Haplotype Motifs: An Algorithmic Approach to Locating Evolutionarily Conserved Patterns in Haploid Sequences

Russell Schwartz
Department of Biological Sciences
Carnegie Mellon University
Pittsburgh, PA 15213 USA
russells@andrew.cmu.edu

Abstract

The promise of plentiful data on common human genetic variations has given hope that we will be able to uncover genetic factors behind common diseases that have proven difficult to locate by prior methods. Much recent interest in this problem has focused on using haplotypes (contiguous regions of correlated genetic variations), instead of the isolated variations, in order to reduce the size of the statistical analysis problem. In order to most effectively use such variation data, we will need a better understanding of haplotype structure, including both the general principles underlying haplotype structure in the human population and the specific structures found in particular genetic regions or sub-populations. This paper presents a probabilistic model for analyzing haplotype structure in a population using conserved motifs found in statistically significant sub-populations. It describes the model and computational methods for deriving the predicted motif set and haplotype structure for a population. It further presents results on simulated data, in order to validate the method, and on two real datasets from the literature, in order to illustrate its practical application.

Keywords: haplotype block, SNP, dynamic programming, expectation maximization, significance testing

1 Introduction

The determination of consensus human genome sequences [7, 19] has revealed much about humanity's common genetic heritage, giving us a great deal of insight into the components of all human genomes, their common functional elements, and their general organizational principles. But while this information on the commonalities in our genomes has proven greatly beneficial, there is also much to be learned by studying how we differ from one another genetically. Studying genetic variations and correlating them with phenotypic variations is a key strategy for locating genes related to human diseases and developing diagnostics to locate people at particular risk for these diseases. Much recent research has therefore focused on the common variations in the human genome, which occur predominantly in the form of common single-base variations known as single nucleotide polymorphisms (SNPs).

It has so far proven difficult to use dense SNP maps directly for fine-scale mapping of complex disease genes, in large part because the many statistical tests performed in correlating the SNPs to a phenotype necessitate strict confidence bounds. There is therefore much interest in enhancing the power of the statistical methods through better modeling of genetic variation, particularly with regards to evolutionarily conserved patterns of polymorphisms known as haplotypes. One can take advantage of correlated SNP sites within conserved haplotypes directly through more sophisticated statistical models for association testing [11, 17, 12, 10]. Other approaches attempt to separate the statistical association problem from the inference of conserved haplotype structure. Most such approaches have been built on the idea of "haplotype blocks." [4] Given a block decomposition, one can use block haplotypes in place of isolated SNPs in association studies or use only subsets of "haplotype tagging" SNPs [8] adequate to characterize the block haplotypes. There are, however, drawbacks to the block idea. There is additional correlation information across blocks [6] as well as likely additional sub-structure within block regions that is lost to analysis when a block decomposition is imposed. Furthermore, the blocks themselves often appear imprecisely defined and may be difficult to derive reproducibly from reasonable sample sizes [16].

We therefore wish to develop alternative methods for understanding haplotype substructure that will take better advantage of data correlations by avoiding an explicit assumption of a global block structure. Two recent approaches at-

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

tempted to identify conserved sequence regions by predicting the likely piecewise ancestry of sequences. The first approach, simultaneously developed by Ukkonen [18] and Schwartz et al. [15], leverages block decompositions but attempts to deduce longer range connectivity by joining together haplotypes across block boundaries. The second approach, developed by Schwartz et al. [15], uses a specialized form of hidden Markov model to represent the haplotype structure in a population and interpret individual sequences. That latter approach addressed many of the problems of the prior work, such as detecting substructure at both finer and broader scales than in a population-wide block decomposition. At the same time, it presented problems in the ultimate application to association studies because of the difficulties in identifying specific regions for association testing. Furthermore, like the other combinatorial optimization methods in the literature, it did not distinguish statistically significant haplotype conservation from that observed only by chance in a particular population sample. Given the small population samples such work must deal with, an inability to distinguish significant from insignificant information could serious limit the method's applicability.

This paper describes a new approach to locating evolutionarily conserved structure in haploid sequences, which we call "haplotype motifs." Like haplotype coloring, but unlike block-based approaches, the haplotype motif method seeks to explain haplotype data on a sequence-by-sequence basis, without the presumption of a global block pattern. Unlike the haplotype coloring work, the haplotype motif approach incorporates a notion of statistical significance as a core element of its construction while still taking advantage of combinatorial optimization methods to simultaneously identify statistically over-represented motifs in haplotype data and interpret individual data in terms of those motifs. Informally, haplotype motifs provide a partial but confident picture of haplotype conservation, whereas haplotype coloring provides a best guess interpretation of all data regardless of confidence. Haplotype motifs are therefore likely to be better suited to the problem of statistical inference given finite datasets that is the central motivation of this work. We show using simulated data that for a wide range of confidence values, the method is very effective at detecting true motifs while avoiding false positives. We further demonstrate some of the practical value of the method by using it to analyze two real data sets.

2 Methods

The goal of the computational methods in this paper is to find a set of statistically overrepresented motifs, or strings of consecutive polymorphic alleles, in a set of aligned haploid sequences. Our input, the haploid sequences, can be treated as a set of binary strings of fixed length. Our output is a set of motifs — substrings of the haplotype strings — with associated frequencies of occurence. These substrings are chosen to satisfy the property that any motif's frequency is significantly higher than that of its best explanation in terms of smaller motifs, a concept further explained below.

We can informally explain the technique with an example. Suppose we sample a population at four polymorphic sites and observe the following haplotypes:

- "AAAA" with frequency $\frac{9}{16}$
- "AATT" with frequency $\frac{3}{16}$
- "TTAA" with frequency $\frac{3}{16}$
- "TTTT" with frequency $\frac{1}{16}$

Assuming an adequate sample size, we might conclude that the substring "AAXX" (A in the first two positions, any value in the last two) is over-represented because its frequency ($\frac{3}{4}$) is significantly higher than the frequency of $\frac{9}{16}$ that would be expected if its two component single-base alleles ("AXXX" and "XAXX", each with frequency $\frac{3}{4}$) were assumed to be sampled independently of one another. We could similarly conclude that "XXAA" is a motif. We would then conclude that "AAAA" is not a motif, even though it is more common than would be expected from its four component single-base frequencies, because its frequency can be explained by assuming it is the conjunction of the motifs "AAXX" and "XXAA" that were previously determined.

The overall approach used in this paper to find haplotype motifs depends on two key sub-methods: motif discovery, which locates overrepresented motifs; and sequence parsing, which optimally interprets sequences based on the conserved motifs. Both sub-methods rely on similar algorithms for finding the probability of constructing a given sequence from independent sub-sequences of known frequencies. These sub-methods are combined into an iterative expectation maximization (EM) algorithm [3] allowing for simultaneous optimization of motif frequencies and individual sequence parses.

2.1 Motif Discovery

To test whether a potential motif is statistically overrepresented, we require a null hypothesis for comparison. We define the expected frequency of a potential motif to be the maximum, over all possible ways of constructing the potential motif from smaller motifs, of the probability of choosing that particular combination of smaller motifs assuming they are sampled independently. Our null hypothesis is that the motif's occurrence can be described a binomial random variable whose probability is the expected frequency of occurrence and for which each observed sequence is one trial. We then ask whether the actual frequency of the motif is above an upper confidence bound on

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

the frequency of occurrence under the null hypothesis, using the approximation recommended by Agresti and Coull [1],

$$u = (\hat{p}+z^2/2m+z\sqrt{\hat{p}(1-\hat{p})/m + z^2/(4m^2)})/(1+z^2/m),$$

where $\hat{p}$ is the expected frequency, m is the population size, and z is the variate value that would be required to give a desired confidence interval for a standard normal distribution. That is, we declare a putative motif real if it is a significantly better explanation for the frequency of its substring than the best explanation in terms of smaller known motifs.

Given our definition, we can efficiently locate all true motifs by repeated application of a dynamic programming algorithm as follows. We first define the following:

m_{ijk} is the k^{th} motif of length i starting at position j; f_{ijk} is its frequency in the population

M_{ij} is the set of all motifs m_{ijk}

s_{ijk} is the sequence of polymorphic values contained in motif m_{ijk}

$s[j]$ is the j^{th} base of sequence s; $s[j,k]$ is the subsequence of s from base j to base k

p_{mut} is a prior probability of recent mutation in a base relative to its motif

$mut(b)$ is $1 - p_{mut}$ if b is true and p_{mut} otherwise

$P(s_1, s_2) = \prod_i mut(s_{1i} = s_{2i})$, the probability of generating s_1 from a motif of sequence s_2

H is the complete set of haploid input sequences

All of these values are trivially calculated from the data with the exception of f_{ijk}, which we can establish by the formula $f_{ijk} = \sum_{h \in H} P(h[j, j+i-1], s_{ijk})$, and p_{mut}, a user-specified parameter. p_{mut} can be estimated, given estimates of the effective population size and the number of generations we wish to consider "recent," using the same type of SNP data as that used as input for the present work [9]. Given the above values, we can find the most probable parse of potential motif $m_{i_0 j_0 k_0}$ by the following dynamic programming algorithm, which we will call MotifScore:

best$[j_0 - 1] \leftarrow 0$
for j = j_0 **to** $j_0 + i_0 - 1$
  best[j] $\leftarrow -\infty$
  for i = 1 **to** $i_0 - 1$
    for each $m_{i,j-i+1,k} \in M_{i,j-i+1}$
      score $\leftarrow$ *best*$[j - i] + \log f_{i,j-i+1,k} + \log P(s_{i_0 j_0 k_0}[j - i + 1, j], s_{i,j-i+1,k})$
      if (*score* > *best[j]*)
        best[j] $\leftarrow$ *score*
        parents[j] $\leftarrow$ *i*
        motifs[j] $\leftarrow m_{i,j-i+1,k}$
      end if
    end for each
  end for
end for

When the algorithm terminates, *best*$[j_0 + i_0 - 1]$ will contain the log probability of the best parse, which we can then compare with the measured frequency to accept or reject the potential motif. We can also backtrack using the *parents* and *motifs* arrays to recover the optimal parse, which is not necessary for motif discovery but will be relevant below. To discover all motifs in a population set, we only need to run this procedure on all potential motifs in order of increasing length, constructing M_{ij} for successively larger values of i using the motif sets already computed for smaller values of i.

Performing this analysis for all possible motifs would potentially require $O(|H|^2 n^5)$ time in the sequence length n and number of haplotypes $|H|$ because there are potentially $O(|H|n^2)$ motifs requiring $O(|H|n^3)$ time each. Fifth order dependence on sequence length would be prohibitive, but we can avoid it by imposing a maximum motif length. We choose a default limit of 20 SNPs, which, based on block-based analyses [6], should be well above the size of almost all true regions of sequence conservation for reasonable marker spacings. Setting a maximum motif length of n_{max} gives an asymptotic run-time of $O(|H|^2 n n_{max}^4)$.

2.2 Sequence Parsing

Given a set of over-represented motifs and their frequencies, we can determine the most probable explanation for any given sequence by a simple modification of the dynamic programming algorithm used in motif discovery. Specifically, running the MotifScore algorithm with the entire sequence under consideration in place of our potential motif $m_{i_0 j_0 k_0}$ will yield the cost of an optimal parse of the sequence, while backtracking provides the parse itself. It is then trivial to count the occurrences of all motifs and further refine our motif list by reapplication of the statistical test used in motif discovery.

There is a subtlety, though, in that the probability model we use in motif discovery differs from that we derive by counting motifs in parsed sequences. In motif discovery, the frequency of a motif is given by the expected number of observed sequences generated by the putative ancestral sequence the motif represents, allowing a given base to count towards potentially many distinct overlapping motifs. When we count in parsed sequences, any given base counts towards at most one motif. This latter definition is the more biologically reasonable one, but cannot be used in initial motif discovery because we do not have parses of sequences at that point. In order to refine our motif discovery using frequencies based on motif counts, we must redefine the expected frequency of a motif to be the sum, rather than maximum, of the probabilities of all parses. The sum is calculated by an algorithm similar to that used for the maximum:

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

sum$[j_0 - 1] \leftarrow 0$
for j = j_0 **to** $j_0 + i_0 - 1$
sum[j] $\leftarrow 0$
for i = 1 **to** $i_0 - 1$
for each $m_{i,j-i+1,k} \in M_{i,j-i+1}$
sum[j] $\leftarrow$ *sum[j]* + *sum*$[j - i](f_{i,j-i+1,k} + f_{i_0 j_0 k_0})P(s_{i_0 j_0 k_0}[j - i + 1, j], s_{i,j-i+1,k})$
end for each
end for
end for

When this algorithm completes, *sum*$[j_0 + i_0 - 1]$ will contain the sum of the probabilities of all possible parses of our putative motif in terms of smaller motifs, which can then be used as our predicted probability in the statistical test in order to screen out motifs found in earlier stages that are no longer sufficiently over-represented. Parsing all $|H|$ sequences, counting motif occurrences, and retesting motifs for significance requires worst case time $O(|H|^2 n n^3_{max})$ using a similar argument to that for asymptotic motif discovery run time.

2.3 Frequency Refinement

One problem with the above approach is that there is no guarantee that the sequence parses based on the estimated frequencies will yield motifs in proportions close to these initial estimates. In order to establish a set of frequencies consistent with the parses they derive, we combine the above motif discovery and motif parsing algorithms into an expectation-maximization (EM) algorithm [3]. We initialize the algorithm by estimating initial frequencies, allowing SNPs to count towards multiple motifs, then performing our motif discovery. We then use these initial motif frequencies as a first guess for motif parsing and iteratively reapply motif parsing, motif counting, and re-screening of motifs until we converge on a consistent set of motif frequencies. The algorithm is described by the following pseudo-code:

count initial frequencies of all observed sub-strings in the population
remove statistically insignificant motifs by the motif discovery algorithm
repeat
parse sequences in terms of conserved motifs
revise motif frequencies based on counts in parsed sequences
remove statistically insignificant motifs
until parse probabilities converge

Upon convergence, detected when parse probabilities stabilize, the procedure yields a final set of parses in terms of motifs and a final set of motif frequencies yielding those parses.

The asymptotic run-time of this procedure can be expressed as a function of the number of rounds of iteration, R, required for convergence; number of haplotypes, $|H|$; number of SNPs, n; and maximum motif size, n_{max}. The initial frequency estimation will require time $O(|H|^2 n n^2_{max})$, the initial motif discovery phase will require time $O(|H|^2 n n^4_{max})$, and each iteration of the EM loop will require $O(|H|^2 n n^3_{max})$, giving a total run time bounded by $O(R|H|^2 n n^3_{max} + |H|^2 n n^4_{max})$. These bounds, will, though generally be extremely pessimistic in practice, because the number of distinct subsequences and true motifs will be much less than the theoretical maximum of $O(|H| n n_{max})$ for real data. We can also empirically assert that R is generally very small (about ten) and that initial motif discovery, not the EM loop, dominates run time. Practical run-time will depend on many other factors — e.g. the significance level, the mutation probability, and the degree of sequence conservation in a given data set — in ways that are difficult to analyze. Practical run times seem reasonable for standard parameter values and realistic sizes of data sets, though. Using a 1.2 GHz Pentium computer and the program default settings (mutation probability 0, p-value 0.0005, maximum motif length 20), a simulated dataset of 1000 random sequences on 100 SNPs requires 54.8 seconds, while a real dataset of 142 sequences on 71 SNPs requires 2.6 seconds.

3 Results

In this section, we describe application of the haplotype motif analysis method to a combination of simulated and real data. We use simulated data to validate the method and characterize its performance under different conditions. The real data illustrates how the method might be useful in practice and allows us to draw some conclusions about the similarity of real data to particular kinds of simulated data. Except where noted below, all tests were performed with a maximum motif length of 20 SNPs and a zero mutation probability.

3.1 Validation on Simulated Data

We first attempted to assess the method's effectiveness at screening out false motifs that are only overrepresented by chance in a particular population by using simulated populations lacking haplotype structure. We created ten simulated populations with 100 SNP sites for each of seven different sizes (10, 25, 50, 100, 250, 500, and 1000 chromosomes). At each site in each population, we defined a major and minor allele (arbitrarily chosen to be "A" and "T"

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

sample size	$p = 10^{-1}$	$p = 10^{-2}$	$p = 10^{-3}$
10	138	0	0
25	359	0	0
50	742	0	0
100	2700	23	0
250	3628	19	0
500	6695	54	0
1000	13159	119	5

Table 1. False positive motifs detected in random data as a function of p-value for different population sizes. The p-value is one half of one minus the confidence interval size. No false positives were found for p-values below 10^{-3}.

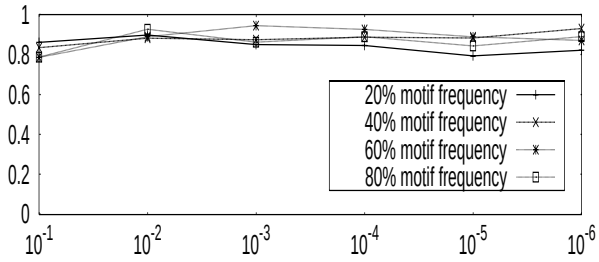

Figure 1. Accuracy on consecutive pairs of sites in detecting the structure of true motifs embedded in random data as a function of p-value for four motif frequencies.

respectively in all cases) with the frequency of the minor allele at site i, f_i, chosen uniformly at random on [0.1,0.5]. We then sampled each site in each sequence independently from these site-specific distributions. Finally, we ran the algorithm with a range of significance levels for each population, recording the total number of distinct motifs detected in the ten data sets of each size for six significance levels. Table 1 shows the results, indicating that for moderate confidence bounds, the method is very effective at avoiding false positives.

In order to assess the false negative rate, we used similar sets of simulated data but embedded conserved motifs within them. For these tests, we generated data sets as before, but with a uniform population size of 250 chromosomes. We then embedded a conserved motif of length 10 at a random position in each data set in a fraction f of the sequences, where f varies in different data sets from 20% to 80%, and tested our ability to detect the motif again using a range of confidence intervals. In all cases, the method predicted a conserved motif at least partially overlapping the true embedded motif, although the boundaries of the true motif were often incorrect; in such cases, assigned motifs were typically subsets of the true motifs. We quantitatively measured accuracy by the fraction of consecutive pairs of bases, starting one base before the true motif and ending one base after it, that were assigned to motifs consistent with the true motif structure. That is, a pair is a true positive if both fall within the true motif and they are assigned to the same predicted motif or if they lie on opposite sides of the motif boundary and the one within the boundary is assigned a motif while the one outside is not. Figure 1 shows the results, indicating high accuracy in detecting true motif structure. One frequent problem not reflected in these statistics is the detection of false "anti-motifs" reflecting the absence of a substring of the true embedded motif; these "anti-motifs" are true motifs given our statistical definition, but are not clearly biologically meaningful.

One final issue of significant practical concern is the ability of the method to detect population-wide block structure. We created a collection of what we call "strongly blocked" data sets, meaning that the data is generated from discrete blocks with complete independence between blocks. To generate each block, we chose a random block length uniformly between 2 and 20 SNPs and a random number of alleles uniformly between 2 and 4. We then generated random sequences of A's and T's representing the distinct alleles, recreating any sequence that was identical to a previous one, and removed any invariant sites. We then selected frequencies for all block alleles such that all possible frequencies were equally likely subject to the constraint of a minimum frequency of 10% for each allele. We generated ten such blocks per sequence for ten data sets for each of six population sizes (10,25,50,100,250,500,1000) and ran our motif detection method for all such populations with a range of p-values. We quantified accuracy by the fraction of pairs of consecutive SNPs assigned motifs consistent with the block structure (i.e. pairs within a true motif assigned the same motif and pairs on opposite sides of a true motif boundary assigned to different motifs). Figure 2A shows the results, demonstrating that reasonable population sizes allow high accuracy in detecting true block structure for a broad range of p-values. Figure 3A shows an example of the motifs detected in a "strongly blocked" population of 100 chromosomes. The image clearly shows the block structure, although there are errors near block boundaries.

Because this conception of "strongly blocked" sequences is deliberately a caricature of the noisier block structure that one would expect in true sequences under the block hypothesis, we conducted further tests using the same populations adding a model of noise. We re-ran the same block tests as above after randomly flipping 5% of the bases in the sequences. Figure 2B shows the quantitative results

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

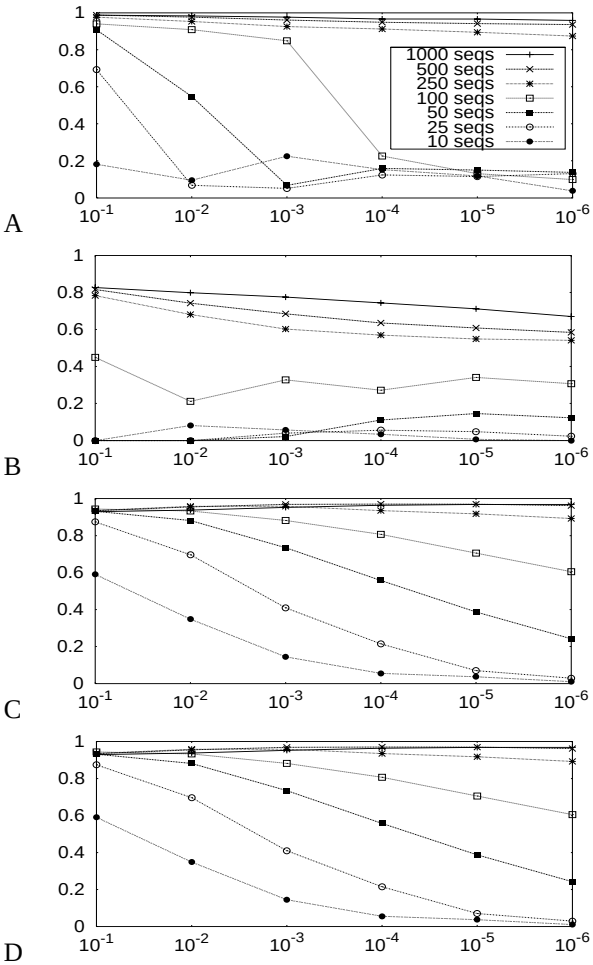

Figure 2. Accuracy on consecutive pairs of sites in detecting true block motif structure as a function of p-value for six population sizes of "strongly blocked" data. A: noise-free, 0% mutation probability; B: 5% noise, 0% mutation probability; C: 5% noise, 5% mutation probability; D: 5% noise, 2.5% mutation probability.

of the noise, displaying a noticeable loss in accuracy in detecting block boundaries. Figure 3B shows the motif coloring produced for the same data set as in figure 3A but with the 5% noise added. Elements of block structure are still evident in frequently-used boundaries between motifs, although many sub-sequences are not assigned motifs and many motifs are not detected or are only partially detected. Activating the mutation model in the haplotype motif detection procedure with a 5% mutation probability, however, largely recovers the block motif structure (Figure 2C, Figure 3C), although it comes at the expense of occasionally conflating similar haplotypes within a block. With the prior mutation probability set to 2.5%, results are nearly as good (Figures 2D, 3D). Thus, even when the prior mutation probability is poorly matched to the true mutation rate, using the mutation model allows us to largely eliminate the effects of moderate random noise on our ability to recover true block structure.

3.2 Application to Real Data

We further applied the methods to two real datasets: a set haploid sequences from 22 biallelic variations (21 SNPs and one two-base deletion polymorphism) in 192 chromosomes for the apolipoprotein E (APOE) gene [13, 5] and a set of computationally inferred haplotypes from 71 SNPs in 142 chromosomes for the lipoprotein lipase (LPL) gene [14]. We used a p-value of 0.001 for both data sets. Due to the possibility of noise in the data from recent or recurrent mutations or errors in the assays, we used a prior mutation probability of 0.01.

Figure 4 shows the motif patterns detected for the two datasets. APOE shows a substantially simpler motif structure than LPL, although this may be an artifact of the computational haplotype inference used with LPL. These images are strikingly different from those produced by "strongly blocked" data in figure 3, with many conserved segments cutting across one another's boundaries. Given the demonstrated effectiveness of the haplotype motif method in finding true motifs and global block patterns and in avoiding false motifs, these differences suggest the real data genuinely contains a considerable amount of substructure that is not captured by the assumption of globally conserved discrete blocks.

A possible partial explanation for this inconsistent block structure is that it reflects population sub-structure. Both data sets used samples from three distinct ethnic groups: African-Americans from Jackson, Mississippi; European-Americans from Rochester, Minnesota; and Europeans from North Karelia, Finland. APOE also used a fourth population of Mayans from Campeche, Mexico. Separating the sequences by the populations from which they are derived gives the images shown in figures 5 and 6. For APOE (Figure 5), the more common motifs (blue and green) are conserved across all four populations while the rarer motifs (brown and white) are enriched in the European and Mayan populations, rare in the European-American, and absent in the African-American. LPL (Figure 6) shows similar patterns of population-specific enrichment or depletion for many individual motifs. In particular, the African-American population seems to have a noticeably lower representation of some of the motifs most common in the other populations.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

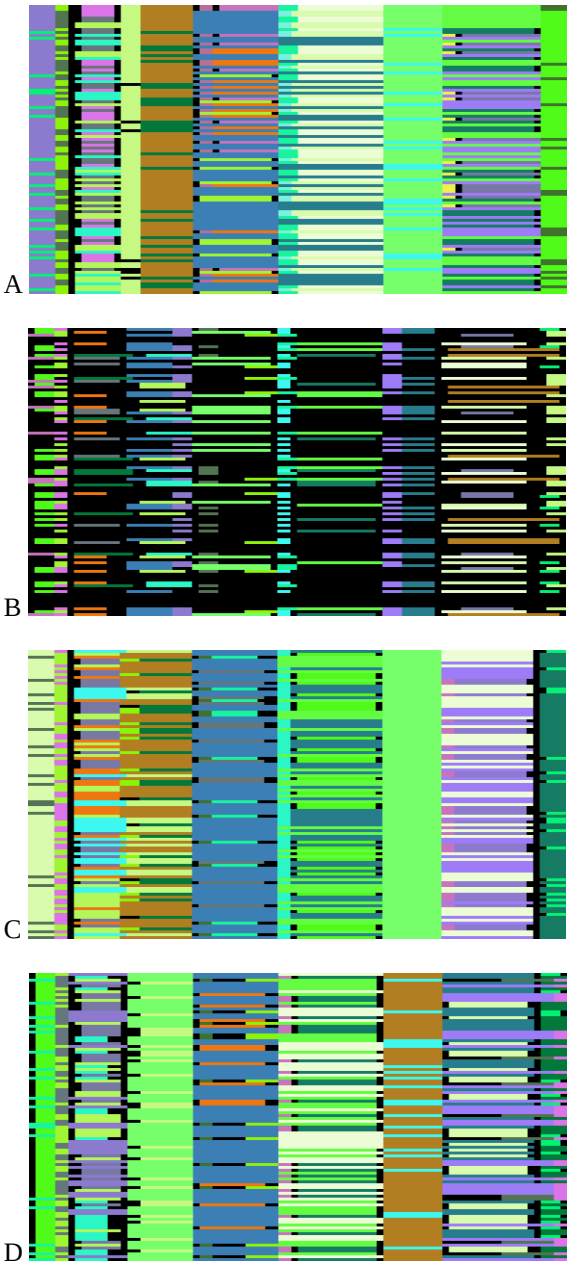

Figure 3. Motifs detected in 100 simulated "strongly blocked" chromosomes. A: noise-free, 0% mutation probability; B: 5% noise, 0% mutation probability; C: 5% noise, 5% mutation probability; D: 5% noise, 2.5% mutation probability. Unbroken segments of uniform color reflect a single motif. There is no correspondence between colors in different images.

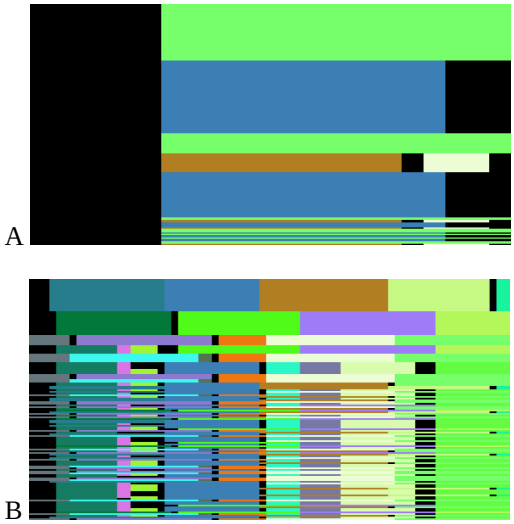

Figure 4. Motif profiles for the A: APOE and B: LPL data sets.

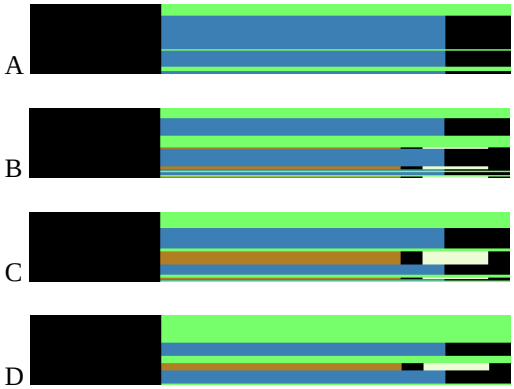

Figure 5. The APOE data set subdivided by population. A: African-American; B: European-American; C: European; D: Mayan.

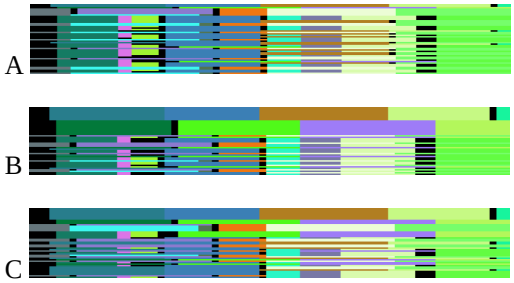

Figure 6. The LPL data set subdivided by population. A: African-American; B: European-American; C: European.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

4 Discussion

We have developed a mathematical model and associated algorithms for detecting conserved substructure in haploid genetic sequences. We have implemented this model in a computational tool and demonstrated its effectiveness on several forms of simulated data. The method is very accurate at finding true motifs, avoiding false motifs, and detecting global block structure over a range of confidence levels. The most frequent mistake it makes appears to be detecting a sub-motif rather than a larger true motif in which it is embedded. We have also applied the method to two real data sets, providing insight into the structure of two genetic regions and into variations in gene structure across different populations.

This technique has several potential uses. The primary intended benefit is to increase the power of case-control association studies by allowing conserved sequence regions to be substituted for their component SNPs in association tests. The method's value for this purpose can only be established by showing it to be effective in locating disease-related variants, a project beyond the scope of this first theoretical work. Another potential application will be in analyzing population histories. As figures 5 and 6 demonstrate, different sub-populations tend to show distinct conserved motifs, making haplotype motifs a potentially useful supplement to existing methods for tracking the migration or intermixing of populations over long time periods. There are also downstream problems for which this technique might prove a useful first step, such as choosing informative "haplotype tagging" SNPs [8, 20, 2].

The results on real data provide some insight into the structure of genetic variation in the human genome. The two genes examined display very different patterns of haplotype conservation, suggesting that such patterns may be more generally variable across the human genome. Furthermore, motif profiles of the real data differ noticeably from those for an idealized conception of haplotype blocks. This result could reflect poor parameter choices, although that seems unlikely given the relative robustness of the methods to parameter choices on simulated data. Even if this observation should prove valid, it would not mean that much of the information in these sequences could not be captured in global haplotype block structures nor that doing so would not be useful. It would, however, suggest that global haplotype blocks may not be the most parsimonious way to explain real sequence data, nor perhaps the way most reflective of actual patterns of evolutionary haplotype conservation.

There are many avenues by which this work might be continued. Algorithmic improvements could likely be achieved in solving the problem as specified above and in finding fast approximations to it. The statistical model could also be improved, for example by developing a notion of confidence intervals better adapted specifically to genetic variations. Furthermore, the bottom-up approach to motif generation seems to produce some undesirable artifacts that argue for experimenting with other models. There is still much to do to apply haplotype motifs to downstream problems, such as informative SNP selection and actual case-control association testing. Finally, there is likely much to be learned about both the haplotype motif approach and haplotype structure in general by applying the method to a wide variety of data sources representing different genome regions and populations.

Code availability: The algorithms described in this paper have been implemented in the program HapMotif. Preliminary source code in C++, precompiled binaries, and associated documentation are available at **http://www-2.cs.cmu.edu/~russells/software/hapmotif.html**. Perl scripts used to generate simulated data sets and run and analyze the simulation experiments will be provided upon request.

Acknowledgments: I thank Andrew G. Clark for providing an electronic version of the APOE and LPL datasets used in this analysis and for helpful comments on the manuscript.

References

[1] A. Agresti and B. A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52:119–126, 2002.

[2] V. Bafna, B. Halldórsson, R. Schwartz, A. G. Clark, and S. Istrail. Haplotypes and informative SNP selection algorithms: Don't block out information. In *Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB-03)*, pages 19–27, 2003.

[3] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probability functions of Markov chains. *Annals Math Stat*, 41:164–171, 1970.

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

[5] S. M. Fullerton, A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor, J. H. Stengaard, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle, and C. F. Sing. Apolipoprotein E variation at the sequence haplotype level: implications for the origins and maintenance of a major human polymorphism. *Am J Hum Gen*, 67:881–900, 2000.

[6] S. Gabriel, S. Schaffner, H. Nguyen, J. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. Lander, M. Daly, and D. Altschuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

[7] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[8] 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*, 29:233–237, 2001.

[9] M. K. Kuhner, P. Beerli, J. Yamato, and J. Felsenstein. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics*, 156:439–447, 2000.

[10] J. S. Liu, C. Sabatii, J. Teng, B. J. B. Keats, and N. Risch. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res*, 11:1716–1724, 2001.

[11] 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 Gen*, 65:858–875, 1999.

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

[13] D. A. Nickerson, S. L. Taylor, S. M. Fullerton, K. M. Weiss, A. G. Clark, J. H. Stengaard, V. Salomaa, E. Boerwinkle, and C. F. Sing. Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res*, 10:1532–1545, 2000.

[14] D. A. Nickerson, S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson, J. H. Stengaard, V. Salomaa, E. Vartiainen, E. Boerwinkle, and C. F. Sing. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat Genet*, 19:233–240, 1998.

[15] R. Schwartz, A. G. Clark, and S. Istrail. Methods for inferring block-wise ancestral history from haploid sequences: The haplotype coloring problem. In *Lecture Notes in Computer Science 2452 (Proceedings of the Second International Workshop on Algorithms in Bioinformatics)*, pages 44–59, 2002.

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

[17] 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 Gen*, 64:1728–1738, 1999.

[18] E. Ukkonen. Finding founder sequences from a set of recombinants. In *Lecture Notes in Computer Science 2452 (Proceedings of the Second International Workshop on Algorithms in Bioinformatics)*, pages 277–286, 2002.

[19] G. Venter, M. A. Adams, E. W. Myers, et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[20] K. Zhang, M. Deng, T. Chen, M. Waterman, and F. Sun. A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci USA*, 99:7335–7339, 2002.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**