# Haplotype Motif Partitioning for Association Studies

Natalie Castellana

Advisor: Russell Schwartz

## Abstract

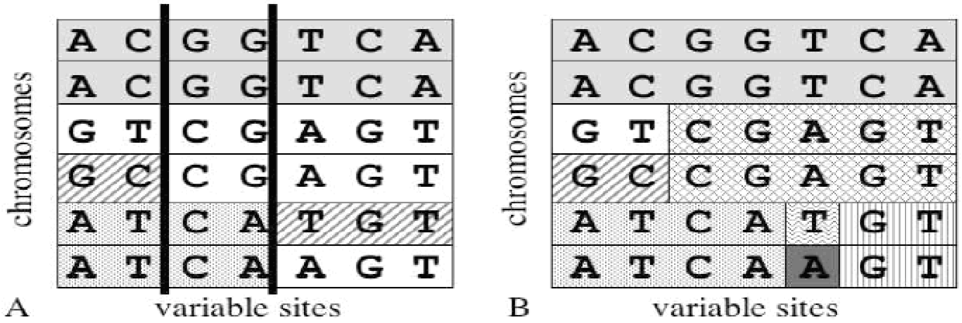
Since the first full genome was sequenced in 1995, the amount of available genomic data has grown exponentially. Utilizing patterns of variation, called haplotypes, has allowed scientists to begin drawing correlations between an organism's genetic code and the characteristics that manifest themselves, i.e. hair color, height, tendency towards depression. While several methods for finding haplotypes have been explored, one model, the haplotype motif model[1], is especially promising. Motifs are intended to capture conserved variation while relaxing some of the constraints imposed by previous methods. The model is designed to test whether correlation information in haplotypes is lost by the more rigid models. Finding the minimum number of motifs is an APX hard problem[2]. So, the focus of my research has been to find an approximation to the solution. One approach is to use an integer programming formulation of the problem. To test how well the approximation algorithm finds useful motifs, this paper looks at compression ability and performance in association testing. The results are compared to a haplotype block model. In the association tests as well as in the compression tests, the haplotype motifs perform marginally better than the blocks in most parameter sets. The sheer size of the data sets, however and the nature of the motifs make both finding partitions and creating statistics for finding association difficult tasks.

## Introduction

The search for effective ways to match genotype to phenotype has been at the heart of many bioinformatics problems. A strong motivation for this research is to correlate genetic variation with complex diseases. Single nucleotide polymorphisms (SNP's) are single bases where variation occurs in a population and have shown great potential in association studies. Complexes diseases, however, rely on multiple genetic markers, making it difficult to distinguish between false associations from the enormous data sets and true but weak signals from several different sites. Single nucleotide polymorphisms (SNP's) are sites in DNA where multiple alleles are observed in a population, and have recently drawn significant attention in association studies[3]. However, the sequencing of many SNPs in a large population is costly, and the enormous number of hypothesis being tested at once pose serious concerns for future studies. Haplotypes are one way of reducing the amount of information to be examined by making use of regions of correlated variation. We expect these regions of correlated variation to occur as a result of recombinations in human history (swapping of whole regions of the genome with other regions). It is therefore very important to determine how haplotype patterns are arranged in the genome.

One family of methods that has performed well in association tests relies on a "haplotype block model"[4]. This model proposes that the genome can be broken up into several regions of variation, between which no significant correlation exists because of past recombinations. The block model can significantly reduce the complexity of the data, but evidence suggests that blocks do not capture the true structure of haplotypes[5].

There has recently been proposed a relaxation of the block model called the "haplotype motif" model[1]. This model allows for a robust fitting of the data[3] without requiring strict correlation region boundaries. Figure 1 demonstrates the difference between the block model and the haplotype model on a hypothetical SNP matrix.



| | | | | | | |
|---|---|---|---|---|---|---|
| A | C | G | G | T | C | A |
| A | C | G | G | T | C | A |
| G | T | C | G | A | G | T |
| G | C | C | G | A | G | T |
| A | T | C | A | T | G | T |
| A | T | C | A | A | G | T |

| | | | | | | |
|---|---|---|---|---|---|---|
| A | C | G | G | T | C | A |
| A | C | G | G | T | C | A |
| G | T | C | G | A | G | T |
| G | C | C | G | A | G | T |
| A | T | C | A | T | G | T |
| A | T | C | A | A | G | T |

**Figure 1. Possible block and motif partitions of hypothetical data set. Each row in the matrix represents a different sample, and each column represents a SNP site. A: A block model B: A motif model**

In this paper, a formulation and implementation of the motif model is explored. Then the partition is compared to a partition derived from the block model. The models are evaluated based on compression (the number of haplotypes found in the data) and performance in association tests.

We began by first procuring both simulated and real data sets. We then continued our studies by inferring haplotype motif structure and haplotype block structure from the data sets. After finding the haplotypes, we then performed association tests with a simulated disease phenotype.

## Computational Methods

*Haplotype Inference*

Haplotype blocks were inferred using a dynamic programming algorithm[6]. This method found the block partition by seeking to minimize the number of haplotypes over the whole data set, while limiting the length of any haplotype to a fixed length. In this paper a length of 15 SNP's was used for the simulated data sets while a length of 10 SNP's was used for real data sets. The discrepancy in maximum block size between the types of data sets results from the difference in size of the data sets.

Motifs in each block partition were inferred using a multi-commodity flow formulation[7] which was then fed into the GNU Linear Programming Kit (REFERENCE?). Figure 2 is an example to explain the formulation below,

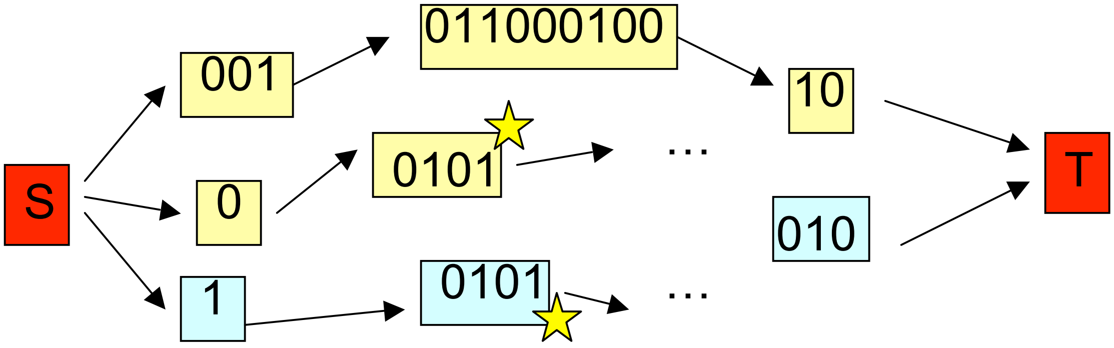


**Figure 2. A visual representation of the IP formulation. The starred motif nodes correspond to two different z variables but the same x variable. The yellow nodes correspond to motifs for one row, while the blue nodes correspond to motifs inferred for a second row.**

We define a motif is a row *r,* a starting column *a*, an end column *b*, and a bit string $s$ or simply (*r,a,b,s*). Let A be our *m* x *n* bit matrix. Then, Z is the set of all distinct motifs from the matrix. Let us collapse Z, such that a motif (*x,a,b,s*) corresponds to the same variable as the motif (*y,a,b,s*). This new set is X. Our structural variables are then of the form $z_{r,a,b}$ and $x_{a,b,s}$. Note that an *x* variable can correspond to the same motif as several *z* variables. Also note that there are $\frac{n(n+1)}{2}m$ *z* variables and hopefully fewer *x* variables. The optimal solution is bounded by $\min\{m,2n\}$, because the trivial solutions (each row is a motif and each column has two motifs "1" and "0") have size *m* and *2n* respectively.

Objective: $\min \sum_{i \le |X|} x_i$

Subject to:

Out Flow Constraint: $\sum_{r<N} z_{i,0,r} = 1 \quad \text{for all } i < m$

This constraint ensures that exactly one unit of flow exits the graph for each row in the input matrix, or in other words, each row may only contribute one sequence of motifs.

Flow Conservation Constraint: $\sum_{j=0}^{k} z_{r,j,k} - \sum_{j=k+1}^{n} z_{r,k+1,j} = 0$ for all $r < m$ and $k < n$-1

This constraint requires that for a row, each motif transitions into a viable next motif for that row.

Arc Capacity Constraint: $z_{r,a,b} - x_{a,b,s} \leq 0$ for all *r* such that the motif at row *r,* starting at *a* and ending at *b* has bit string *s*

This constraint ensures that a motif's variable cannot contribute to the objective function unless the corresponding unique motif has been selected for a row in the partition.

Integrality Constraint: $z_{r,a,b} \in \{0,1\}, x_{a,b,s} \in \{0,1\}$ for all *r,a,b,s*

This constraint requires that whole motif's are selected or not selected (no half motifs).

As earlier stated, finding the motif partition is an APX hard problem. Therefore the IP formulation is a subroutine in a motif approximation algorithm. The algorithm is similar to the block partitioning method. Haplotype motifs were found using the same dynamic programming framework. The method follows from the work of Zhang et al[4] but within each block partition found, the data set is further partitioned into motifs. Then in the algorithm, the score of a block is again the number of haplotypes and the partition seeks to minimize the score over the whole data set.

## Method Validation and Results

### *Data Processing*

We evaluated the two methods on both simulated and real data sets. The simulated data was generated using the *ms* program[8] which uses a coalescent scheme adopting the principles of the Wright-Fisher neutral model. We used a mutation rate, recombination rate, and effective population size of $2.5\times10^{-8}$ per nucleotide per generation[9], $10^{-8}$ per pair of sites per generation[10], and 10,000[11] based on realistic human values. We generated 220 data sets for a region of a 10,000 site chromosome in 2,000 individuals. These sequences were randomly paired to simulate the homologous chromosomes of individuals. SNP's with minor allele frequency less than 10% were removed from the sequences.

A 500 bp region of 7q21.13 was downloaded from the ENCODE resequencing project[12]. The data was derived from four population groups determined by the HapMap[13]: CEPH (Utah Residents with Northern and Western European ancestry); Han Chinese in Beijing, China; Japanese in Tokyo, Japan; and Yoruba in Ibadan, Nigeria. To process the real data, first we removed the SNP's that weren't observed in all populations. Due to the complexity of the inference and analysis of the methods, the ENCODE data sets were reduced to the first 50 bp.

A disease phenotype was simulated in these data sets to separate the samples into healthy individuals (controls) and diseased individuals (cases). One SNP was selected as the causal site. The minor allele frequency of this SNP was required to be between 40%

and 60%. The site was then chosen from the candidates in the middle 20% of the sequence. A model of the disease was chosen based on a penetrance parameter *p*. Individuals homozygous in the disease allele (both chromosomes had a "1" at the causal site) had a probability *p* of being in the case data set, and a 1-*p* probability of being in the control data set. Similarly, individuals homozygous for the normal allele had a probability 1-*p* of being in the case data set and a probability *p* of being in the control set. Heterozygous individuals (chromosomes had different alleles) had equal probability of being in the case and control set. Data sets were segregated on each penetrance between 0.55 and 1 at 0.05 increments. Ten case/control sets were also generated from simulated data using penetrance of 0.5 for specificity testing. For the real data sets, each population group was segregated independently and then pooled into one data set. The process requires equal numbers of cases and controls from each population group.

The association tests for both haplotype blocks and haplotype motifs were applied to all simulated data sets. A method was measured by how many data sets in a penetrance class yielded an association above LOD value 3 (p-value of 0.001), LOD value 2.5 (p-value of 0.0032), LOD value 2 (p-value of 0.01), and LOD value 1.5 (p-value of 0.032). False positive rates were measured on randomly segregated case and control sets.

## *Association Testing*

Association for both haplotype blocks and haplotype motifs was determined using a statistic which was normalized each haplotype based on its frequency of occurrence in the data set. For each SNP in the data set, we counted the occurrences of each haplotype over that site in the case group and the control group. We then computed a statistic of association for the SNP using the formula:

$$\alpha = \sum_{haplotypes} \frac{\left|A_i - B_i\right|}{\sqrt{A_i + B_i}}$$

where $A$ the count of cases with the haplotype, and $B$ is the count of the controls with the haplotype.

Since the motif method does not divide the data sets into discrete, uncorrelated regions as the block does, a permutation test was used to determine significance in the form of p-values. For each data set, the test randomly permuted samples into cases and controls. Then the haplotypes were evaluated using the above statistic, and the max value over all the SNP's was added to a table. This was done 1000 times. P-values were determined by the position of the statistic value in the table of 'random' values.

## *Results*

We began our analysis of the motif method with the simulated data sets. The haplotype motif and haplotype block structures were inferred for 5 data sets per penetrance in the range of 55% to 70%. We measured association by examining the negative log of the p-value (LOD) at each site. We found that at penetrance above 70%, both the block method and the motif method found an association at LOD cutoff 3 in 100% of the data sets. The power of each method, as determined by fractions of data sets with significant values, is plotted in Figure 3. For each LOD score, 1, 2, and 3 we determined the fraction of data sets at each penetrance that found a correlation of at least the selected score. We found that for diseases simulated with penetrance values greater than 0.7, both the block method and the motif method found significant correlation. With a maximum haplotype size of 10 SNP, blocks and motifs perform equally well at penetrance above 65%. Only in the LOD 3 graph, do motifs outperform blocks at penetrance 60%. This indicates that the two methods are only distinguishable for a very narrow range of weakly correlated diseases. Overall, it appears that the motifs performed at least as well as the blocks in all cases.

Because the block and motif methods perform similarly for the higher penetrance cases, we decided to look closely at the methods for penetrance values 0.51 through 0.60. Figure 4 depicts the results. At LOD cutoff of 1, the block method outperforms the motif method in association at penetrance 0.54. In all other parameter sets, the haplotype motif partitions find an association at least as often as the block partitions.

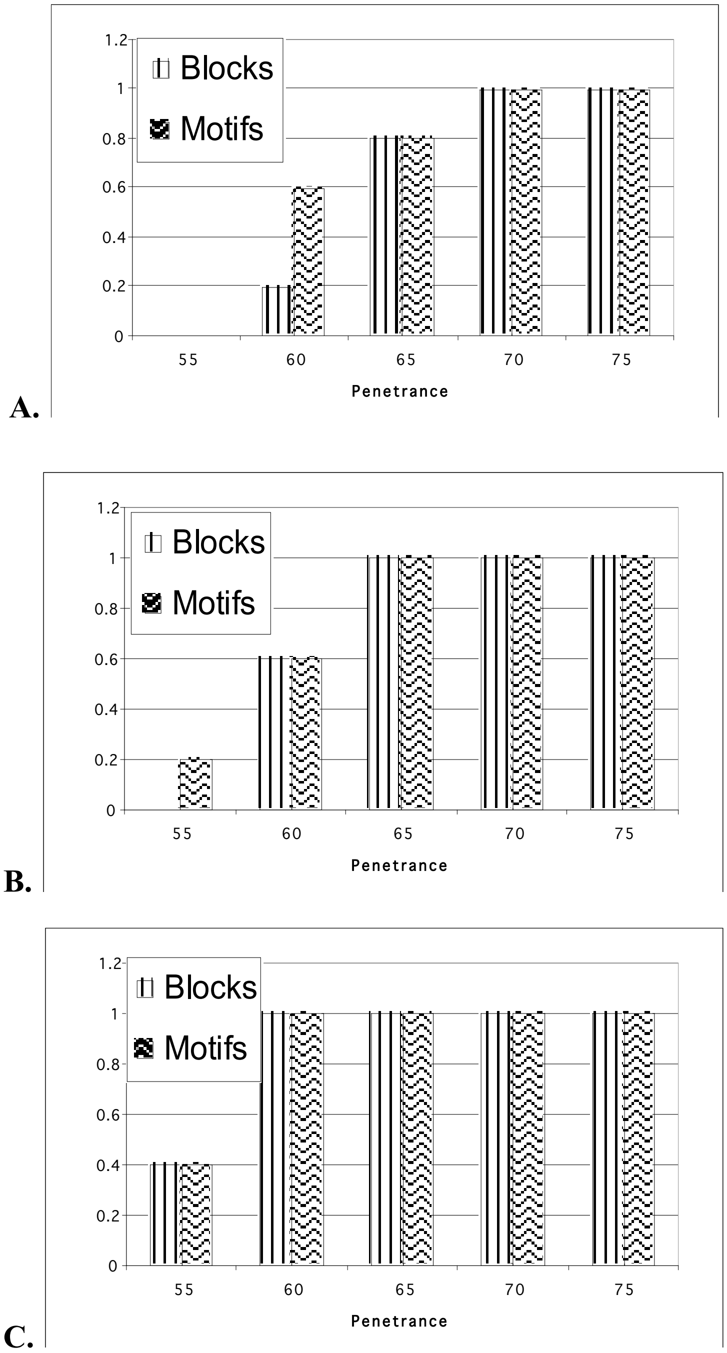


**Figure 3. Power of associations in simulated data sets at maximum haplotype length 10. A: LOD score of 3; B: LOD score of 2; C: LOD score of 1.**

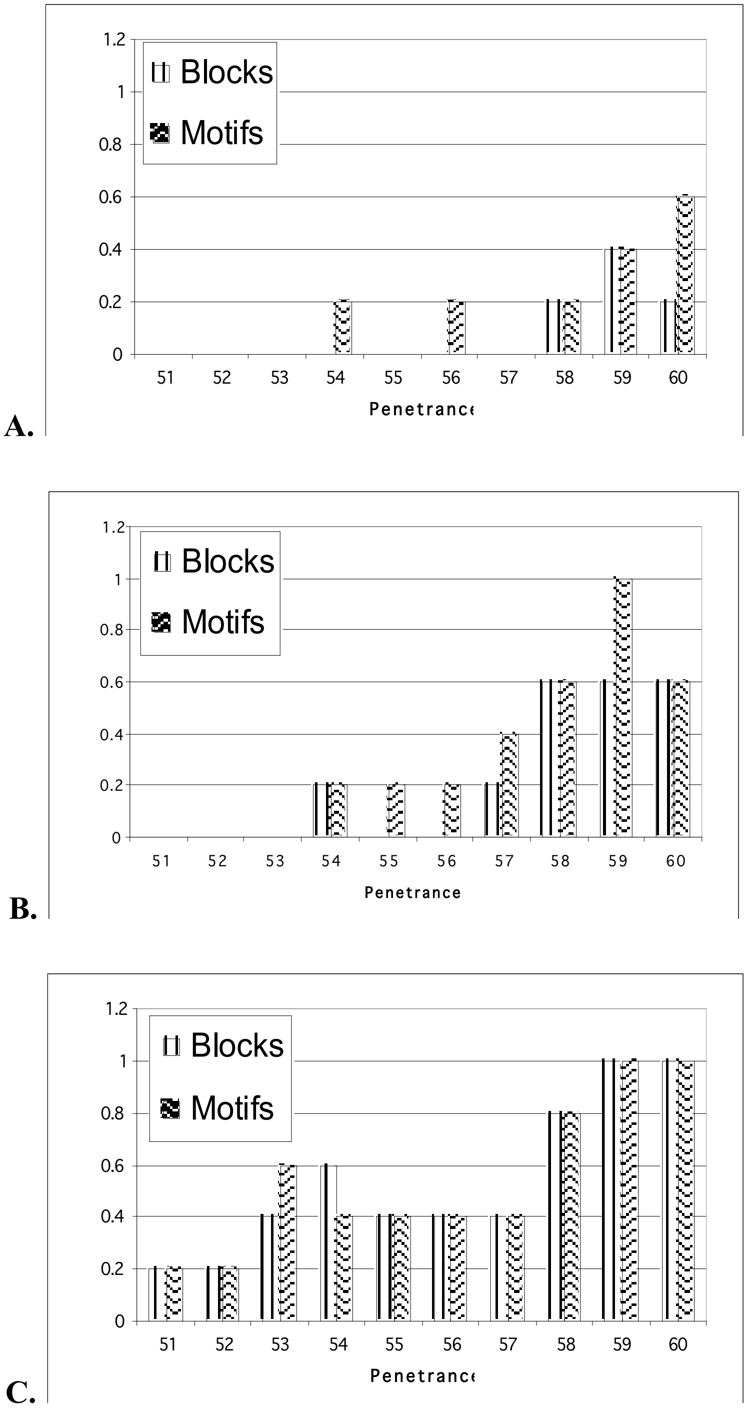


**Figure 4. Power of association in simulated data at penetrances 0.51-0.60 and maximum haplotype length 10. A: LOD score 3; B: LOD score 2; C: LOD score 1**

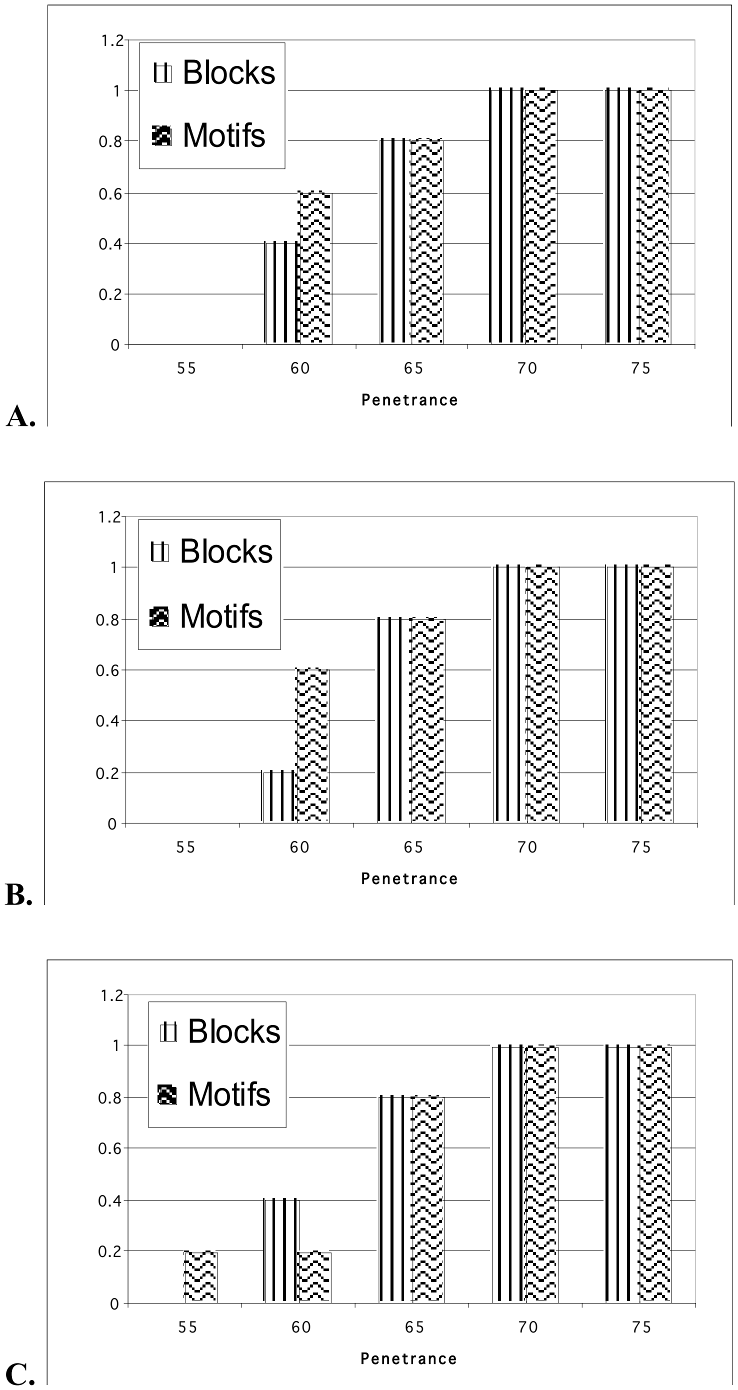


**Figure 5. Power of association at LOD 3 for simulated data sets. A: Max haplotype length 5; B: Max haplotype length 10; C: Max haplotype length 15.**

The simulated data sets were also used to determine the effect of maximum haplotype length on association tests. We repeated the above tests for maximum lengths of 5 and 15. The best illustration of the effect this parameter change has on the association tests can be observed in Figure 5. The block method's power appears consistent despite the change in allowed haplotype length, dipping only by one data set at max length 10. The motif method also demonstrates no change between maximum length of 10 and maximum length of 5. At length 15, however, the method picks up more correlation at penetrance .55, while the power is reduced for penetrance .60. This suggests that increasing the maximum length of haplotypes allows the motif method to more accurately infer the true haplotype structure and consequently find weaker signals in the data.

To test the specificity of the methods, we ran the association tests on randomly segregated, simulated case and control data sets (penetrance of 0.5). We then measured the fraction of data sets for which each method falsely indicated a correlation at LOD scores of 3, 2.5, 2, 1.5, and 1. At maximum haplotype length 10, the motif method found no correlation at any LOD score. The block method however, found weak false positives in one data set.

We also compared the block and motif methods in terms of compression ability. Our methods for optimizing both partitions are based on the notion of maximum parsimony. Specifically, we assume that the best model of a sequence's evolution is one that allows the fewest mutations and recombinations. Therefore, we designed both partitioning algorithms to seek the minimum number of haplotypes. We test compression then by how many haplotypes are inferred from how many partitions in the data sets. Figure 6 displays a representative image of the comparison. The tables contain the haplotype count and partition count for 5 data sets at penetrance 0.7 for maximum haplotype lengths 5, 10, and 15.

**A.**

| | | Data Sets | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| Block method | haplotypes | 22 | 50 | 54 | 20 | 73 |
| | partitions | 5 | 11 | 12 | 4 | 18 |
| Motif method | haplotypes | 22 | 50 | 54 | 20 | 73 |
| | partitions | 5 | 11 | 12 | 4 | 17 |

**B.**

| | | Data Sets | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| Block method | haplotypes | 21 | 46 | 47 | 18 | 55 |
| | partitions | 3 | 7 | 6 | 3 | 11 |
| Motif method | haplotypes | 20 | 46 | 47 | 17 | 55 |
| | partitions | 3 | 6 | 6 | 2 | 10 |

**C.**

| | | Data Sets | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| Block method | haplotypes | 21 | 46 | 43 | 16 | 48 |
| | partitions | 3 | 7 | 5 | 2 | 7 |
| Motif method | haplotypes | 20 | 46 | 43 | 16 | 48 |
| | partitions | 2 | 3 | 4 | 2 | 6 |

**Figure 6. Compression data for simulated data sets. A: Max haplotype length 5; B: Max haplotype length 10; C: Max haplotype length 15.**

At maximum motif length 5, the partition methods perform identically in terms of the number of haplotypes inferred. In data set 4, however, the motif method only requires 17 partitions whereas the block method requires 18. Though the discrepancy is small, it perhaps suggests that while the same number of haplotypes is inferred, the motif method better captures the true structure of the data set by allowing larger partitions with variable length haplotypes within them. Also, it is possible that there are few true haplotypes of size less than 5, so the motif partitions with this parameter set look very similar to the block partitions.

We also looked at the compression performance as the maximum haplotype length increases. The number of haplotypes inferred by either method decreases as the maximum length increases. In addition, for three of the data sets, the motif method found fewer haplotypes than did the block method. This suggests that not only are some haplotypes of length greater than 5 or 10, but also that haplotypes can vary in length even within the same partition. The discrepancy between the number of partitions required by the motif method and the number of partitions required by the block method also increases with the maximum haplotype length. This is further evidence of the variation of haplotype length within a partition.

After our analysis of the methods on the simulated data sets, we examined the real data sets. We ran association tests on 5 data sets at each penetrance from 60% to 100% at 5% increments. Figure 7 shows the plot of the power as a function of penetrance. In the real data sets, the motif method performs at least as well as the block method in 12 of the 20 parameter sets that any significance was found. At high penetrances and low LOD cutoffs, the motif method and block method are indistinguishable. However, at lower penetrances, it is difficult to determine which method performs better. At LOD 3, the motif method appears to find more correlations at penetrance less than 90%. At LOD 2, the block method begins to pick up the signals at penetrances 80% and 85%, while the motif method is more powerful at penetrance 75%.

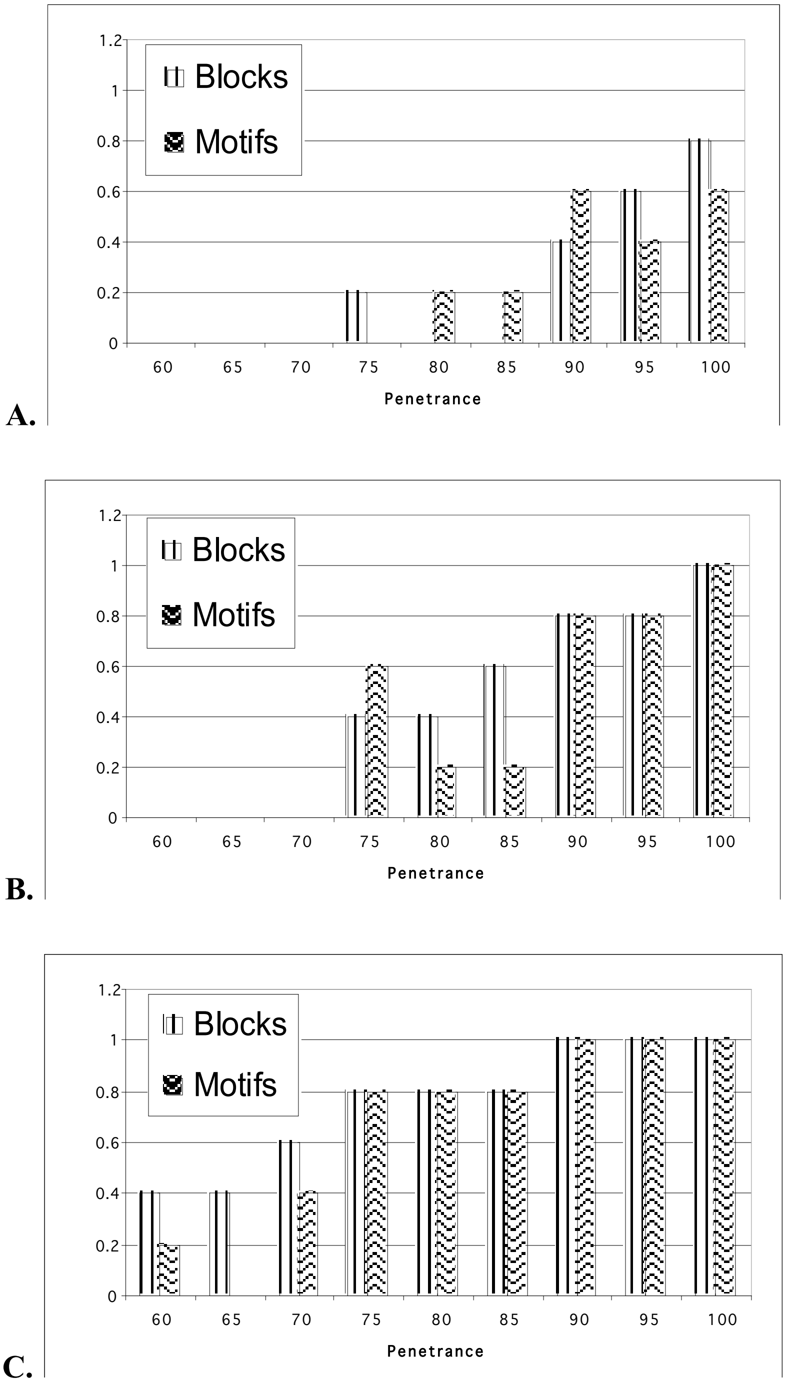


**Figure 7. Power of association for ENCODE data. A: LOD score 3; B: LOD score 2; C: LOD score 1**

The block method assumes that frequent recombinations occur DNA at "hot spots". This is captured in the partition by the hard block boundaries. The motif method on the other hand hypothesizes that recombinations occur at various locations in DNA over time, therefore haplotypes end and begin at different sites for different samples.

## Conclusions

The motivation for this study was to measure the performance of haplotype motif methods against the haplotype block model. Our results have shown that in most simulated instances the motif method performs at least as well as the block method in both association tests and compression ability. The greatest disparity in the simulated data sets between the methods was observed when maximum haplotype length was large and the disease correlation was the weak. For the real data sets, neither method completely dominated the other in the power calculations. While we expect the motif partitions to capture more accurately the true haplotype structure, due to their definition, the motifs cannot be tested as a series of independent hypotheses, and therefore finding a method for calculating association can be elusive.

The integer programming formulation of the motif partitioning algorithm allowed faster inference of haplotypes and longer maximum haplotype lengths, which explains in part the improvement over previous implementations of the method. Unfortunately, the hardness of the problem continues to hinder the testing of approximate haplotype motif inferences on larger data sets, such as whole chromosomes, or exact partitions with no limit on haplotype length. Overall, the results of this study indicate that the relaxation of haplotype boundaries affords the motif method some advantage over the block method.

## Acknowledgements

We extend thanks to G. Lancia for helpful discussion and suggestions on this work. This work was supported in part by Merck Program for Computational Biology and Chemistry at Carnegie Mellon University and the ALADDIN Center's REU program.

[1] R. Schwartz, *Proc. IEEE Comp. Sys. Biotech. Conf.,* 306 (2003).

[2] G. Lancia, R. Rizzi. Complexity results and approaches to the tiling of binary matrices. unpublished manuscript. 2006.

[3] N.J. Risch and K.R. Meikangas. *Science*. **273**. 1516 (1996).

[4] M.J. Daly, J.D. Rioux, S.F. Schaffner, and T.J. Hudson, *Nat. Genet.* **29,** 229 (2001).

[5] R. Schwartz, B.Halldórsson, V. Bafna, A.G. Clark, and S.Istrail. *J. Comp. Biol.* **10,** 13 (2003).

[6] K. Zhang, M.Deng, T.Chen, M.S. Waterman and F. Sun. *Proc. Natl. Acad. Sci. USA*. **99**, 7335 (2002).

[7] G. Lancia. Tiling of a binary matrix. Unpublished manuscript. 2006.

[8] R.H. Hudson. *Bioinform.* **18**, 337 (2002).

[9] M.W. Nachman and S.L. Crowell. *Genetics*. **156**, 297 (2000).

[10] 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. *Genome Res.* **14**, 528 (2004).

[11] B. Rannala and Z. Yang. *Genetics.* **164**, 1645 (2003).

[12] The ENCODE Project Consortium. *Science*. **306**, 636 (2004).

[13] The International HapMap Consortium. *Nature*. **426**, 789 (2003).