Gibbs Sampling and Centroids for Gene Regulation

Lee A. Newberg
NY State Dept. of Health Wadsworth Center
@ Albany Chapter
American Statistical Association
11/30/2011
Acknowledgments

Team:
- Sean P. Conlan (National Institutes of Health)
- Travis J. Desell (University of North Dakota)
- Charles E. Lawrence (Brown University)
- Lee Ann McCue (Pacific Northwest Nat'l Lab)
- Thomas M. Smith (MIT Lincoln Laboratory)
- William A. Thompson (Brown University)

Resources:
- Wadsworth Center – including the Laboratory of Computational and Structural Biology Core Facility
- Rensselaer Polytechnic Institute
- Brown University –including the Center for Computational Molecular Biology
- NIH/NHGRI: K25 Mentored Career Award “Quantitative Cross-Species Approaches to Gene Regulation” (LAN)
- DOE: “Bayesian computational approaches for gene regulation studies of bioethanol and biohydrogen production” (CEL, LAM, LAN)
Gibbs sampling and centroids for gene regulation

- Gibbs Sampling
- Centroids
- DNA cis-regulatory elements
- Computational Prediction: Inputs, Outputs
- Results
- Methods
Gibbs Sampling

• Simultaneous sampling of $\theta_1$, $\theta_2$, $\ldots$, $\theta_n$ from posterior distribution $p(\theta_1, \theta_2, \ldots, \theta_n \mid D)$ can be hard.

• But often, if $\theta_2$, $\ldots$, $\theta_n$ are fixed then it isn’t hard to sample $\theta_1$ from

$$p(\theta_1 \mid \theta_2, \ldots, \theta_n, D) = \frac{p(\theta_1, \theta_2, \ldots, \theta_n \mid D)}{\sum_{\theta_1''} p(\theta_1'', \theta_2, \ldots, \theta_n \mid D)}$$

and similarly for each of $\theta_2$, $\ldots$, $\theta_{n-1}$, and $\theta_n$.

• This is often good enough!
Gibbs sampling algorithm

- Repeatedly loop through the parameters, sampling $\theta_1, \theta_2, \ldots, \theta_n, \theta_1, \ldots$ in turn.
- Repeated iterations converge to desired distribution. Keyword: detailed balance
Find the *Centroid* — or why we want the posterior distribution

We focus on the region of solution space containing the most posterior probability, rather than on the single solution that has the most joint probability.

→ Ignore nuisance variables; build the centroid from marginal probability of relevant features.
DNA *cis*-regulatory elements

Importance of DNA cis-regulatory elements

Important … for the understanding of cell function, differentiation, and pathology

… because the elements affect both the products of genes and when and to what extent the genes are expressed

Typically vary species to species, but not individual to individual, except pathologically.

Howard-Ashby, Materna, Brown, Tu, Oliveri, Cameron, & Davidson, *Dev Biol*, 2006

Lee A. Newberg
11/30/2011

Gibbs sampling for Gene Cis-Regulatory Elements
Outputs
- Element sites
- Motif (pattern) description

Input data
- Promoter sequences: aligned when feasible (e.g., 20 sequences × 5 species)
- Phylogenetic tree and model: or ad hoc substitute

Input parameters
- Motif model type: consensus w/ deviations vs. probabilistic
- Motif size: fixed, varying. (6-36 nts.)
- Motif shapes: palindromic, off positions
- Site frequency: per promoter, genome
- Site positioning: nucleosomes, relative to +1
**Effectiveness in simulations**

Verifying known results

<table>
<thead>
<tr>
<th>Motif Models</th>
<th>Possible Sites</th>
<th>Total Predictions</th>
<th>True Positives</th>
<th>False_positives</th>
<th>False Negatives</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>Mean Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>103</td>
<td>57.7</td>
<td>47.3</td>
<td>10.3</td>
<td>55.7</td>
<td>0.82</td>
<td>0.46</td>
<td>0.20</td>
</tr>
<tr>
<td>2</td>
<td>128</td>
<td>74.3</td>
<td>57.3</td>
<td>17.0</td>
<td>53.7</td>
<td>0.77</td>
<td>0.45</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>132</td>
<td>79.3</td>
<td>61.3</td>
<td>18.0</td>
<td>70.7</td>
<td>0.77</td>
<td>0.46</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Novel predictions (by others)

*E.g.: Driscoll et al. (2007) Carbon utilization pathway in Shewanella*
Gibbs sampling for gene regulation

Random walk through posterior probability space. We iterate, re-sampling:

• Missing Data
  – Element sites
  – Tree sequence alignment

• Parameters
  – Motif model logos

Record key features as we walk.
Gibbs sampling needs conditional distributions

- **Plausible, workable, statistical model components:**
  - Position weight matrices for motif models, etc.

\[
Q = \begin{pmatrix}
* & \pi_C & K\pi_G & \pi_T \\
\pi_A & * & \pi_G & K\pi_T \\
K\pi_A & \pi_C & * & \pi_T \\
\pi_A & K\pi_C & \pi_G & *
\end{pmatrix}
\]
Centroids for gene regulation

Determine the centroid set of DNA element sites using marginal probability of each possible site. (The MAP/MLE of the walk is inferior.)
Single CPU / Small Cluster / DNA@home.
Payoff

Plausible statistical model components
+ Gibbs sampling
+ Centroid
→ Robust predictions

• Theory: Newberg et al. (2007), *Bioinformatics*, Pubmed17488758
• Use: Thompson et al. (2007), *Nucleic Acids Research*, Pubmed17483517