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 θ which represents the motif
- Idea:
  - If we knew θ, we could find the motif locations
  - If we knew the motif locations, we could compute θ
- Goal: Find a θ 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
  - $\text{Precision} = \dfrac{\text{True Positives}}{\text{True Positives + False Positives}}$
  - $\text{Recall} = \dfrac{\text{True Positives}}{\text{Known Instances}}$
- 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

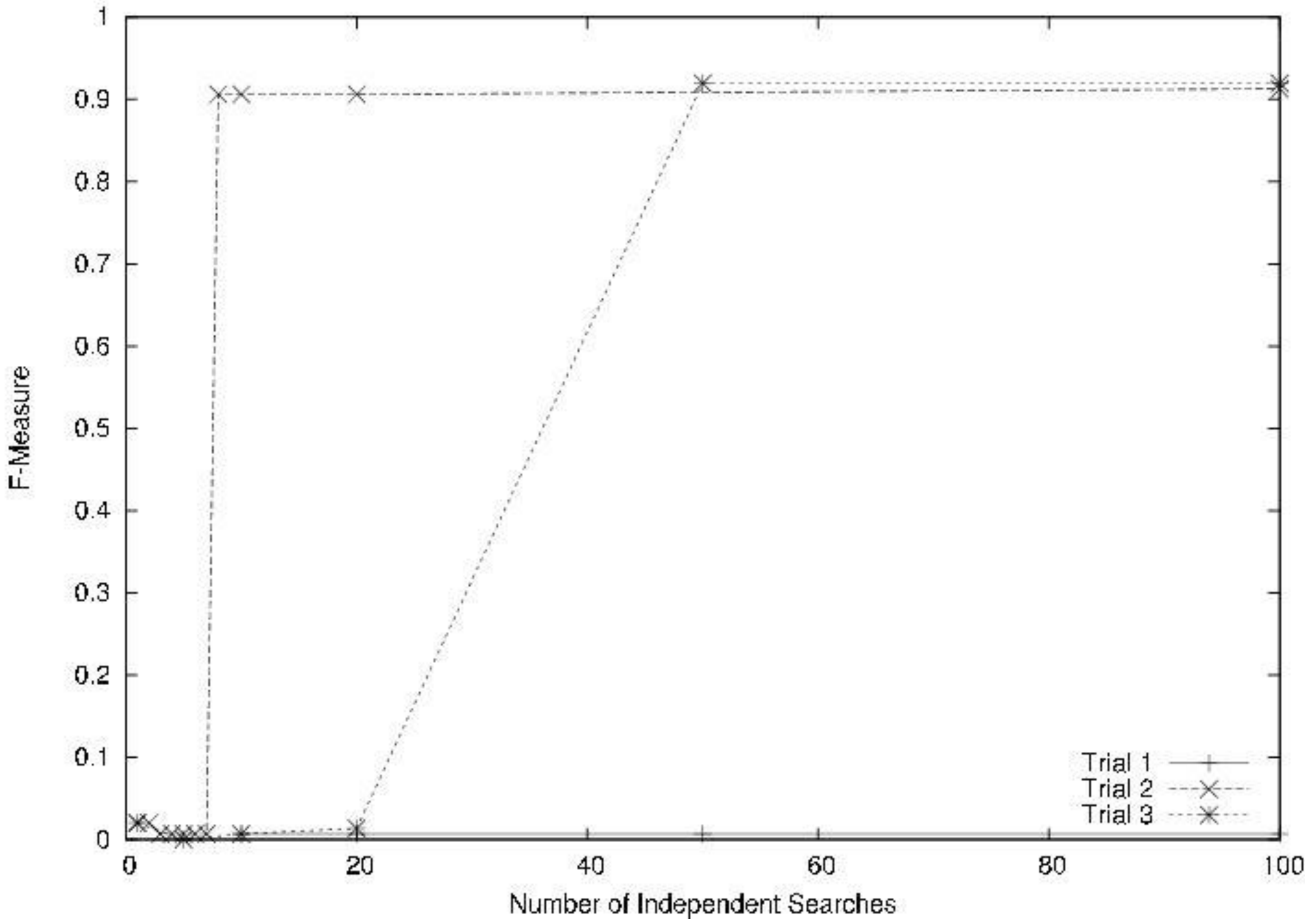
Quick and Dirty

	Gibbs			EM		
Dataset	**Precision**	**Recall**	**Shift**	**Precision**	**Recall**	**Shift**
Myb 1	0.9333	0.9333	0	0.9333	0.9333	0
Cytochrome P450	0.9778	0.9778	1	0.9778	0.9778	1
Zinc Protease	0.0201	0.0201	3	0.9933	0.9933	0
Zf Ring 1	0.9848	0.9848	0	0.9848	0.9848	0

Intialization: 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

Dataset	Gibbs	MEME	Factor
Myb 1	2	6.55	3x
Cytochrome P450	5	33.04	7x
Zinc Protease	45	225.95	5x
Zf Ring 1	2	100.23	50x

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

Method	Searches	Found Motif	Known Motif	Precision	Recall
Gibbs	1	MOTIF C	Cytochrome P45	0.0526	0.0111
	1	MOTIF C	Zinc Protease	0.0294	0.0076
	10	MOTIF A	Cytochrome P45	0.2308	0.0333
	100	MOTIF A	Zinc Protease	0.4809	0.4773
	500	MOTIF A	Zinc Protease	0.4809	0.4773
EM	n/a	MOTIF 1	Zf Ring 1	0.9847	0.9773
	n/a	MOTIF 2	Zinc Protease	0.9851	0.8859
	n/a	MOTIF 3	Cytochrome P45	1.0000	0.9556

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