Comparison of Haplotype Motif and Block Models using the Principle of Minimum Description

Srinath Sridhar, Kedar Dhamdhere, Guy E. Blelloch, R. Ravi and Russell Schwartz
Department of Computer Science and Biological Sciences, Carnegie Mellon University

Haplotype Structure
- Haplotypes: Contiguous DNA segments between recombination sites
- Popular models of haplotype structure:
  - Blocks
  - Haplotype boundaries are aligned
  - Built on the recombination hypothesis assumption
  - Motifs
  - Overlapping haplotype boundaries
  - Robustness of the block model

Algorithm for Motifs
- Step 1 - Initial Solution
  - Construct a generative Markov model $M$ of all possible motifs with a “start” state
  - Initialize transition probabilities of $M$
  - Repeat maximization-maximization step:
    - For each row $i$ in input $I$
    - Find maximum likelihood path $P_i$ (explanation) of $i$ in $M$
  - Perform maximum likelihood estimate for transition probabilities based on number of times the transitions occur in $P$

- Step 2 - Simulated Annealing
  - Define motif as a triple $(x, e, b)$
    - where $x$, $e$ are columns and $b \in \{0, 1\}$
  - Let current solution $S$ be a set of motifs
  - Neighbors of $S$ are solutions that can be obtained from $S$ by one of the following operations:
    - Select a column $z$; exchange all $w \in (x, e, b)$ with $w' = (z, e, b)$ and add $w, w'$ to $S$
    - Select a column $z$; let $S_z \subseteq S$, $z \in S_z 
      x, e \in S_z \setminus c$ and $e, c \notin S_z$; select a subset $S_z$, of $S_z$, remove $T_i$ from $S_z$ for every $(x, e, b) \in T_i$
      add $(x, e, b)$ and $(e, f, c)$ and $(c, z, e, z)$ to $S$, where $f = b \cup b$

Minimum Description Length (MDL)
- Popular measures for comparing models:
  - MDL, information content & compression
- Let
  - $M$ represent the parameters of the model
  - $I$ represent the input
  - $E$ be the “explanation” of $I$ using $M$
  - $l$, be the length of encoding
- Objective:
  - Minimize $L(M) + \lambda \cdot I(M)$
  - Complicated models are penalized, prevents overfitting

Example: Blocks and Motifs
- Blocks: Boundary aligned
- Motifs: Overlapping boundaries
- Generative Markov model representing the motifs

Coalescent Simulation using the ms program (Hudson 2002)
- Length of DNA under simulation: 1000bp
- Mutation rate per nucleotide per generation: $2.5 \times 10^{-8}$
- Recombination rate per pair of sites per generation:
  - Low rate: $2 \times 10^{-3}$
  - High rate: $2 \times 10^{-3}$
- Effective population size: 10,000
- SNP density (number of SNPs/physical length of DNA) varied between 0.002 to 0.0035

Motifs and block haplotypes in Daly et al 2004
- Rows: SNPs
- Columns: DNA segments
- Identical columns indicate same motif or block haplotype

Algorithm for Blocks: Dynamic Program
- Dynamic Program (Korivisto et al. 2003):
  $$ F(i) = \min_{j \neq i} (F(j) + C(j, j)) $$
- where $C(j, j)$ is the cost of creating a single block from $j$ to $j$
- Running time: $O(n^2)$
- Work space: $O(n)$

Conclusions
- Motifs better capture haplotype conservation than blocks in most instances
- Results less pronounced in real data sets than in simulations that use the assumption of uniform recombination rate
- Blocks can be easily inferred and used in applications such as association testing
- Motifs are harder to infer but could possibly improve the power of association testing
- Are there better models “in between” blocks and motifs?

References