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

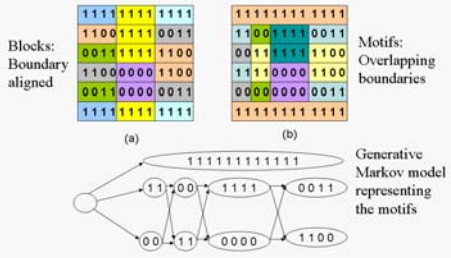
Haplotype Structure

- Haplotypes: Contiguous DNA segments between recombination sites
- Popular models of haplotype structure:
  - Blocks
    - Haplotype boundaries are aligned
    - Built on the recombination hot-spots assumption
  - Motifs
    - Overlapping haplotype boundaries
    - Relaxes the rigidity of the blocks model

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 $L(M) + L(E(I)|M)$
- Complicated models are penalized, prevents over-fitting

Example: Blocks and Motifs

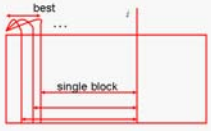


Algorithm for Blocks: Dynamic Program

- Dynamic Program (Koivisto et al. 2003):

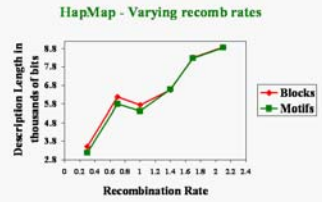
$$F(i) = \min_{1 \le j < i} (F(j) + C(j+1, i))$$

- where $C(j+1, i)$ is the cost of creating a single block from $j+1$ to i.
- Running time: $O(n^2)$
- Work space: $O(n)$

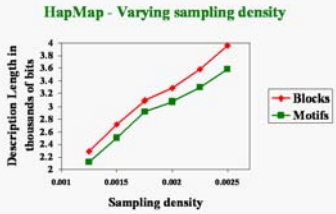


Algorithm for Motifs

- 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 r in input I
    - Find maximum likely path P_r (explanation) of r in M
  - Perform maximum likelihood estimate for transition probabilities based on number of times the transitions appear in P
- Step 2 – Simulated Annealing
  - Define motif as a triple (s, e, b)
    - where s, e are columns and $b \in \{0, 1\}^{e-s+1}$
  - 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 c; concatenate all $m = (\_, c, b)$ with $m' = (c+1, \_, b')$; add m, m' to S
    - Select a column c; let $S_c \subseteq S$, s.t $\forall (s, e, b) \in S_c$, $s \le c$ and $e > c$; select a subset T_c of S_c; remove T_c from S; for every $(s, e, b) \in T_c$, add (s, c, b_1) and $(c+1, e, b_2)$ to S, where $b = b_1 \circ b_2$

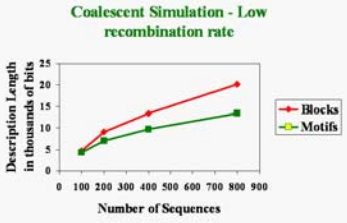


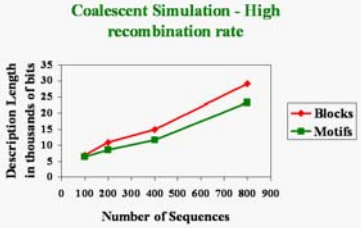
Recombination rates obtained from Jensen-Seaman et.al, 2004



Coalescent Simulation using the ms program (Hudson 2002)

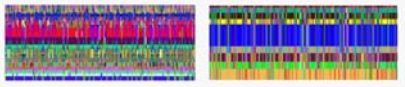
- Length of DNA under simulation: 100kb
- Mutation rate per nucleotide per generation: 2.5×10^{-8}
- Recombination rate per pair of sites per generation:
  - Low rate: 10^{-8}
  - High rate: 2×10^{-8}
- Effective population size: 10,000
- Selected SNP's with minor allele frequency at least 0.1
- SNP density (number of SNPs/physical length of DNA) varied between 0.002 to 0.0025





Motifs and block haplotypes in Daly et.al 2004

Rows: SNPs
Columns: DNA sequences
Identical colors indicate same motif or block haplotype



Motifs Description Length: 6554.93 bits Blocks Description Length: 7342.71 bits

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

- M. Daly, J. Rioux, S. Schaffner and T. Hudson. High resolution haplotype structure in the human genome. *Nat Genet* 29:229-232, 2001.
- E. Eskin, E. Halperin, R. Karp. Large Scale Reconstruction of Haplotypes from Genotype Data. In Proc *RECOMB*, 104-113, 2003.
- R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337-8, 2002.
- M. I. Jensen-Seaman, T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C.-F. Chen, M. A. Thomas, D. Haussler, and H. J. Jacob. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res* 14:528--538, 2004.
- M. Koivisto, T. Kivioja, H. Mannila, P. Rastas and E. Ukkonen. Hidden Markov modelling techniques for haplotype analysis. In Proc. *ALT* 2004
- M. Koivisto, M. Perola, T. Varilo, W. Hennah, J. Ekelund, M. Lukk, L. Peltonen, E. Ukkonen, and H. Mannila. An MDL method for finding haplotype blocks and estimating the strength of haplotype block boundaries. In Proc. *PSB*, 502--513, 2003.
- R. Schwartz. Haplotype motifs: an algorithmic approach to locating evolutionarily conserved patterns in haploid sequences. In Proc. *CSB*, 2003.