Inferring Phylogenetically Conserved Motifs

Bernard Suh

Department of Biomolecular Engineering
University of California, Santa Cruz
Biological Motivation

Figure 1: Wray, et al. 2003
Motif Search Problem

- Given $n$ sequences of length $l$ with alphabet $\{A,C,G,T\}$
- Given a tree topology with $n$ leaves
  - Star
  - Binary
- Specify motif length $k$
- Assume each sequence contains 1 motif
- Define background/null model
- Given a method of “scoring” motifs
- Given an optimality criterion (relative entropy)
- Find the position $d$ within each sequence of the “best” motif
- Consensus motif $p_i(A), p_i(C), p_i(G), p_i(T)$
Gibbs Sampler (Lawrence, et al. 1993)

- HMM Representation vs. Scanning
- Position $d$ of motif in sequence is hidden
- Learning via Stochastic EM
- Assumes star tree
FootPrinter (Blanchette, et al. 2000)

- Standard Parsimony
  - Match - 0
  - Mismatch - 1
- MAX–PRODUCT with logs (Viterbi)
- Time complexity: $O(nk(4^k)^2)$
Probabilistic Models for DNA

Jukes-Cantor, 1969 (JC69) [extension: F81]

\[ R = \begin{pmatrix}
- & \alpha & \alpha & \alpha \\
\alpha & - & \alpha & \alpha \\
\alpha & \alpha & - & \alpha \\
\alpha & \alpha & \alpha & - \\
\end{pmatrix} \]

Kimura 2-parameter, 1980 (K2P or K80) [extension: HKY85]

\[ R = \begin{pmatrix}
- & \beta & \alpha & \beta \\
\beta & - & \beta & \alpha \\
\alpha & \beta & - & \beta \\
\beta & \alpha & \beta & - \\
\end{pmatrix} \]
Hybrid Approach for Motif Search

- Combine approaches from Gibbs Sampler and FootPrinter
- Enhance FootPrinter with SUM–PRODUCT
- Probabilistic model is a phylo-HMM
  - Replace emission prob. with tree likelihood prob
  - (Felsenstein & Churchill, 1996), (Siepel & Haussler, 2003)
- MCMC via Metropolis updates for parameters
  - $\theta^* = \theta + N(0, 1)$
Synthetic Data

- generate $n$, $0^{th}$ order Markov background sequences of length $l$
- Given a tree with $n$ tips, evolve a motif of length $k$
- Implant motif into background sequence
Sample Results

- $n = 4, k = 6, l = 50, HKY 85$
- Gibbs (right), Hybrid (left)
- Summarize?
  - PME, Credible Interval
  - Consensus motif
Sample Results continued

- 2 motifs implanted
Conclusions and Future Work

- Hybrid method for motif search
- Works on simulated data
- Many areas for improvement
- Real data
- Further simulated benchmarking
- MCMC issues (burn in, mixing, termination, etc)
- Expand and contract motif length $k$
- More complicated evolutionary models