Evaluation of the Haplotype Motif Model using the Principle of Minimum Description

Srinath Sridhar [1], Kedar Dhamdhere [1], Guy E. Blelloch [1], R. Ravi [2] and Russell Schwartz [3]

October 13, 2004
CMU-CS-04-166

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[1]{srinath, kedar, blelloch}@cs.cmu.edu, Computer Science Department, CMU
[2]ravi@cmu.edu, Tepper School of Business, CMU
[3]russells@andrew.cmu.edu, Dept. of Biological Sciences, CMU

This research was sponsored by National Science Foundation (NSF) grant number CCR-0122581. R. S. is supported in part by a grant from the Samuel and Emma Winters foundation.

Keywords: single nucleotide polymorphism, haplotypes, minimum description length

Abstract

We apply minimum description length (MDL) principles to evaluate the merit of relaxing the rigidity of block models of haplotype structure. We accomplish this by developing an MDL formulation of the more general "haplotype motif" haplotype structure similar to an approach proposed independently by Koivisto et al. [K+04]. Comparison of equivalent block and motif MDL models on real and simulated data reveal that the more flexible motif models can yield substantial reductions in data explanations, suggesting that motifs are more accurately capturing the true nature of haplotype conservation. These benefits are less pronounced in real than in simulated data, however, and depend on coverage level, marker density, and intrinsic recombination rates of specific data sets.

1 Introduction

The completion of the human genome sequence has given hope that we will soon be able to understand the molecular bases of a broad range of human diseases. So far, though, the genetic factors underlying the so-called "complex diseases," which likely depend on multiple genetic and environmental factors, have proven difficult to uncover. Single nucleotype polymorphism (SNP) association studies, in which one correlates common single-base genetic variants with disease occurrence in large populations of diseased (case) and healthy (control) individuals, provide a possible avenue for progress [RM96]. These studies have so far been hampered both by the costs of sequencing many SNPs in many people and the statistical difficulties of finding faint but real signals in these large data sets. Both the experimental and statistical problems can be addressed by exploiting haplotypes, local regions of sequence that take on only a few possible variations in a population. Haplotype patterns can be used to can reduce the number of genetic sites that need to be sequenced, through a process known as "haplotype tagging" SNP (htSNP) selection [J+01]. They can also be built directly into statistical association tests [MS99, S+99, M+00] and are useful for various related problems in study design. All such methods depend, either implicitly or explicitly, on models of how haplotype correlation patterns are structured in the genome.

There has recently been great interest in "haplotype block" models [D+01], which assume that sequences can be decomposed into short regions of low diversity conserved across populations. Block representations are convenient for analysis purposes, as they can be efficiently inferred [Z+02] and provide a restricted model for subsequent computational inference. Despite their advantages, though, evidence suggests haplotype blocks cannot be robustly inferred and only imperfectly reflect true haplotype conservation patterns [NT02, S+03, WP03]. Other models of haplotype structure make no assumptions about conservation of haplotype boundaries between distinct individuals, including various Markov model representations [M+00, S+01, S+02] and parsimony models that seek to explain a population with a minimum number of recombinations [S+02, U02]. Such models have the advantage of making few prior assumptions about the nature of haplotype conservation but are too flexible to be fit reliably to realistic sizes of data sets or to easily yield efficient optimization methods for problems in association study design. The recent "haplotype motif" approach [S03] was an attempt to fill the gap between these two paradigms by capturing a broader range of possible conservation patterns than block models while still being

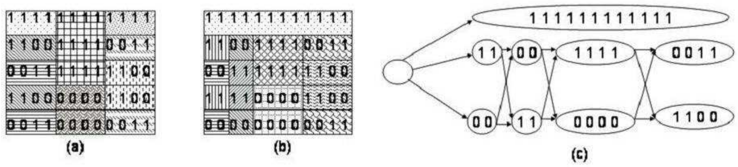

Figure 1: Figure (a) is one possible partition of a hypothetical haplotype-SNP matrix into blocks. Figure (b) is a possible motif partition and figure (c) represents the DAG/Markov Model corresponding to Figure (b). In all the figures, identical shades on different rows between the same columns denote the same conserved segment (motif or block haplotype)

sufficiently constrained to allow robust inference and efficient use in downstream analysis problems. Motif models assume that haplotype sequences can be explained as concatenations of conserved DNA segments, much like in block models, but without the assumption that boundaries between conserved segments are shared across the population. Figure 1 compares block and motif approaches using a small hypothetical haplotype-SNP matrix.

We note that Koivisto et al. [K+04] recently proposed a Markov model approach identical to that used in haplotype motifs [S03] and developed an EM algorithm with an MDL objective function. The experimental results of Koivisto et al. [K+04] do not compare their blocks model with the haplotype motifs model. In this paper, we present some preliminary experimental results using real and simulated data on the description lengths of the two models.

Even for a particular general modeling framework, various metrics can be used to fit individual data sets. The minimum description length (MDL) principle has proven promising with block models as a way to avoid overfitting with the sparse data sets currently available [K+03, An03]. MDL is based on Ockham's razor: the best explanation of a data set is the simplest model that explains it. In MDL, we minimize $L(M) + L(E(I)|M)$, where $L(M)$ is the length of the encoding of the model M in terms of a universe of possible models and $L(E(I)|M)$ is the length of the explanation of input I in terms of M. By accounting for both model and data description costs, MDL prevents one from producing a concise definition of the data by using an overly complicated model. Any encoding scheme can be applied to minimize the objective function. While its impossible to determine if an encoding scheme is optimal, when comparing any two models care is taken to make the encoding for both models as efficient as possible.

While much attention in the computational field has been focused on applications of haplotype correlation patterns, comparatively little attention has been paid to the problem of how best to represent these patterns. We focus here on the problem of characterizing haplotype structure per se, rather than its use in any particular downstream application, specifically comparing prevailing block models to the more relaxed motif framework. The MDL principle provides an objective basis for considering whether a more complicated or a simpler model is better able to concisely capture the true information content of haploid sequences, which is ultimately what one requires of a haplotype structure model. We define an MDL variant of haplotype motifs, develop associated computational optimization methods, and compare this method to an equivalent block MDL method. We then analyze a combination of real and simulated data, considering how concisely the various data sets can be represented in the two frameworks and how this depends on characteristics of specific data sets.

2 Definitions and Notations

We first define a motif model and the block model as a special case of it. We use the following definitions and notations throughout the paper.
Input: *Input* I consists of an $n \times m$ 'haplotype-SNP' matrix. The set of rows $R = \{1, \cdots, n\}$ of I correspond to DNA sequences and the set of columns $C = \{1, \cdots, m\}$ observations over SNP sites. We assume that all the SNP sites are bi-allelic, yielding a boolean matrix. The notation $I_{r,s,e}$ denotes the bits of row r of I between columns s and e (both inclusive).
Motif model: A *motif model* is defined by a set $H = \{h_1, h_2, \cdots, h_k\}$ of *'haplotype motifs'* or simply *motifs*. Each motif h_i is a quadruple (s_i, e_i, b_i, p_i). The start and end columns of the haplotype motifs w.r.t the input matrix are defined by s_i, e_i ($s_i, e_i \in C$). The bits of a haplotype motif are defined by b_i ($b_i \in \{0,1\}^{e_i - s_i + 1}$). Associated with each haplotype motif is a probability p_i. We require that $\forall j, \sum_{h_i \in H | s_i = j} p_i = 1$, i.e. the probabilities of all the motifs that start at location j sum to 1.

We can also view the motif model as a directed acyclic graph $G = (V, E)$ (or a Markov model) where $t_1 \in V$ is a special 'start' vertex. By definition, there exists a bijection $\phi : V \backslash \{t_1\} \mapsto H$. Every vertex $v \in V \backslash \{t_1\}$ such that $\phi(v) = h_i$ is annotated with the quadruple $(s_v, e_v, b_v, p_v) = (s_i, e_i, b_i, p_i)$. Therefore all the properties of a vertex v are identical to the properties of motif $\phi(v)$. Two vertices u and v are connected by a directed edge (u, v) if and only if $s_v = e_u + 1$. Directed edges (t_1, v) are added if and only if $s_v = 1$. Figure 1(c) shows a sample DAG/Markov model corresponding to the motif decomposition in Figure 1(b).

Exact conformation and explanation of I: A motif model *conforms* to input I if for every row r of I, there exists a directed path $\mathcal{P}_r$ in G that produces all the bits of the row in order. More formally, the path $\mathcal{P}_r$ is defined as an ordered set of vertices $\mathcal{P}_r = \{t_1, v_1, v_2, \cdots, v_l\}$, where $e_{v_l} = m$ and $(v_i, v_{i+1}) \in E$ and $b_{v_1} \circ b_{v_2} \circ \ldots \circ b_{v_l} = I_{r,1,m}$. Any row r could have several such candidate paths, where each path $\mathcal{P}_r$ constitutes an *explanation* for row r. Our definition for the probabilities p_i and the explanation for the rows of I are mutually dependent. To explain a row $r \in R$, the set $\mathcal{P}_r$ that is most likely (that minimizes $-\sum_{v \in \mathcal{P}_r} \log(p_v)$) is *selected* (ties are broken arbitrarily). A motif model *exactly conforms* to input I if every haplotype motif of the model appears in the explanation of at least one row, $\forall v \in V\ \exists r \in R,\ v \in \mathcal{P}_r$. We define the probability of a vertex v as $p_v = (|\{r \in R | v \in \mathcal{P}_r\}|)/(|\{r \in R | u \in \mathcal{P}_r \text{ and } e_u = s_v - 1\}|)$. Although we could construct a more general Markov-model that associates probabilities with edges instead of nodes, that would complicate association study applications and might overfit the input.

Block model: A *block model* can be viewed as a constrained version of a motif model in which motif boundaries do not overlap. More formally, a block model is defined by a set $H = \{h_1, h_2, \cdots, h_k\}$ of *block haplotypes* consisting of quadruples (s_i, e_i, b_i, p_i) defined identically to motifs. The *blocks* of H are contained in an ordered set B of *block boundaries* sorted by the start locations, $B = \{(s, e) | \exists h = (s, e, \_, \_) \in H\}$. We further require that $\forall (s_i, e_i), (s_{i+1}, e_{i+1}) \in B,\ (e_i + 1 = s_{i+1})$. The definitions of exact conformation and explanation extend to the block model in the obvious manner. *Note*: If a block model exactly conforms to I, $\forall h \in H, p = \frac{1}{n} |\{r \in R | I_{r,s,e} = b\}|$. *In this paper, we are only interested in discovering block and motif models that exactly conform to I.*

3 Haplotype Blocks

We apply here an encoding and algorithm for blocks similar to that of Koivisto et al. [K+03]. A naive encoding of the block haplotypes has cost $\sum_{i=1}^{k} \log(s_i) + \log(e_i)\ + ((e_i) - (s_i) + 1) + L(p_i)$, where $L(p_i) = \log |\{r \in R | I_{r_i, s_i, e_i} = b\}|$ and k is the total number of motifs. To encode the model more efficiently, the block haplotypes are sorted based on the start (and therefore end) locations. For any column j of I, let s'_j be the number of block haplotypes that start at location j, $s'_j = |\{h_i \in H | s_i = j\}|$. Define the set $T = \{(j, s'_j) | s'_j \neq 0\}$ of tuples consisting of start locations with corresponding (non-zero) number of block-haplotypes starting at that location. T is sorted to yield an ordered set of block haplotypes, $T = \{(t_i, s'_{t_i})\}$ such that $t_i < t_{i+1}$. Encoding the set of start locations makes encoding the end locations unnecessary. The start locations are difference-encoded with cost $2\log(t_1) + 2\sum_{i=2}^{|T|} \log(t_i - t_{i-1})$. The cost for encoding $\mathcal{P}_r$, the ordered set of block haplotypes corresponding to the explanation of a row r, is computed as $L(E(r)) = -\sum_{h_i \in \mathcal{P}_r}, \log(p_i)$. The total cost of our encoding is therefore:

$$2\log(t_1) + 2\log(s'_{t_1}) + \sum_{i=2}^{|T|}(2\log(t_i - t_{i-1}) + 2\log(s'_{t_i})) +$$

$$\sum_{i=1}^{k}((e_i - s_i + 1) + L(p_i)) + \sum_{r=1}^{n} L(E(r))$$

. This encoding differs from Koivisto et al. [K+03] in our use of difference encoding and lack of a noise model. These changes allow fairer comparison to our motif MDL model, which also uses difference encoding and lacks a noise model.

We find the least cost solution using the Koivisto et al. [K+03] dynamic programming algorithm. Their algorithm uses the recurrence $F(b) = \min_{1 \leq a \leq b}(F(a-1) + f(a, b))$, where $F(b)$ is the cost for the optimal block partition of the input up to and including column b and $f(a, b)$ is the cost of a single block from columns a to column b, both inclusive. In practice, we find that the optimal solution obtained by the above algorithm can be encoded with slightly lower cost by using the following scheme. Instead of encoding the start locations and the number of block haplotypes per start location with cost

$$2\log(t_1) + 2\log(s'_{t_1}) + 2\sum_{i=2}^{|T|}(\log(t_i - t_{i-1}) + 2\log(s'_{t_i})),$$

we define multi-set

$$\chi = \{t_1\} \cup \{s'_{t_1}\} \cup (\cup_{i=2}^{|T|}\{t_i - t_{i-1}\} \cup \{s'_{t_i}\})$$

. The multi-set χ is encoded using entropy such that if x appears in χ for n_x times, then cost for encoding all the x's in χ is $-n_x \log(n_x/|\chi|)$. The cost incurred by using the above encoding is referred to in the tables of Figures 3 and 5 as *extra cost* and the rest of the cost as *main cost.*

4 Haplotype Motifs

To encode motifs efficiently, we first sort them by start locations and then by end locations (among those that have identical start locations). Let $T = \{c \in C | \exists v \in V,\ s_v = c\}$ be the set of columns that are the starting locations for at least one motif. The set T is ordered such that $\forall t_i, t_{i+1} \in T, t_i < t_{i+1}$, so the start locations can be difference encoded with cost $2\log(t_1) + 2\sum_{i=2}^{|T|} \log(t_i - t_{i-1})$.

Define an ordered set of tuples $T_c = \{(t_{c,i}, u_{c,i})\}$ consisting of an end index $t_{c,i}$ and non-zero number of motifs starting at c and ending at $t_{c,i}$. Set T_c is ordered such that $t_{c,i} < t_{c,i+1}$. For a specific start location c, the end locations can now be difference encoded with cost $2\log(t_{c,1}) + 2\sum_{i=2}^{|T_c|} \log(t_{c,i} - t_{c,i-1})$. The cost for encoding the bits of the motif and the cost for specifying the number of motifs within each pair of start and end indices is computed as: $\sum_{i=1}^{|T_c|}((t_{c,i} - c) \times u_{c,i} + 2\log(u_{c,i}))$. The cost for encoding the probability of a vertex v (motif $\phi(v)$) is computed as $L(p_v) = 2\log(|\{r \in R | v \in \mathcal{P}_r\}|)$. Note that this information is enough to determine the probability of a vertex, since the denominator in the definition of p_v is $\sum_{u \in V | (e_u = s_v - 1)} |\{r \in R | u \in \mathcal{P}_r\}|$.

The cost for encoding the ordered set $\mathcal{P}_r$ corresponding to the explanation for row r is given by $L(E(r)) = -\sum_{v \in \mathcal{P}_r} \log(p_v)$. Therefore the total cost for the motif model with explanation is

$$2\log(t_1) + \sum_{c \in T}(2\log(t_{c,1}) + 2\log(u_{c,1}) + (t_{c,1} - c) \times u_{c,1}) + \sum_{i=2}^{|T|} 2\log(t_i - t_{i-1}) +$$

$$\sum_{c \in T} \sum_{i=2}^{|T_c|}(2\log(t_{c,i} - t_{c,i-1}) + 2\log(u_{c,i}) + (t_{c,i} - c) \times u_{c,i}) + \sum_{v \in V} L(p_v) + \sum_{r=1}^{n} L(E(r))$$

An important point to note here is that if none of the motif boundaries overlap (all the motifs are block haplotypes), then the term $2\sum_{c \in T} \log(t_{c,1})$ would be the extra cost that motif model incurs as compared to the blocks model.[1] Also, instead of coding the differences in start location, the number of end locations per start location and the number of motifs within a pair of (start, end) indices as shown above, we add the values into a multi-set χ and encode them using entropy as with the case of blocks.

[1] We need to specify end locations *in addition* to start location of motifs.

4.1 Algorithm

We use a two-step heuristic method to find low-cost motif models. We first use a two stage maximization heuristic to find a good initial solution. We then apply a simulated annealing local search method to improve the cost and find a local minimum of the potential function.

Step 1: Our starting method alternates between two maximization steps. We initialize a set H of all possible haplotype motifs in I (all possible substrings in I) and construct the DAG corresponding to H as described in the previous section. The probability p_v associated with vertex v is initialized to reasonable non-zero values. The maximization step finds the least cost path $\mathcal{P}_r$ that explains each row r separately, assuming that the probabilities of the vertices are fixed. After completing the first maximization step on all rows, the second performs a maximum likelihood estimate by re-normalizing the probabilities of all the vertices such that the vertices that were used to explain more rows obtain proportionally higher probabilities. During each iteration of the algorithm the size of G and therefore the size of H decreases. See the pseudo-code in Figure 2 for more details. The algorithm requires $O(nm^2)$ space to store all possible motifs and their associated attributes and the time per iteration is bounded by $O(nm^2)$.

Step 2: Our simulated annealing method improves on the solution by considering two perturbations to the input: *merge*(j) and *split*(j) where $j \in C$. If the current solution is G with cost $c(G)$ a merge operation creates a new solution G' by merging *all* the motifs that end at j with motifs that begin at $j+1$. To support *split* consider the set of motifs that contain bits at columns j and $j+1$, $S_j = \{v \in G \mid s_v \leq j \text{ and } e_v > j\}$. Just as in the case of blocks we assume that $|S_j|$ is bounded by a constant. The split operation considers *all* possible subsets of $S'_j \subseteq S_j$ and computes $2^{|S_j|}$ new solutions G' corresponding to splitting all the motifs in S'_j between locations j and $j+1$. If a sequence is explained using a motif u in G, then it is explained with the motifs u_l and u_r of G', where u_l and u_r are motifs obtained by splitting u between j and $j+1$. Note that after performing the splits, some motifs become identical w.r.t their start, end locations and the bits they represent. Consider two motifs u and u' of G' that are identical. Let E_u and $E_{u'} \subseteq R$ be the set of rows that contain the motifs u and u' in their explanation respectively. Then a motif v replaces u and u' in G' such that it is identical to u and u' and $E_v = E_u \cup E_{u'}$. The Markov model where the states represent current solutions and edges the transitions between states by performing a split/merge operation is ergodic. The simulated annealing procedure randomly selects a candidate G' that is obtained either by a split or merge. If $c(G') < c(G) + \alpha Z$, where Z is an exponential random variable with mean 1 and $\alpha \geq 0$ is the simulated annealing parameter, then G' replaces G as the current solution. The simulated annealing parameter α is slowly reduced to 0. **Note:** We use several additional heuristics to both steps to improve the running time, space requirement and accuracy.

5 Experimental Results

We analyze a combination of real and simulated haplotype data. We applied the *ms* program of Hudson [H02] to simulate samples of a 100 kb region of DNA under a Wright-Fisher neutral model by the coalescent method. We used a mutation rate of 2.5×10^{-8} per nucleotide per generation to match that estimated for humans [NC00]. We use two different recombination rates — a low recombination rate of 10^{-8} per pair of sites per generation and a high recombination rate of 2×10^{-8} — to capture the range of variation estimated between human chromosomes [J+04]. The effective population size N_0 is set to $10,000$, as estimated for modern humans [RY03]. We used 100-800 sequences in our experiments. From the *ms* output we selected only those sites with minor allele frequency at least 10%, consistent with typical SNP selection guidelines. After this step, we obtain about 200-250 SNPs on average, giving a frequency of about one SNP per 400-500 bp.

The table in Figure 3 compares the effectiveness of blocks and motifs in concisely describing different simulated data sets. Motifs consistently yield substantially smaller explanations, with a greater relative reduction for low recombination rates. Figure 4 shows the dependence of results on data set size. The advantage of motif models increases with increasing population sizes. We note that at unrealistically low and high recombination/mutation rates, block models perform better than motif models (data not shown).

We show results from two real data sources produced by Daly et al. [D+01] and the HapMap project [H04]. Large scale phase known haplotype data is not available and therefore we used the phasing algorithm of Eskin et al. [E+03] as provided by the **HAP** webserver. To mimic the data used for association test, we selected one per trio of the Daly et al. data set. For the HapMap data, we present results for two input matrices of which one exhibits small blocks and the other longer blocks. Figures 6 and 7 illustrates motif and block patterns using the Daly et al. haplotype-SNP matrix, transposed such that the vertical lines denote haplotype sequences and the horizontal lines SNPs. We can observe that some boundaries match across the two models but motifs also find overlapping conserved segments that would be impossible for the blocks model to discover.

Figure 5 summarizes description length results for the real data sets. The real data sets are noticeably more favorable to blocks than are simulated data sets. Blocks yield more concise explanations for the short-block haplotype map data. Motifs yield shorter description lengths when the block sizes are comparatively large. Small block sizes in some data sets might result from either high local mutation/recombination rates or low SNP sampling densities. In either event, the advantage of using haplotypes at all would be limited for such data sets. We note that in the Gabriel et al. [G+02] data-set, which is much smaller in both input size and the average block length discovered, the motif model was competitive but not better than the block model (data not shown).

6 Conclusions

We have presented a novel method for inferring haplotype structure by combining MDL principles with the haplotype motif framework and used this to evaluate the merits of relaxing assumptions of haplotype block models. It is currently a matter of debate how much human haplotype structure has been shaped by recombination hotspots versus random recombinations. Either would yield data that could be fit to block models [NT02], although patterns produced predominantly by recombination could likely be much better fit to more general haplotype structure models. Our results on simulated data lacking hotspots suggest that such data could be much more parsimoniously explained by motif than by block models. Real data shows a less pronounced and more inconsistent benefit to using motif models, although the advantage of motif models becomes more pronounced as data set sizes and marker densities grow to levels likely to be desirable for association studies. Together, these results suggest potentially large advantages to relaxing some of the restrictions of block models if we wish to best characterize the information content of human genetic sequences and apply that information to solving pressing problems in human medicine.

References

[An03] E. C. Anderson and J. Novembre. Finding haplotype block boundaries by using the minimum-description-length principle *Am J Hum Genet* 73:336–354, 2003.

[D+01] M. Daly, J. Rioux, S. Schaffner and T. Hudson. High resolution haplotype structure in the human genome *Nat Genet* 29:229-232, 2001.

[E+03] E. Eskin, E. Halperin, R. Karp. Large Scale Reconstruction of Haplotypes from Genotype Data In Proc *RECOMB*, 104-113, 2003.

[G+02] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altschuler. The structure of haplotype blocks in the human genome *Science* 296:2225.2229, 2002.

[H04] The International HapMap Consortium. The international HapMap project. *Nature* 426:789–796, 2003.

[H02] R. R. Hudson Generating samples under a Wright-Fisher neutral model of genetic variation *Bioinformatics* 18:337-8, 2002.

[J+04] 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.

[J+01] G. C. Johnson, L. Esposito, B. J. Barret, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbrigde, R. C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes *Nat Genet* 29:233237, 2001.

[K+04] M. Koivisto, T. Kivioja, H. Mannila, P. Rastas and E. Ukkonen. Hidden Markov modelling techniques for haplotype analysis In Proc. ALT 2004

[K+03] 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.

[KPU04] M. Koivisto, P. Rastas and E. Ukkonen. Recombination Systems *LNCS 3113:159-169, 2004*

[MS99] M. S. McPeek and A. Strahs. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping *Am J Hum Genet*, 65:858–875, 1999.

[M+00] A. P. Morris, J. C. Whittaker, and D. J. Balding. Bayesian fine-scale mapping of disease loci by hidden Markov models *Am J Hum Genet*, 67:155–169, 2000.

[NC00] M. W. Nachman and S. L. Crowell. Estimate of the mutation rate per nucleotide in humans *Genetics*, 156: 297–304, 2000.

[NT02] M. Nordborg and S. Tavare. Linkage disequilibrium: what history has to tell us *Trends Genet*, 18:83–90, 2002.

[R78] J. Rissanen. Modelling by shortest data description *Automatica* 14 465–471, 1978.

[RM96] N. J. Risch and K. R. Meikangas. The future of genetic studies of complex human diseases *Science*, 273:1516–1517, 1996.

[RY03] Bruce Rannala and Ziheng Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci *Genetics*, 164:1645-1656, 2003.

[S03] R. Schwartz. Haplotype motifs: an algorithmic approach to locating evolutionarily conserved patterns in haploid sequences In Proc. *CSB*, 2003.

[S+02] R. Schwartz, A. G. Clark, and S. Istrail. Methods for inferring block-wise ancestral history from haploid sequences *LNCS*, 2452:44–59, 2002.

[S+03] R. Schwartz, B. Halldórsson, V. Bafna, A. G. Clark, and S. Istrail. Robustness of inference of haplotype block structure *J Comp Bio*, 10:13–21, 2003.

[S+01] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data *Am J Hum Genet*, 68:978–989, 2001.

[S+99] S. K. Service, D. W. Temple Lang, N. B. Freimer, and L. A. Sandkuijl. Linkage disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations *Am J Hum Genet*, 64:1728–1738, 1999.

[U02] E. Ukkonen. Finding founder sequences from a set of recombinants *LNCS*, 2452:277–286, 2002.

[WP03] J. D. Wall and J. K. Pritchard. Assessing the performance of the haplotype block model of linkage disequilibrium *Am J Hum Genet*, 73:502–515, 2003.

[Z+02] K. Zhang, M. Deng, T. Chen, M. S. Waterman and F. Sun. A dynamic programming algorithm for haplotype block partitioning In Proc. *Nat Aca Sci USA*, 99:7335–7339, 2002.

function **runMotifMinCost**

1. create a set of all possible motifs: $H = \{h_1, h_2, \cdots, h_k\}$. Initialize a DAG $G = (V, E)$ corresponding to H

2. create a new vertex t_2 and add directed edges (v, t_2) iff $e_v = m$

3. for each $v \in V(G)$ let $c_v := |\{r \in R | I_{r,s_v,e_v} = b_v\}|$

4. reNormalizeProbabilities

5. for each iteration of the algorithm

   (a) for $r := 1, \cdots, n$

      i. consider subgraph G', s.t $\forall v \in V(G')$, $I_{r,s_v,e_v} = b_v$

      ii. compute a shortest path $\mathcal{P}_r$ in G' from t_1 to t_2
      (i.e) find $\mathcal{P}_r$ minimizing $-\sum_{v \in \mathcal{P}_r} \log(p_v)$ is minimized

      iii. for each $v \in \mathcal{P}_r$ increment c_v

   (b) reNormalizeProbabilities

function **reNormalizeProbabilities**

1. for $i := 1, \cdots, m$

   (a) sum $:= \sum_{v \in V | s_v = i} c_v$

   (b) for all $v \in V$ s.t $s_v = i$

      i. $p_v := c_v/\text{sum}$

      ii. $c_v := 0$

Figure 2: Algorithm for haplotype motifs

#Seqs/ Rcmb Rt	Num SNPs	DL: Motifs	DL: Blocks main + extra
100/low	208	4342.48	4715.89+81.35
200/low	247	7028.01	9167.05+44.55
400/low	231	9710.4	13460.37+35
800/low	221	13353.9	20218.22+20
100/high	205	6298.56	6953.77+75.35
200/high	276	8577.02	10892.17+67.06
400/high	223	11640.3	14965.19+64.30
800/high	231	23403	29094.79+46.55

Figure 3: Results of motif and block models on simulated data

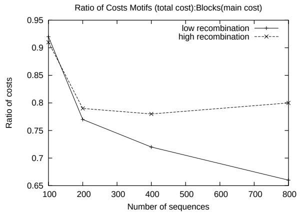

Figure 4: Ratios of MDL costs between motif (total cost) and block (main cost) models

Src, Num Seqs	Num SNPs	Desc Length: - Motifs	Desc Length: Blocks - main + extra cost	Avg Block Length
[D+01] 258	103	6554.93	7342.71 + 74.44	11.44
[H04] Chr22 180	150	8349.06	9245.28 + 148.20	8.33
[H04] Chr1 180	150	10135.9	9868.23 + 199.41	5.77

Figure 5: Description lengths of motif and block models using real data

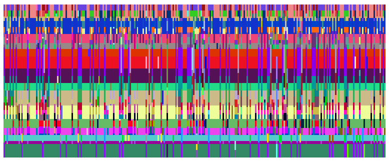

Figure 6: Motifs of Daly et al. [D+03] as discovered by our algorithm. Each color represents a unique motif.

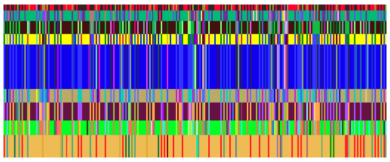

Figure 7: Block haplotypes of Daly et al. [D+03] as discovered by our implementation of the algorithm by Koivisto et al.