Decoding gene regulation based on DNA accessibility data

Xiaoyu Chen

UW CSE 2009-10
Annual Industrial Affiliates Meeting
Oct 29th, 2009
Transcriptional regulation

Transcription Factor

RNA polymerase

Regulatory Element

Coding region
Chromatin structure

- double helix
- nucleosome
- compaction

Transcription factor complex

http://biomserv.univ-lyon1.fr/baobab/images/epi.jpg
Protein binding footprints

Accessibility tags

DNA

DNA accessible region

Transcription factors

Footprints
Problem definition

• Given
  – T: the number of tags at each position
  – $k_{\text{min}}, k_{\text{max}}$: size range of footprints (8 - 30)

• Return
  – A ranked list of non-overlapping footprints, each associated with a statistical confidence score

Yeast intergenic regions
  – 4.7M bp
  – 13.0M tags
Scoring a candidate footprint

- Assume that tags are uniformly distributed in the background.
- Given a background window of size $b$ containing $B$ tags, the probability that a window of size $a$ contains $x$ or fewer tags:

$$
Pr(S \leq x|H_0) = \sum_{t=0}^{x} \binom{B}{t} \left(\frac{a}{b}\right)^t \left(1 - \frac{a}{b}\right)^{B-t}
$$

Greedy selection

Dynamic Bayesian network

\[
P[T_i \mid S_i] \sim \text{Normal or CPT}
\]

- **Segment**: 
  - \(S_{i-1}\) to \(S_i\)
  - \(T_{i-1}\) to \(T_i\)

- **Missing**: 
  - \(M_{i-1}\) to \(T_{i-1}\)
  - \(M_i\) to \(T_i\)

- **TagCount**: 
  - Low tag-counts, low variance, large length
  - High tag-counts, high variance, medium length
  - Low tag-counts, low variance, small length

- **Position**: 
  - Position \(i-1\)
  - Position \(i\)
Add duration constraints

Hidden integer variable, counting down segment length

Hidden binary variable, indicating whether segment state should be changed

Force state2 segments have a length $\sim [8, 30]$

Force state0 segments have a length $> 30$
Acknowledgements

• William S. Noble
• Martin Tompa
• Jay Hesselberth
• Stanley Fields
• John A. Stamatoyannopoulos