihood of a regulatory sequence $y$ is:

$$p(y|\Theta) = \sum_x p_g(x|\Theta_g)p_l(y|x,\Theta_l)$$

$$= \sum_x p_g(x|\Theta_g)p_l(A|x,\Theta_l)p_b(y - A|x,\theta_b)$$
Inference and parameter estimation

- The inference and learning problems:
  - Maximum a posteriori (MAP) prediction of motif locations
  - Bayesian estimation of motif PWMs

A variational EM algorithm for LOGOS

**Variational “E” step:** Compute the $\langle h \rangle$, the expected count of each nucleotide, for each motif and each site within it, via inference in the global motif model given $\langle \ln \theta \rangle$ and sequence $y$.

**Variational “M” step:** Compute the Bayesian estimation of the PWM, in natural parameter form $\langle \ln \theta \rangle$, via inference in the local motif alignment model given $\langle h \rangle$. 
Experiments:

- Semi-synthetic sequence: random background (300–600bp) with genuine motif instance(s) planted at known positions.
- Real genomic sequences: 500-5000 bp fragments in the genomic sequences flanking the motif instances.
Detecting a single motif using HMDM model

- Result: different HMDMs can be used for different “classes” of motifs.
  - can use $p$-value to identify best-performing model.

<table>
<thead>
<tr>
<th></th>
<th>abf1</th>
<th>gal4</th>
<th>gcn4</th>
<th>gcr1</th>
<th>mat</th>
<th>mcb</th>
<th>mig1</th>
<th>crp</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMDM(f)</td>
<td>0.81</td>
<td>0.82</td>
<td>0.71</td>
<td>0.65</td>
<td>0</td>
<td>0</td>
<td>0.73</td>
<td>0.58</td>
</tr>
<tr>
<td>HMDM(e)</td>
<td>0.77</td>
<td>0.89</td>
<td>0.63</td>
<td>0.12</td>
<td>0.67</td>
<td>0</td>
<td>1.00</td>
<td>0.54</td>
</tr>
<tr>
<td>HMDM(b)</td>
<td>0.12</td>
<td>0</td>
<td>0</td>
<td>0.59</td>
<td>0.92</td>
<td>0.72</td>
<td>0.91</td>
<td>0.04</td>
</tr>
<tr>
<td>PM</td>
<td>0.14</td>
<td>0</td>
<td>0</td>
<td>0.41</td>
<td>0.79</td>
<td>0.13</td>
<td>0.64</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Detecting multiple motif instances using LOGOS

- We compare three variants of LOGOS, ordered with decreasing model expressiveness,
  - HMDM+HMM \( \text{LOGOS}_{hh} \)
  - PM+HMM \( \text{LOGOS}_{ph} \)
  - PM+UI \( \text{LOGOS}_{pu} \)

<table>
<thead>
<tr>
<th>motif name</th>
<th>( \text{LOGOS}_{hh} )</th>
<th>( \text{LOGOS}_{ph} )</th>
<th>( \text{LOGOS}_{pu} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>abf1</td>
<td>0.31 0.21</td>
<td>0.68 0.20</td>
<td>0.79 0.91</td>
</tr>
<tr>
<td>gal4</td>
<td>0.16 0.16</td>
<td>0.19 0.15</td>
<td>0.29 0.79</td>
</tr>
<tr>
<td>gcn4</td>
<td>0.18 0.24</td>
<td>0.61 0.28</td>
<td>0.0 0.96</td>
</tr>
<tr>
<td>gcr1</td>
<td>0.20 0.21</td>
<td>0.34 0.20</td>
<td>0.33 0.94</td>
</tr>
<tr>
<td>mat</td>
<td>0.07 0.03</td>
<td>0.36 0.0 0</td>
<td>0.50 0.96</td>
</tr>
<tr>
<td>mcb</td>
<td>0.37 0.09</td>
<td>0.36 0.08</td>
<td>0.33 0.94</td>
</tr>
<tr>
<td>mig1</td>
<td>0.08 0</td>
<td>0.09 0</td>
<td>0.98 0.10</td>
</tr>
<tr>
<td>crp</td>
<td>0.38 0.34</td>
<td>0.27 0.53</td>
<td>0 0.95</td>
</tr>
</tbody>
</table>
## Analyzing Yeast promoter sequences using LOGOS

<table>
<thead>
<tr>
<th>set name</th>
<th>LOGOS</th>
<th>MEME</th>
<th>AlignACE</th>
<th>seq no.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FP</td>
<td>FN</td>
<td>FP</td>
<td>FN</td>
</tr>
<tr>
<td>abf1</td>
<td>0.79</td>
<td>0.65</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>csre</td>
<td>0.44</td>
<td>0.17</td>
<td>0.78</td>
<td>0.50</td>
</tr>
<tr>
<td>gal4</td>
<td>0.13</td>
<td>0.07</td>
<td>0.17</td>
<td>0.29</td>
</tr>
<tr>
<td>gcn4</td>
<td>0.35</td>
<td>0.18</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>gcr1</td>
<td>0.29</td>
<td>0.61</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>hstf</td>
<td>0.86</td>
<td>0.55</td>
<td>0.60</td>
<td>0.56</td>
</tr>
<tr>
<td>mat</td>
<td>0.42</td>
<td>0</td>
<td>0.38</td>
<td>0.56</td>
</tr>
<tr>
<td>mcb</td>
<td>0.47</td>
<td>0.25</td>
<td>0.20</td>
<td>0.33</td>
</tr>
<tr>
<td>mig1</td>
<td>0.81</td>
<td>0.28</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>pho2</td>
<td>0.90</td>
<td>0.50</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>swi5</td>
<td>0.76</td>
<td>0.50</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>uash</td>
<td>0.83</td>
<td>0.68</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Analyzing *Drosophila* promoter sequences

- Motif patterns derived from biologically identified motif instances (Berman et al. [2002]).

- Motif patterns detected by **LOGOS** in the promoter regions (600-5000bp) of 9 *Drosophila* genes.

- Motif patterns detected using conventional models.
Future work

- Background model: higher order HMM
- Incorporating expression, binding, comparative genomic data
- Analyzing eucaryotic cis-regulatory networks