Finding regulatory modules

December 17, 2003
CSE527 Computational Biology Project Presentation

Raphael Hoffmann
Agenda

• Importance of module discovery
• MCAST
• My Implementation
• Testing on simulated data
• Testing on real data
• Discussion
Importance of module discovery

- Full understanding of gene functions requires understanding the regulatory machinery
- TFBSs are usually small and appear frequently by chance
- True binding sites appear in clusters
MCAST (part of MetaMEME)

MEME → Motifs, PSPMs occurrences → dna sequences

HMM → Viterbi & Scoring Function → EM

MCAST → modules
MCAST – new ideas

- Sophisticated scoring function
- EM
- HMM
My Implementation

MEME → Motifs, PSPMs, occurrences → dna sequences → Linear HMM → Viterbi → modules
My Implementation

• Written in JAVA
• Includes Random Sequence generator (for simulation data according to HMM)
• Linear HMMs
Testing on simulated data

- In real datasets the results depend on a large number of factors, so testing on simulated data first is recommendable.
- I generated a sequence of 16,000 base pairs distributed according to a simple HMM and Drosophila background model (described later).
Testing on simulated data

Log-odds score of sliding window

Background Model

HMM Model
Testing on real data

• Available datasets that contain known modules: Drosophila and human genome
• Motivation: Berman et. al. used a sliding window of 700bp and counted motif occurrences in Drosophila sequences. (Successfull though simple)
Testing on real data

- Drosophila: ~ 20 modules known upstream of the *even-skipped* gene
- Many contain transcription factor binding sites for Bcd, Cad, Hb, Kr, Kni
- This data has been used by many researchers
Testing on real data

- MEME could not correctly identify the transcription factor binding motifs
- I used PSPMs which were identified by a research team and aligned them to known modules using MAST
Testing on real data

MAST alignment of known modules and their transcription factor binding motifs
Testing on real data

- Hardly any common pattern
- But Hb-sites often appear in pairs
- Created a linear HMM
Testing on real data
Discussion

Modules not reliably detectable. Why?

- Only regarded motifs on one strand
- Complete Topology or Star Topology HMM instead of Linear HMM
- Geometric distribution of spacer lengths
- Models not accurate
- Log-odds score
Discussion

- Average log-odds scores were higher on the real data than on the simulated data, across the whole sequence. Why?

Background model which assumes independence might be too simple.
Future Work

• Use EM to train the parameters of the HMM (transition probabilities)