A Comparison Of Expectation Maximization and Gibbs Sampling Strategies for Motif Finding

Michele Banko
CSE 527
Final Project
Outline

- Introduction to the Task
- Review of Methods: EM and Gibbs
- Tools, Data, and Evaluation
- Performance Analysis
- Robustness Analysis
- Conclusions
Motif-Finding

- Wish to identify similar subsequences over a set of nucleotide or protein sequences
  - Of any length
  - Having zero or more occurrences per sequence
  - Allowing for insertions/deletion (ideally)

- Two well-studied automated approaches
  - Expectation Maximization (Bailey and Elkan)
  - Gibbs Sampling (Lawrence, et al.)
The EM Approach

- **Input:**
  - n sequences having zero or more instances per sequence
  - The desired length of the motif
  - Background model

- **Model:** a WMM $\theta$ which represents the motif

- **Idea:**
  - If we knew $\theta$, we could find the motif locations
  - If we knew the motif locations, we could compute $\theta$

- **Goal:** Find a $\theta$ such that the log-likelihood of the data is maximized

- Guaranteed to improve after each step, but may get stuck in local optimum
The Gibbs Sampling Approach

- Again, have n sequences
- For each sequence, build a WMM from the remaining sequences, compute probability that the motif starting at a position given what we know about the other sequences
- Maximize ratio of pattern probability relative to the background probability
- Not guaranteed to improve after each iteration
Goals of Evaluation

- Performance
  - How well can each method find the optimal solution?
  - How sensitive is each method to different initializations?
  - How long does the algorithm take to converge?

- Robustness
  - How well can each method cope with noisy data?
  - With small training sets?

- Overall ease of use?
Data

- Use Prosite to extract protein sequences containing 4 known transcription factors present in both the mouse and human species:
  - **Myb 1**, a retroviral oncogene, which has been implicated in regulation of the cell cycle.
  - **Cytochrome P450**, a group of enzymes involved in the metabolism steroids, fatty acids, drugs and carcinogens.
  - **Zinc protease**, a zinc-binding region signature, part of the family of neutral zinc metallopeptidases.
  - **ZF Ring 1**, a zinc finger RING-type signature.
Data

Factors chosen because they possess the following properties:

- Small number of samples (MYB 1)
- Large number of known false positives (MYB 1)
- Large number of known false negatives (Zf Ring 1)
- Several with same motif length (Zf Ring 1, Zinc Protease, Cytochrome P)
- No gaps
Evaluation Metrics

- Site-Level Precision and Recall
  - Precision = \[ \frac{\text{True Positives}}{\text{True Positives + False Positives}} \]
  - Recall = \[ \frac{\text{True Positives}}{\text{Known Instances}} \]
- Best = the motif with the highest recall
- Shift up to w/2 positions in either direction
Implementations

- EM: MEME Toolkit from SDSC
- Gibbs: From Jun Liu
- Strictly off-the-shelf, no modifications to source code
# Quick and Dirty

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gibbs</th>
<th>EM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Myb 1</td>
<td>0.9333</td>
<td>0.9333</td>
</tr>
<tr>
<td>Cytochrome P450</td>
<td>0.9778</td>
<td>0.9778</td>
</tr>
<tr>
<td>Zinc Protease</td>
<td>0.0201</td>
<td>0.0201</td>
</tr>
<tr>
<td>Zf Ring 1</td>
<td>0.9848</td>
<td>0.9848</td>
</tr>
</tbody>
</table>
Initialization: Gibbs

- Gibbs very sensitive to seed values
- Run several independent searches from each starting point
- Zinc Protease motif improvements from $F=0.0201$ to
  - $F=0.9128$ (20 searches with another seed)
  - $F=0.9195$ (50 searches with one seed)
Gibbs over Several Starts and Searches
Initialization: EM

- Insensitive to starting position

- Options
  - Vary fuzziness of sampling function
  - Override start sampling using knowledge of known motif

- Experimented with settings for lowest-performing dataset, found no difference
Seconds to Reach Best Alignment

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gibbs</th>
<th>MEME</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myb 1</td>
<td>2</td>
<td>6.55</td>
<td>3x</td>
</tr>
<tr>
<td>Cytochrome P450</td>
<td>5</td>
<td>33.04</td>
<td>7x</td>
</tr>
<tr>
<td>Zinc Protease</td>
<td>45</td>
<td>225.95</td>
<td>5x</td>
</tr>
<tr>
<td>Zf Ring 1</td>
<td>2</td>
<td>100.23</td>
<td>50x</td>
</tr>
</tbody>
</table>

While Gibbs is relatively faster, time does not account for possible number of restarts needed.
Simultaneous Discovery: Setup

- How well can each algorithm locate several motifs at once?
- One dataset
  - CYTOCHROME + ZINC PROTEASE + ZF RING
  - All Motifs are 9 units long
- Guide the searches, specifying how many instances to expect for each motif
- Several starts/searches for Gibbs
## Simultaneous Discovery: Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Searches</th>
<th>Found Motif</th>
<th>Known Motif</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gibbs</td>
<td>1</td>
<td>MOTIF C</td>
<td>Cytochrome P45</td>
<td>0.0526</td>
<td>0.0111</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>MOTIF C</td>
<td>Zinc Protease</td>
<td>0.0294</td>
<td>0.0076</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>MOTIF A</td>
<td>Cytochrome P45</td>
<td>0.2308</td>
<td>0.0333</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>MOTIF A</td>
<td>Zinc Protease</td>
<td>0.4809</td>
<td>0.4773</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>MOTIF A</td>
<td>Zinc Protease</td>
<td>0.4809</td>
<td>0.4773</td>
</tr>
<tr>
<td>EM</td>
<td>n/a</td>
<td>MOTIF 1</td>
<td>Zf Ring 1</td>
<td>0.9847</td>
<td>0.9773</td>
</tr>
<tr>
<td></td>
<td>n/a</td>
<td>MOTIF 2</td>
<td>Zinc Protease</td>
<td>0.9851</td>
<td>0.8859</td>
</tr>
<tr>
<td></td>
<td>n/a</td>
<td>MOTIF 3</td>
<td>Cytochrome P45</td>
<td>1.0000</td>
<td>0.9556</td>
</tr>
</tbody>
</table>
Small Samples: Setup

- Claim: EM can discover a motif even when as little as 20% of the sequences contain an instance
- Corpus Construction:
  - Randomly select 5% of sequences containing occurrences of the motif.
  - Select the remainder of the sequences at random from the total genome, keeping the entire size of the dataset fixed.
- For 10% known occurrences, select another 5% of the known sequences, ensuring no overlaps with the previous set.
- Add it to the previous set of 5%, and select the remaining 80% at random from the total genomes.
- Do this procedure for up to 20%.
Small Samples: Results

- EM: unable to find any instances of the motif when data has few instances
- Gibbs: Using the best seed value from the previous 3 trials, had at best a precision of 0.1250 and recall of 0.1429, which came when seeing only 5% of actual occurrences.
Conclusions

- EM and Gibbs implementations able to find non-gapped motifs quickly with relative ease
- Gibbs faster, yet may require many trials to find the best alignment
- EM better at finding >1 motif at a time
- Neither method able to cope with noisy data