Identification of polymorphic motifs using probabilistic search algorithms

Analabha Basu,[1] Probal Chaudhuri,[2] and Partha P. Majumder[1,3]

[1]*Human Genetics Unit and* [2]*Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata, 700108 India*

The problem of identifying motifs comprising nucleotides at a set of polymorphic DNA sites, not necessarily contiguous, arises in many human genetic problems. However, when the sites are not contiguous, no efficient algorithm exists for polymorphic motif identification. A search based on complete enumeration is computationally inefficient. We have developed probabilistic search algorithms to discover motifs of known or unknown lengths. We have developed statistical tests of significance for assessing a motif discovery, and a statistical criterion for simultaneously estimating motif length and discovering it. We have tested these algorithms on various synthetic data sets and have shown that they are very efficient, in the sense that the "true" motifs can be detected in the vast majority of replications and in a small number of iterations. Additionally, we have applied them to some real data sets and have shown that they are able to identify known motifs. In certain applications, it is pertinent to find motifs that contain contrasting nucleotides at the sites included in the motif (e.g., motifs identified in case-control association studies). For this, we have suggested appropriate modifications. Using simulations, we have discovered that the success rate of identification of the correct motif is high in case-control studies except when relative risks are small. Our analyses of evolutionary data sets resulted in the identification of some motifs that appear to have important implications on human evolutionary inference. These algorithms can easily be implemented to discover motifs from multilocus genotype data by simple numerical recoding of genotypes.

[Supplemental material is available online at www.genome.org. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: A. Chowdhury.]

Single nucleotide polymorphisms (SNPs) are abundant in the human genome and occur at roughly 1 per 2 kb spacing on the average (Balasubramanian et al. 2002). Alleles at SNP loci are often nonrandomly associated. Various evolutionary mechanisms, including drift and natural selection, maintain the association of specific nucleotides at two or more sites, which may not be contiguous. The search for nucleotides that exhibit association at a set of polymorphic sites is of interest in studies of common diseases (Sabeti et al. 2002) and in evolutionary genetics (Tateno et al. 1997; Daly et al. 2001). We define a set of nucleotides that occurs with a high frequency at multiple polymorphic DNA sites, not necessarily contiguous, in a group of individuals as a "polymorphic motif." We note that our definition of a motif differs from the conventional definition, for example, that is used for finding regulatory sequences in promoter regions of genes (Keiler and Shapiro 2001), in two ways; (1) the sites included in our definition are polymorphic, and (2) the sites need not be contiguous. In conventional motif-identification problems, search is made for evolutionarily conserved nucleotide sequences at a contiguous set of nucleotide positions (Gupta and Liu 2003). In case-control studies of common diseases, it is of interest to find polymorphic motifs and to test whether there are differences in motif frequencies between cases and controls (Khani-Hanjani et al. 2002). Motifs that are found in significantly higher frequencies among cases are associated with the disease under study. If variants in multiple genes are indeed involved in the disease, the sites in such a motif may not be contiguous. Similarly, the discovery of polymorphic motifs is important in evolutionary genetics. Indeed, such motifs have been used to define subhaplogroups of specific clades (haplogroups) of the human mitochondrial (mt) DNA (Bamshad et al. 2001).

In the context of evolutionary or human genetic studies, there are two related issues. First, to identify motifs or haplotypes that occur at high frequencies in subsets of a large data set, such as those sampled from specific geographical regions or groups, or from individuals afflicted with a specific disease. Having identified such motifs, the second problem is to decipher the biological or population genetic processes (e.g., linkage, drift, selection, epistasis) that have resulted in the existence of these high-frequency motifs. In this study, we shall only address the first issue, viz., how to identify high-frequency motifs. To address the second issue, collection of further data (e.g., family data), statistical modeling, investigations of metabolic pathways, wet-laboratory experimentation, etc., may be required.

It is theoretically possible to discover polymorphic motifs in a set of N-aligned DNA sequences, each of length L nucleotides, by examining frequencies in all possible $k \times k$ tables, $k = 2,3,\ldots,L$. However, this is computationally infeasible. The purpose of this study is to propose a set of computationally fast probabilistic search algorithms that may be used for motif finding, and to evaluate their efficiencies using both synthetic and real data sets. Keeping SNP loci in mind, which are usually biallelic, we formulate, describe, and assess these algorithms using sequences of binary characters. However, there is no inherent restriction in these algorithms that the search has to be confined to binary sequences. These algorithms can also be used on multilocus genotype data of diploid individuals. When genotype data are used, the distinct genotypes only need to be numerically recoded, as discussed later. Thus, the proposed algorithms are fairly general in nature, and can be put to diverse uses.

We first propose an algorithm for identifying a motif of a given length. We then extend this algorithm when the length is

unknown. Finally, we propose a modification for identifying "variant" motifs. The problem of identifying a variant motif arises when, given a collection of DNA sequences derived from a set of individuals, it is of interest to identify whether an appropriately defined subset of individuals in this collection possesses a motif that is different from that possessed by the remaining subset of individuals. For example, in a case-control study, it is pertinent to identify whether the cases possess a motif at a certain number of sites that comprise nucleotides, each of which is different from the nucleotide possessed by the controls at the corresponding sites. Identification of such a variant motif can help in identifying SNPs associated with the disease in question. The problem of identifying variant motifs in subsets of a collection of sequences at the hypervariable segment-1 (HVS1) of human mtDNA has received a lot of attention (Quintana-Murci et al. 1999). In particular, efforts have been made to to identify contrasting motifs in HVS1 in subsets of individuals belonging to different haplogroups (HGs) that are defined by the presence of specific nucleotides at sites outside of the HVS1.

In each of these problems, a variant motif is defined in relation to another. For example, for case-control data, the variant motif among cases is defined in contrast to the one found among the controls. In the evolutionary analysis of mtDNA HVS1 sequences, search for a motif is made in contrast to the Cambridge Reference Sequence (CRS) (Anderson et al. 1981). In such cases, for motif searching, not only do we have to find a high-frequency motif, but this motif should contain nucleotides that are completely or largely different from those present in the reference sequence at the corresponding nucleotide positions.

Methods

Consider a data matrix $((a_{ij}))_{N\times L}$, where a_{ij} denotes a nucleotide (A,T,G, or C) at the j^{th} polymorphic site ($j = 1,2,\ldots,L$) for the i^{th} individual ($i = 1,2,\ldots, N$). The data matrix is generated from aligned DNA sequences of a specific genomic segment of N individuals, from which all monomorphic sites have been removed. We note that if these N individuals belong to a case-control study, then the data matrix needs to be initially created by pooling all cases and controls, and subsequently separated into two matrices, one for cases and another for controls. A similar strategy is also required in evolutionary studies, while simultaneously dealing with two populations. We also note that if disjoint segments of DNA are to be simultaneously examined for motif finding, then appropriate segments may be separately aligned, and the aligned segments concatenated in the data matrix.

Let $V = \{1,2,\ldots,L\}$ denote the set of all L polymorphic sites in the data. Let Π_p denote the set of all possible combinations of p sites in V. To fix ideas, consider the data matrix given in Table 1. In this matrix, $N = 4$, and $L = 7$. Thus, $V = \{1,2,\ldots,7\}$. For $p = 2$, $\Pi_2 = \{\{1,2\}, \{1,3\},\ldots, \{1,7\}, \{2,3\}, \{2,4\},\ldots, \{6,7\}\} = \{V_2^k\}$, $k = 1,2,\ldots\binom{7}{2}$. In general, $\Pi_p = \{V_p^k\}$, $k = 1,2,\ldots\binom{L}{p}$ and $V_p^k = \{x_1^k, x_2^k,\ldots,x_p^k : x_i^k \in V\}$. For a fixed k, we define the *modal sequence* on V_p^k as that particular combination of nucleotides at the sites $\{x_1^k, x_2^k,\ldots,x_p^k\}$ included in V_p^k, $k = 1,2,\ldots\binom{L}{p}$ which has the highest frequency. In the data matrix of Table 1, the modal sequence, for example, on $V_2^1 = \{1,2\}$ is AG with frequency 2, on $V_2^2 = \{1,3\}$ is AT with frequency 3, etc. We define a motif of length p as the maximally frequent modal sequence on Π_p; that is, the sequence that occurs with the highest frequency (globally modal) among modal sequences on $V_p^1, V_p^2,\ldots, V_p^{\binom{L}{p}}$. In our example, the motif of length 2 is AT on $V_2^2 = \{1,3\}$ with frequency 3.

In general, the problem of finding a motif of length p from an $N\times L$ data matrix reduces to identifying the set V_p^k, $k = 1,2,\ldots,\binom{L}{p}$, from Π_p, such that the modal sequence on V_p^k is globally modal. With an $N\times L$ data-matrix, the search space Π_p has $\binom{L}{p}$ elements. Obviously, each element of Π_p is a string, S, comprising the identities of those specific p sites chosen out of L. There are $\binom{L}{p}$ such strings in Π_p. An exhaustive search of this space Π_p is computationally very expensive, and perhaps infeasible. We propose a stochastic search method, similar in spirit to the Metropolis-Hastings version (Metropolis et al. 1953) of simulated annealing, which is computationally fast and efficient. In this method, we maximize an objective function, $G(S)$, that is naturally defined as the "frequency of the modal sequence on the string $S \in \Pi_p$." By our definition, maximizing the frequency of the modal sequence on Π_p leads to identification of the motif of length p. Thus, the search comprises choosing both sites and characters at these sites, so that the chosen set of characters at the chosen set of sites has the maximum frequency in the data set.

Table 1. An example of a data matrix

Sequence/ Individual no.	Variant site no.						
	1	2	3	4	5	6	7
1	A	A	T	T	G	C	C
2	A	G	T	C	G	C	T
3	A	G	T	T	A	C	T
4	G	G	C	C	A	T	T

Algorithm for finding a motif of a given length and assessing its statistical significance

Although in real problems, the motif length is usually unknown, for ease of exposition, we first describe an algorithm for a known motif length p, and then generalize it to the case of an unknown motif length. Instead of maximizing $G(S)$, we shall consider the equivalent problem of minimizing a monotonically decreasing function, $H(S)$, of $G(S)$. The algorithm is iterative. We start with an arbitrary string S of length p; that is, a set of p distinct nucleotide sites drawn randomly from the L polymorphic sites. In each iterative step, an element (a specific site) of the string S is updated. The updating procedure requires the computation of $G(S)$, which is done from the frequency distribution of all unique sequences at the sites included in the string S. For this purpose, given a specific string of sites, S, of length p, we enumerate from data the frequencies (f_l) of all unique nucleotide sequences ξ_l ($l = 1,2,\ldots$), at the sites included in S. At each iterative step, we update a single site, and after p such iterative steps, we get a completely updated string. The procedure of updating a string completely is called a *sweep*. Thus, a sweep comprises p iterative steps. Let S_t denote the updated string after t sweeps.

We shall use the following notations:

1. Let $S_t = (x_t^{(1)}, x_t^{(2)},\ldots, x_t^{(p)})$, where $x_t^{(i)}$ denotes the i^{th} element (a site) of the string at the $(t+1)^{th}$ sweep.
2. Let $S_t^{(i)}$ denote a string in the $(t+1)^{th}$ sweep, whose first i ($0 \le i \le p-1$) elements have already been modified.
3. Let $S_t^{(i)}(y)$ denote a string in the $(t+1)^{th}$ sweep, whose first i ($0 \le i \le p-1$) elements have already been modified and the $(i+1)^{th}$ element is replaced by element y.
4. Let $H_t^{(i)}$ denote the minimum value of $H(S)$ after completion of the i^{th} iterative step in the $(t+1)^{th}$ sweep step.

5. Let $M_t^{(j)}$ denote the string of elements (array of sites) corresponding to $H_t^{(j)}$.

We initially set $H_0^{(0)} = 0$, and $M_0^{(0)}$ as a "null" string, that is, a $1 \times p$ vector whose elements are all set to zero. The updating procedure for the i^{th} element in the $(t+1)^{th}$ sweep uses the idea underlying the Metropolis-Hastings algorithm (Metropolis et al. 1953), which can be described as follows:

We first calculate $\beta_t = c.1n(t+1)$; where c is a constant and $c > 0$. One element (x) is selected at random from the set $V \backslash S_t^{(i-1)}$; that is, from the set $V = \{1,2,\ldots, L\}$, from which the elements included in the set $S_t^{(i-1)}$ have been removed. We then probabilistically update $x_t^{(i)}$ to $x_{t+1}^{(i)}$ according the following rule:

$$x_{t+1}^{(i)} = \begin{cases} x & \text{with probability } min(\Lambda, 1) \\ x_t^{(i)} & \text{with probability } 1 - min(\Lambda, 1) \end{cases}$$

where, $\Lambda = e^{-\beta_t |H(S_t^{(i-1)}(x)) - H(S_t^{(i-1)}(x_t^{(i)}))|}$.

Obviously, the transition probability from one string to another depends only on the outcome of the current string (Markov property). As is easily understood from the above updating rule, at any step of the iteration, a new string that yields a smaller value of $H(S)$ is always accepted, but to avoid being trapped at a local minimum, the new string with higher value of $H(S)$ may also be retained with a small probability (that crucially depends on the preassigned control parameter c and the corresponding sweep step t). It may be noted, however, that as the number of sweeps, t, increases, the process stabilizes. In other words, the probability of accepting a worse string decreases as t increases. The algorithm converges to the global minimum if β_t increases to infinity logarithmically (Winkler and Lutz 2003), and the speed of convergence is determined by the local oscillations of the function H at various coordinates of its argument. (Detailed results on convergence of nonstationary Markov chains can be found in Winkler and Lutz [2003]). It is clear that our choice of β_t satisfies this general property. In practice, it is important to start with a small value of c (say, five), but also to try with larger values of c to examine convergence to the same optimal value of the function and the rate of convergence. Large values of c can substantially speed up convergence, but can also result in the algorithm being trapped in a local minimum, and a large number of sweeps may be required to get out of the trap.

After each iteration, we compare $H(S_t^{(i-1)}(x))$ with $H_t^{(i-1)}$. If, $H(S_t^{(i-1)}(x)) < H_t^{(i-1)}$, then $H(S_t^{(i-1)}(x))$ is the new value for $H_t^{(i-1)}$ and $M_t^{(i)}$ is the updated string $S_t^{(i-1)}(x)$. Otherwise, we do not change $H_t^{(i-1)}$ and $M_t^{(i)}$. In each iteration, therefore, we compare the value of the objective function with the smallest value it has attained thus far. (If that smallest value remains the same over a large number of consecutive sweeps, then the entire procedure may have to be restarted with a new randomly chosen string. This is standard in most numerical optimization procedures.) This introduces the concept of elitism in our algorithm, which is popular in evolutionary computation (Goldberg 1989), and is done to retain the best value that was achieved during the entire run. Using available convergence results (Liu 2001; Winkler and Lutz 2003), it can be shown that as the number of sweeps goes to infinity, the value of the objective function converges to the global minimum. In practice, however, the procedure needs to be terminated after a finite number of sweeps. We have terminated when an upper bound (usually taken to be a large number; we have used the value of 5000 in our analyses) on the total number of sweeps was reached.

We note that, as with all numerical optimization procedures, it is desirable to repeat the procedure a certain number of times from different starting strings, and examine whether convergence to the same optimal value is obtained. The number of repetitions of the procedure that is practically feasible obviously depends on the availability of computing resources.

Having discovered a motif of a given length p in a data set, it is important to assess the statistical significance of the discovery. For this, we need to estimate the probability of existence of a motif of length p in a "random" data set of "similar" structure as the real data set in terms of nucleotide composition (as explained in detail in the Results section), that has a frequency higher than the motif discovered in the real data set. If this probability is smaller than a preassigned value (say, 0.05), then the motif that has been discovered can be declared to be statistically significant. To estimate this probability, we created a large number of random data sets, by randomly permuting the elements of each column of the real data set. For each random data set thus created, we used our algorithm to discover the motif of length p with the highest frequency, that is, the "best" motif. The proportion of random data sets in which the best motif had a frequency higher than that of the motif discovered in the real data set provided an empirical estimate of statistical significance. We note that for this purpose, ideally, the best motif in each random data set should be identified by a complete enumeration search, and not by using the algorithm proposed by us. However, this is infeasible unless the real data set is small. (We have actually carried out the complete enumeration search in many small data sets; the results are presented later.)

Extension of the algorithm when the motif length is unknown and assessment of statistical significance

In practical applications, the motif length (p) will usually be unknown. When p is unknown, one can start with a small value of p and increase this value sequentially, examining for each value of p the extent of decrease in the value of $G(S)$. One can stop with that value of p when an increase to $(p+1)$ results in a "substantial" drop in the value of $G(S)$. In practice, two values, p_{min} and p_{max} may be specified, and search for p may be made in the interval $[p_{min}, p_{max}]$. We now need a measure to evaluate whether the drop in the value of $G(S)$ for two consecutive values of p is substantial to stop the iterative algorithm.

For any given value of the motif length $p \in [p_{min}, p_{max}]$, we can use the algorithm described for identifying a motif of a given motif length, and obtain the (maximum) value of $G(S)$ given p, which we shall denote as $G(S|p)$. We, therefore, calculate $G(S|p_{min})$, $G(S|p_{min}+1),\ldots,G(S|p_{max})$. Let $d(p_i)$ denote the value of $G(S|p_i) - G(S|p_i+1)$, where $p_i \in [p_{min}, p_{max}-1]$.

To assess the statistical significance of a decrease in $G(S|p)$ as the motif length (p) is increased, we propose the following criterion. Let

$$\overline{d(p_j)} = \sum_{p_j=p_{min}}^{p_j-1} \frac{d(p_j)}{(p_j - p_{min})},$$

and

$$\sigma^2(p_j) = \sum_{p_j=p_{min}}^{p_j-1} \frac{[d(p_j)]^2}{(p_j - p_{min})} - [\overline{d(p_j)}]^2; p_{min} < p_j < p_{max}.$$

If $(d(p_j) - \overline{d(p_j)}) > 2.\sigma^2(p_j)$, then we declare the decrease from $G(S|p_j)$ to $G(S|p_j+1)$ as significant, and stop with the motif length

p. The idea underlying this criterion is that we declare a drop in the value of the objective function to be statistically significant if this drop differs from the mean of all previous drops by more than two times the variance of all previous drops. In the rare event that $\sigma^2(\ p_i) = 0$, we use the stopping criterion $G(S|p_i - 1) > 2.G(S|p_i)$, and declare the length of the motif as p_i.

Although the above method of assessment of statistical significance is intuitively appealing, the choice of the value of the constant (=2) in the stopping criterion is somewhat arbitrary. Further, in the above search procedure, it is possible that the sets of sites included in motifs of length p and $(p + 1)$ are disjoint. In many practical applications, this may not be desirable. Therefore, after the initial stage, new sites should be added to the set of sites included in the motif discovered thus far. Such an addition is made by searching for a site from among those sites not included in the identified motif. This strategy is not only more meaningful in many practical applications, but is also computationally less expensive. However, there is a trade-off. After convergence of this procedure, it is possible that the identified motif of length q (say) is suboptimal among all motifs of length q. When this procedure is adopted, we suggest the use of the criterion described below to assess statistical significance of increase of motif length from p to $(p + 1)$. Let π_p and π_{p+1} denote the probabilities of occurrence of the motifs of lengths p and $(p + 1)$, respectively. Let θ_{p+1} denote the probability of the nucleotide at the new site included in the motif as its length is increased from p to $p + 1$. We now wish to test the null hypothesis $H_0 : \pi_{p+1} = \pi_p \times \theta_{p+1}$, versus the alternative hypothesis $H_1 : \pi_{p+1} > \pi_p \times \theta_{p+1}$. In other words, we wish to test whether the additional site and the nucleotide at this site that were included to expand the motif of length p to $p + 1$ is associated, to a greater degree, with those sites and nucleotides already included in the motif (of length p) than is expected by chance. The level of significance of the test is given by: $\Sigma_{i=n}^{N} \binom{N}{i} (\hat{\pi}_p \times \hat{\theta}_{p+1})^i (1 - (\hat{\pi}_p \times \hat{\theta}_{p+1}))^{N-i}$, where 'hats' denote the relative frequency estimates of the parameters and $n = N \times \hat{\pi}_{p+1}$.

Starting with a small motif length, one can continue to increase its length until the level of significance falls below a preassigned value (say, 0.05).

If the structure of a data set is such that sequential addition of sites leads to the same motif at every stage, compared with the direct procedure of identifying a motif of a certain length, then, as we shall show later, the use of these two procedures of testing statistical significance yield concordant inferences.

Identification and statistical significance of variant motifs

In a standard case-control study, a set of N individuals (cases) possessing a characteristic (e.g., a specific disease) and another set of N individuals (controls), usually matched for age and gender with the cases, not possessing that characteristic, are chosen. DNA sequence data are generated on these $2N$ individuals, and polymorphic sites identified. If the data are diploid, appropriate analyses are carried out (Stephens et al. 2001) to estimate the frequencies of distinct haploid sequences (haplotypes). The objective is to identify a haplotype—polymorphic motif—that occurs at a high frequency among cases, but in low frequency among controls, resulting in a high degree of association of the haplotype with disease. If, indeed, the association is due to causality, then it is expected that there will simultaneously exist a haplotype at a high frequency among controls that comprises alternative nucleotides at the same sites as those found in the high-frequency haplotype among cases. In other words, there will exist a variant motif occurring at a high-frequency among controls compared with that among cases. To identify such motifs among cases and controls, we need to maximize an objective function with respect to three parameters, which may be written in a general form as: $G(S_p) = g(f_1, f_2, m)$, where f_1 and f_2 are, respectively, the frequencies of sequences of nucleotides at the sites in S_p among cases and controls, and m is the number of mismatches between the nucleotide sequences considered for cases and controls. The objective function is so chosen that it is monotonically increasing in f_1, f_2 and m and is to be maximized with respect to these three parameters. The idea is that, since we are searching for a variant motif among controls, we need to find a high-frequency motif among them that simultaneously exhibits a large number of mismatches with a high-frequency motif occurring among cases. Except for this natural modification in the objective function, no change in the search algorithm described earlier was made. An example and details of its implementation are given in Supplemental text 1.

Upon termination of the algorithm, we test whether the odds-ratio estimated from the 2×2 table comprising the frequencies of the two motifs identified among cases and controls (or in the two data matrices under consideration) was significantly different from unity (Breslow and Day 1993).

Following the same spirit as for a single data set discussed and described earlier, one may also assess the statistical significance of the discovered motif in case-control data by using a permutation algorithm to generate a large number of "random" data sets of a structure similar to that of the controls. We have done this. For each case-control data set, synthetic or real, after having identified a motif in the case data by using the variant-motif algorithm, we generated a large number of control data sets by permuting the elements of each column of the control data matrix. We then used the algorithm, and empirically estimated the probability that the odds-ratio obtained for the real data sets of cases and controls is lower than the odds-ratio obtained from the real case data and a randomly generated control data set. We have used this probability as a measure of statistical significance (p-value) of the motif discovered from the real data sets.

In data sets pertaining to evolution, the method of finding a variant motif is simpler because a specific reference sequence is generally given. In this setup, given a string, S_p, of length p, we enumerate from the data all possible sequences $\xi_{l,p}$ $(l = 1,2,\ldots)$ of nucleotides, at the sites included in S_p. For each such sequence $\xi_{l,p}$, we calculate its frequency $f_{l,p}$. We then calculate the number of mismatches, $m_{l,p}$, of each of these sequences $\xi_{l,p}$ $(l = 1,2,\ldots)$ with the reference sequence. The objective function is obviously to be modified as

$$G(S_p) = g(f_2, m).$$

It is evident from the objective function that the value of $m_{l,p}$, for which $G(S_p)$ is maximized is $\leq p$. This indicates that, if the value of $m_{l,p}$ realized at the maximum value of the above objective function is less than p, then there may exist sequences of length p with more than $m_{l,p}$ mismatches with the reference sequence. But the frequency of such a sequence will be much smaller than $f_{l,p}$, resulting in a drop in the value of $G(S_p)$. One effective strategy that we have used in implementing the above objective function is to start the algorithm with a large value of p. This enables us to find a sequence with a considerably high frequency, where $m_{l,p}$ out of the p sites differ from the reference sequence. By keeping track of the sites at which the sequence differs from the reference sequence, we can find the sequence at the sites constituting the

variant motif. Another advantage of using the algorithm is that, even without any prior on the actual length of the motif (discussed in detail in the previous section), the objective function obtains its maximum at some value of $m_{l,p}$, which enables us to get the motif length, the best estimate of which is $m_{l,p}$, from a single run.

To assess the statistical significance of the discovered motif, we generated a large number (10,000) of "random" data sets of a structure similar to the original. If the length of the motif discovered in the original data set was p, we restricted the search algorithm to maximize only over those sequences for which $m_{1,p}$ was equal to p. That is, in the randomly generated data, given a string S_p, the frequency of a sequence was set to 0 if it had less than p mismatches with the reference sequence.

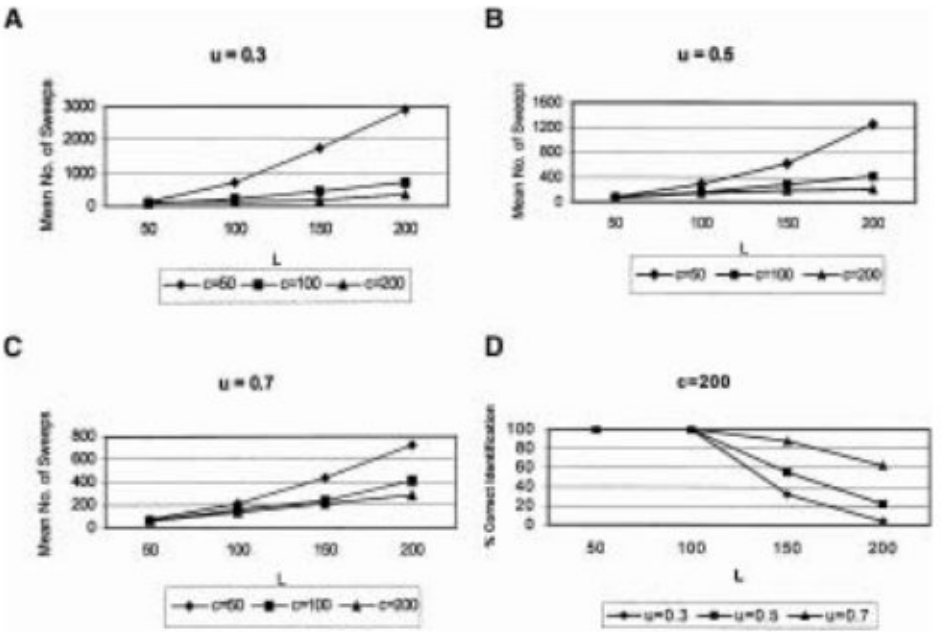


Figure 1. Summary of results of synthetic Data Set 1 with motif length, $p = 10$: (*A–C*) Effect of the control parameter c on time to convergence for three values of u, and (*D*) the effect of increase of the number of polymorphic sites on the probability of correct motif identification (with $c = 200$) for various values of u.

Results

Performance of the algorithms: Assessment using synthetic data sets

Data Set 1

We designed various synthetic data sets, so that the motif in each data set was known, to assess the performance of our algorithm. In our synthetic Data Set 1, a data matrix ($N \times L$) was created, and a known motif of a fixed length (p) was planted in a proportion u of individuals. Data sets were created with different values of relevant parameters; details are given in Supplemental Text 2. The algorithm was applied on each synthetic data matrix, with different values of the control parameter c. As stated earlier, instead of maximizing $G(S)$, we consider an equivalent problem of minimizing a monotonically decreasing function $H(S)$ of $G(S)$. We have taken

$$H(S) = \frac{1}{1+G(S)}$$

in this and in all of the remaining analyses. This choice was subjective and was guided by its simplicity. However, any other monotonically decreasing function, $H(S)$, of $G(S)$ will also obviously work. The results are presented in Figure 1A–C, for $N = 200$, $p = 10$, $L = 50$, 100, 150, and 200, $u = 0.3$, 0.5, and 0.7, and $c = 50$, 100, and 200. (More detailed results for various other values of the parameters are presented in Supplemental Table 2.) For every combination of values of N, L, and p, with an appropriate choice of the control parameter c, our proposed algorithm correctly identified the planted motif in 100% of simulation runs. (Although, for brevity, we have presented results only for $N = 200$ and $p = 10$, results are similar for other values of N and p.) The role of the control parameter c is that it speeds up convergence with larger values, but the convergence may not be to the correct optimum. In our simulation experiments, while for values of $c = 50$, 100, and 150, the planted motif was correctly identified in 100% of simulation runs for any set of values of L and p, the proportion of correct identification was substantially smaller for $c = 200$ (Fig. 1D). For $c = 200$, the mean number of sweeps to convergence was the lowest compared with the other values of c (Fig. 1A–C). Thus, there is a trade-off between speed of convergence and convergence to the correct value. In any application of our algorithm, we recommend that multiple values of c be used, starting with a small value. In other words, we recommend that some experimentation on the convergence behavior of our algorithm with respect to c be done before accepting the results obtained by using a specific value of c.

For every synthetic data set (for different values of N and L) on which the algorithm was used to discover a motif of length p (= the length of the planted motif), we generated 10,000 random data sets of similar structure to test the statistical significance of the discovered motif, as explained earlier. In every case, the estimated probability that a random data set has a motif of frequency higher than that of the discovered motif was $<10^{-7}$. Thus, in every case, the discovered motif was statistically significant at a level $<10^{-7}$.

We have also assessed the levels of significance as the motif length was increased. The significance levels were all <0.005 as the motif length was increased from 2 to 10, but were >0.5 when the motif length was increased from 10 to 11. (Statistical significance was assessed using both the criteria described in an earlier section—assessing the significance of a "drop" in frequency with increase in motif length and also of the addition of a site. Both criteria yielded concordant inferences in every simulation run.) This indicates that our algorithm was not only able to discover the planted motif of length 10, but the discovery was statistically significant. Further, increase of length to 11 was not statistically significant. Detailed results are presented in Supplemental Table 2.

Some general results on the validity and good performance

of the proposed method of assessing statistical significance of a motif discovered by our algorithm are presented in Supplemental text 3.

To examine the limits to which our algorithm can perform well, we constructed new data sets. The descriptions of the data sets and results are given in Supplemental texts 4 and 5.

Data Set 2

We created a synthetic data set analogous to that generated by a case-control study. Two separate data matrices, each of size $N \times L$, corresponding to the cases and controls, were created. We planted, in the case data matrix, a motif of length p in a proportion u_1 of individuals. Under the common-disease, common-variant model (Collins et al. 1998), each of the p sites (SNPs) carries a small relative risk (RR) to the disease, that collectively results in a large haplotype (motif) relative risk. Hence, in the data matrix corresponding to the controls, we changed the proportion of the motif relative to the cases in such a way that the relative risk conferred by the high-risk variant at the i^{th} site of the motif was >0. Details of the methodology for creation of Data Set 2 are given in Supplemental text 6.

In creating synthetic data sets, we have used various values of u_1 and RR. The algorithm for finding variant motifs was used. Statistical significance was assessed by testing the null hypothesis of the odds-ratio being equal to unity, as described earlier.

The values of the parameters used in generating the synthetic data sets were as follows: $N = 100$; $L = 100$, 200, 300; $p = 4$ and 6; $u_1 = 0.2$, 0.4, and RR = 1.2, 1.5, and 2.0. For each combination of L and u_1, 1000 synthetic data sets were generated with each of the various combinations of the other parameters. The algorithm was run on each data set for values of the control parameter $c = 50$, 75, and 100. The results are given in Table 2, for $c = 100$. (For $c = 50$ and 75, the results, not shown, were virtually identical.) In general, our algorithm correctly identified the planted motif in a large proportion of simulation runs only when the RR attributable to a single site was high. The probability of correct identification decreased with decrease in RR. Further, for fixed values of the parameters u_1 and RR, this probability decreased with increase in the motif length, p, but was found to be not strongly dependent on the value of L. Although, for several combinations of simulation parameter values, the probability of correct identification was small or zero, we note that the number of sites and nucleotides that matched between the planted and identified motifs was large, except for RR = 1.2. This indicates that just by chance there may exist motifs with haplotype (motif) relative risks higher than that of the planted motif. However, it is clear that unless the relative risk is small, the true motif will share many sites and nucleotides with the identified motif.

Whether or not the identified motif matched with the planted motif in a synthetic data set, we carried out a test of statistical significance of the identified motif by generating 10,000 random data sets of a similar structure as the control data and estimating the odds-ratios, as explained earlier. The p-values corresponding to the identified motif in the real data, are given in Table 3. None of the identified motifs for the various combinations of the parameter values (motif-length, p; u_1; and the number of polymorphic sites) was statistically significant when RR was small (=1.2). However, when the RR was 1.5 or 2, the identified motifs were all statistically significant at the 5% level.

Data Set 3

This data set was constructed to mimic an evolutionary scenario. When two populations that have diverged from an ancestral population evolve separately, the daughter populations accumulate separate sets of mutations that increase in frequencies because of natural selection or other evolutionary forces. Thus, one may find motifs in the daughter populations, with some motif sites being shared between the two populations, while some being unshared (Schwaiger and Epplen 1995). The shared sites are presumably those sites that belonged to a motif that was present in the ancestral population, while the unshared sites are those that have arisen and increased in frequency since the divergence of the two populations from the ancestral population. We constructed a synthetic data set to mimic this evolutionary scenario (details are provided in Supplemental text 7) and applied our algorithm to assess whether it is possible to discover the relevant motifs. In this data set, the parental population (D_1) carried a motif of length 10, while each of the two daughter populations (D_2 and D_3) carried motifs of length 15, with the 10 parental sites and five additional sites in each motif.

We carried out 1000 independent simulation runs using the procedure described above, with $c = 200$. Detailed results for five runs are provided in Table 3, which show that our probabilistic search algorithm always converged and identified the correct motifs of correct lengths in the parental and in the daughter populations in a small number of sweeps. The final motifs were statistically significant at levels <0.005, as assessed by the procedure in which 10,000 random data sets were generated. As a matter of fact, correct convergence was achieved in every one of the 1000 runs (detailed results not provided) and the convergence using the proposed algorithm was fairly fast (Supplemental Table 6).

Identification of variant motifs: Applications to real data

Gilbert's syndrome: Case–control study

In an ongoing study on Gilbert's syndrome (OMIM #143500), we have generated DNA sequence data of the promoter of UGT1A1 gene among affected individuals and normal controls. The syndrome, characterized by elevated levels of unconjugated serum bilirubin, is caused primarily due to the homozygous insertion of a pair of nucleotides T and A at specific sites in this promoter (Bosma et al. 1995). However, a small fraction of normal individuals also carry these insertions in heterozygous form. In addition to these insertions, in our study, we have found a subset of affected individuals to carry an additional trinucleotide (CAT) insertion at a specific site in the promoter. This insertion has not been found in any of the unaffected control individuals. Haplotypes and their frequencies were estimated from the sequence data, separately for the cases and controls. The size (80×456) of the data matrix was the same for both cases and controls. We have used the algorithm for finding variant motifs (with $c = 100$) and were able to identify the 5-site motif (corresponding to the TA dinucleotide and the CAT trinucleotide insertions) correctly. The 5-site motif was estimated to be present in 11.25% of the cases and 0% of controls, which agreed with the actual count. Since none of the controls possessed this motif, the relative risk cannot be computed, but the finding is obviously significant

Table 2. Percentage of simulation runs indicating matches between planted and identified motifs pertaining to case-control data set 2, and the significance levels of the identified motifs

			RR																						
			2.0								1.5								1.2							
Planted Motif length	u_1	No. of polymorphic sites	Number of matches[a]								Number of matches								Number of matches							
			0	1	2	3	4	5	6	p-value	0	1	2	3	4	5	6	p-value	0	1	2	3	4	5	6	p-value
4	0.2	100	0	0	0	0	100	—	—	$<10^{-7}$	0	0	0	55.2	44.8	—	—	0.007	43.5	33.9	12.2	10.4	0	—	—	0.522
		200	0	0	0	0	100	—	—	$<10^{-7}$	0	0	0	57.4	42.6	—	—	0.008	46.8	31.5	12.7	9.0	0	—	—	0.481
		300	0	0	0	0	100	—	—	$<10^{-7}$	0	0	0	60.3	39.7	—	—	0.011	49.2	35.8	10.1	4.9	0	—	—	0.497
	0.4	100	0	0	0	0	100	—	—	$<10^{-7}$	0	0	0	56.1	43.9	—	—	0.011	25.3	31.6	28.1	14.8	0.2	—	—	0.927
		200	0	0	0	0	100	—	—	$<10^{-7}$	0	0	0	56.8	43.2	—	—	0.011	28.1	32.2	25.7	13.9	0.1	—	—	0.931
		300	0	0	0	0	100	—	—	$<10^{-7}$	0	0	0	59.1	40.9	—	—	0.013	28.8	33.6	24.1	13.4	0.1	—	—	0.919
6	0.2	100	0	0	0	0	0	0	100	$<10^{-7}$	0	8.2	13.1	23.7	35.3	18.9	0.8	0.010	19.8	37.2	21.9	18.7	2.4	0	0	0.498
		200	0	0	0	0	0	0	100	$<10^{-7}$	0	9.5	16.8	22.6	32.2	18.6	0.3	0.011	22.7	34.0	20.8	18.2	4.1	0	0	0.489
		300	0	0	0	0	0	0	100	$<10^{-7}$	0	11.2	12.1	23.9	31.9	20.6	0.3	0.013	22.8	35.3	21.6	18.6	1.7	0	0	0.490
	0.4	100	0	0	0	0	13.3	56.8	29.9	$<10^{-7}$	0	3.2	8.1	18.6	51.2	18.5	0.4	0.017	35.2	28.3	27.6	8.3	0.6	0	0	0.917
		200	0	0	0	0	14.9	60.3	24.8	$<10^{-7}$	0	3.2	8.4	16.3	52.7	18.9	0.5	0.018	36.5	25.9	28.5	8.6	0.5	0	0	0.901
		300	0	0	0	0	14.7	61.1	24.2	$<10^{-7}$	0	3.7	8.9	16.1	53.0	17.9	0.4	0.021	37.8	22.7	31.9	7.3	0.3	0	0	0.886

[a]Number of matches indicate the number of sites and the nucleotides at the sites that match between the motif identified by the algorithm and the planted motif.

Table 3. Detailed results pertaining to synthetic data set 3 for five independent simulation runs

Characteristics of Synthetic Data Matrices					
Population	Motif length	Simulation number	Sites in motif[a]	Frequencies of "1" at motif sites	Number of sweeps to convergence	Whether converged to "correct" motif
Population 1 (D_1)	10	1	17 22 24 27 28 36 39 40 43 50	855 845 847 837 855 847 843 847 827 841	38	YES
		2	1 4 8 15 19 20 22 37 39 44	855 832 848 867 861 874 850 858 851 849	27	YES
		3	6 7 8 9 11 16 20 21 35 47	876 846 837 852 830 855 843 851 859 849	64	YES
		4	3 11 13 27 35 39 42 46 47 50	846 868 851 834 874 874 837 848 832 856	35	YES
		5	5 7 10 16 22 29 30 33 42 44	830 846 827 840 851 860 853 851 858 826	84	YES
Population 2 (D_2)	15	1	17 22 24 27 28 36 39 40 43 50 *18 31 34 42 48*	863 845 867 854 875 846 851 858 844 832 845 827 816 888 855	13	YES
		2	1 4 8 15 19 20 22 37 39 44 *7 12 21 31 46*	857 821 852 882 863 887 838 866 858 853 828 890 886 830 850	37	YES
		3	6 7 8 9 11 16 20 21 35 47 *2 27 36 37 41*	873 835 839 859 828 853 863 851 876 848 845 846 890 824 816	77	YES
		4	3 11 13 27 35 39 42 46 47 50 *5 15 26 37 38*	844 874 848 831 867 860 819 862 834 856 827 809 886 826 856	11	YES
		5	5 7 10 16 22 29 30 33 42 44 *1 8 27 35 46*	843 840 821 848 854 880 838 836 847 820 821 845 876 860 881	28	YES
Population 3 (D_3)	15	1	17 22 24 27 28 36 39 40 43 50 *3 9 15 33 44*	870 843 847 841 824 857 859 864 822 845 882 828 811 892 852	42	YES
		2	1 4 8 15 19 20 22 37 39 44 *9 18 26 29 49*	865 812 831 868 863 877 833 855 836 824 833 837 890 886 854	23	YES
		3	6 7 8 9 11 16 20 21 35 47 *12 25 30 39 42*	860 833 845 873 821 839 852 872 857 846 886 826 876 854 843	58	YES
		4	3 11 13 27 35 39 42 46 47 50 *10 20 21 32 50*	846 861 826 838 882 867 834 847 837 872 829 889 883 826 840	61	YES
		5	5 7 10 16 22 29 30 33 42 44 *6 13 20 37 47*	844 842 826 849 852 840 851 834 879 825 811 827 885 864 867	12	YES

[a]The sites indicated in italics are the five new sites that are specific to the daughter population, (D_2 and D_3) in each simulation run, in addition to the 10 sites of the ancestral population (D_1).

(11.25% among cases, vs. 0% among controls). The exact p-value computed from the binomial distribution (using the estimate of the probability of the motif from the pooled data of cases and controls) is 9.74×10^{-3}.

LDL receptor haplotypes among individuals of European and African descent: The PARC study

In an ongoing project entitled "Pharmacogenomics and Risk of Cardiovascular Disease" (PARC) at the University of Washington, Seattle, data on haplotypes of individuals belonging to African descent ($n = 48$) and European descent ($n = 46$), pertaining to the LDL receptor gene (located on human chromosome 19p13.3), have been made available in the public domain (http://droog.gs.washington.edu/parc/data/ldlr/welcome.htm). The number of polymorphic sites (L) in this data set is 117. We have used our algorithm to find whether there are any high-frequency contrasting (variant) motifs present among individuals of European and African descent. We have used our algorithm for finding variant motifs (with $c = 100$) and discovered that the motif TTTGGTAGC of length 9 occurs at the nucleotide sites 26, 34, 41, 43, 50, 54, 57, 58, and 61 with a frequency of 19 (39.5%) among individuals of African descent, and a completely contrasting motif CCGACCCAT occurs at these sites with a frequency of 34 (73.9%) among individuals of European descent. The degree of association in the corresponding 2×2 table is statistically significant at a level <0.001. Both of these motifs were statistically significant with estimated p-values $< 10^{-7}$. (To test the significance of the discovered motif among Europeans, we generated 10,000 random data sets of a structure similar to that of African-descent individuals, and vice-versa for testing the significance of the discovered motif among individuals of African descent.)

Mitochondrial DNA haplogroups M and U

Extensive sequence data on the hypervariable segment 1 (HVS1) of the mitochondrial DNA (mtDNA) have been generated (Handt et al. 1998) and analyzed (Macaulay et al. 1999) in various global populations (http://www.hvrbase.org). Based on the presence or absence of specific restriction sites outside of the HVS1 in the mtDNA, two of the major haplogroups (HGs) identified are M and U (Wallace 1995). Within these haplogroups, specific motifs have been found in the HVS1, some of which are in contrast to those found in the CRS (Anderson et al. 1981). These motifs have been used to define haplogroups within the HGs (Kivisild et al. 1999). We have used data of 528 individuals from various ethnic populations of India (Basu et al. 2003). The total number of polymorphic sites in this data set was 153, and the numbers of individuals belonging to HGs M and U were, respectively, 338 and 115. We applied the motif-finding algorithm (with $c = 100$) separately on the HVS1 sequence data of HGs M and U. An objective function, $G(S_p)$, that gives considerable weightage to the number of mismatches was used, that is,

$$G(S_p) = \max_I (f_{I,p}^{m_I,p}).$$

For HG-M, $G(S_p)$ attained a maximum value with $m_I = 4$, for all values of $p \geq 4$. For $m_I = 4$, the sites at which nucleotides differed from the CRS were S = (16223, 16270, 16319, 16352). The frequency of this string, f_p, was 21 (= 6.21% of the total number of samples), and the nucleotides at the relevant positions were T, T, A, and C, respectively. The next most frequent string was (16223[T], 16274[A], 16319[A], and 16320[C]) with a frequency of 17 (5.03%). These two motifs belong to known subhaplogroups M* (defined by C→T transition at the site 16223) and M2 (defined by C→T transition at the site 16223 and a G→T transition at the site 16319), which are prevalent in Indian populations (Bamshad et al. 2001).

For HG-U also, the objective function, coincidentally, attained a maximum at $m_I = 4$, and the motif identified was (16051[G], 16206[C], 16230[G], 16311[C]), with a frequency of 18 (=15.65% of the total number of samples). The vast majority of HG-U individuals in India belong to HG-U2i and U7. The U2i is the Indian-specific subcluster of U, as opposed to the Western-Eurasian subcluster U2e (Kivisild et al. 1999). Interestingly, the motif GCGC at nps 16051, 16206, 16230, and 16311, respectively, has been found on the U2i background, which is present in 18 of the 115 individuals. This motif is found almost exclusively among tribal, middle- and lower-caste populations, but not among the upper-caste populations or the Muslims (of Uttar Pradesh). This motif is also present in many of the Pakistani samples screened by Kivisild et al. (1999). Our bootstrap procedure for testing statistical significance of a motif indicated that in all of the above cases, the identified motifs were statistically significant at level <0.05.

These examples demonstrate that the proposed algorithm was able to identify previously discovered motifs, and therefore, can be profitably used in evolutionary studies to identify new motifs. The anthropological implications of our findings on HGs M and U presented above have already been described in Basu et al. (2003).

The !Kungs of Botswana, Africa

We have also applied our algorithm (with $c = 100$) on mtDNA HVS-1 data sets (Handt et al. 1998) of various African populations. The algorithm identified a motif of length 5 in the !Kung population of Botswana, which contrasts with the CRS. The motif is constituted by the sites 16129, 16223, 16230, 16294, and 16311. The nucleotides in the respective sites in the CRS are G, C, A, C, and T respectively, while in the !Kung population, the motif is ATGTC. The motif is present in 17 (68.0%) of 25 !Kung sequences. Using the procedure suggested earlier, with 10,000 replications, no variant motif of length 5 with a frequency higher than that of the identified motif was found, indicating that the statistical significance of the identified motif is very high. The uniqueness of the motif is not only characterized by its difference from that present in the CRS, but also because this motif is not present in any other African population (Table 4), and has probably risen to the present high frequency among the !Kungs by genetic drift.

Discussion

The problem of identifying motifs in genetic data arises commonly in human genetical research. Such data include DNA sequence data, haplotype data, and genotype data. Motif identification is necessary to draw inferences on evolutionary histories of populations or lineages, to examine associations in case-control studies, etc. More recently, with the initiation of the HapMap project (Couzin 2002), the problem of finding motifs within

Table 4. Motifs in mtDNA HVS1 discovered in various African populations using the proposed algorithm

Population	Sample size	Frequency of the most frequent sequence (%)	Motif[a]
Egypt (Assiut)	23	8 (35.8)	G–C–A–C–T
Egypt (Cairo)	10	3 (30.0)	G–C–A–C–T
Egypt (Manasoura)	46	13 (28.3)	G–C–A–C–T
Sudan (Kerma)	42	12 (28.6)	G–C–A–C–T
Sudan	86	23 (26.7)	G–**T**–A–C–**C**
Ethiopia	10	3 (30.0)	G–**T**–A–C–T G–**T**–A–C–**C**
Somali	27	7 (25.9)	G–**T**–A–**T**–T
Somali (Kenya)	15	4 (26.7)	G–**T**–A–**T**–T
Turkana (Kenya)	37	9 (24.3)	G–**T**–A–C–**C**
Kikuyu (Kenya)	25	8 (32.0)	G–**T**–A–C–**C**
Tanzania	17	10 (58.8)	G–**T**–A–**T**–**C**
!Kung (Botswana)	25	17 (68.0)	**A**–**T**–**G**–**T**–**C**

[a]Nucleotides denoted in boldface are different from the nucleotides in the CRS at the corresponding sites.

haplotype blocks, which probably occur because of variation in recombination rates across the human genome, arise naturally. In most of these applications, it is pertinent to identify motifs of nucleotides at a set of polymorphic sites, which may not be contiguous. For example, in research on complex diseases, often data are generated on multiple unlinked genes, and if, indeed, genotypes or haplotypes at a subset of these genes determine the susceptibility to the disease, then motifs will exist at a set of noncontiguous polymorphic sites. A search based on complete enumeration for such motifs can be computationally extremely time consuming and inefficient—it might not even be feasible in practice for large data sets. To the best of our knowledge, no computationally efficient algorithms exist for finding motifs at noncontiguous polymorphic sites. We have, therefore, devised a set of computationally fast and efficient algorithms based on probabilistic methods. We have first devised a search algorithm when the length of the motif is specified a priori, and have then extended it to take into account the possibility of the motif length not being known a priori. The specific functions (e.g., β_t, $H(S)$) used by us were chosen not only to satisfy the criteria required for convergence of this class of probabilistic search algorithms (Winkler and Lutz 2003), but also because of their simplicity and intuitive appeal. Our algorithms are not tied to these specific choices of functions; users may try other functions satisfying the general conditions required for convergence. For a given motif length, we have proposed a statistical criterion of assessing the significance of the motif discovery using a bootstrap procedure. When the motif length is not specified, we have devised a statistical criterion for determining the motif length from the data simultaneously with the search for a motif. We have proposed an alternative criterion of assessing statistical significance when the motif length is extended by sequential addition of sites and nucleotides. Finally, we have proposed methods for assessment of statistical significance of a discovered motif in a real data set, in relation to a random data set of similar structure. Using various synthetic data sets to mimic real-life applications, we have demonstrated that the proposed methods work well. We have also applied these methods to several real data sets—pertaining to case-control data on complex phenotypes and evolutionary data—and obtained many useful inferences.

Through our simulations, we have discovered some of limitations of our algorithm as well. In particular, when we assessed (Supplemental text 4) whether our algorithm converges correctly in a search space that contains exactly one global maximum, and also a large number of local maxima with values not very different from the global maximum, our algorithm failed to converge to the global maximum. This limitation is, of course, inherent to all numerical search procedures that do not use complete enumeration. Further, in simulated case-control data, our algorithm failed to identify the correct motif, especially when the relative risk attributable to a site included in the motif was small (Table 2). For a small relative risk, the identified motif was also statistically nonsignificant (Table 2). However, in most simulation runs, the identified motif shared several sites in common with the planted motif. The reason for nonconvergence to the correct motif was due to the fact that in realistic case-control data sets, there may be multiple motifs with high haplotype (motif) relative risks just by chance, especially when individual sites (SNPs) do not confer a large relative risk to the disease. This finding is consistent with published observations (e.g., Cardon and Bell 2001) that significant findings of haplotype associations from case-control studies are often not replicable. Our simulation results also underscore the need for replication of findings of case-control association studies.

We would finally like to emphasize that the convergence properties of the proposed algorithms are critically dependent on the control parameter, c. While from the user's point of view it is desirable to be able to prescribe some universal and objective guidelines for the choice of c, this is not possible. In specific applications like those presented here, one can identify a range of values of c that makes the algorithm computationally feasible, with a high probability of convergence to the true optimum. In practice, this range of c needs to be identified by trial and error. We first note that the speed of convergence is directly proportional to the value of c. Further, the probability of convergence to the true optimum for a specific choice of c is more dependent on the value of L than on N. Using these two facts, the user should make a judicious choice of c, but try with multiple values. We strongly recommend that some experimentation on the convergence behavior of the algorithm with respect to c in multiparameter settings be done to make a judicious choice of c. We have found that with N in the range of from 200 to 500 and L in the range of from 200 to 500, any value of c in the range of from 50 to 100 works very well.

Although we have formulated our algorithms keeping haplotype or haploid DNA sequence data in mind, there is no inherent limitation to use these methods on genotype data. Genotype data need only be recoded in order to apply these algorithms. For example, at a biallelic locus, with alleles A and a, the genotypes AA, Aa, and aa may be recoded as 1, 2, and 3. We finally note that there are other classes of probabilistic search algorithms—such as genetic algorithm (Goldberg 1989), Gibbsean annealing (Winkler and Lutz 2003), and evolutionary Monte Carlo (Liang and Wong 2001)—that may also be applicable to the problem considered in this study. We have not explored these classes of algorithms in any detail, and therefore, make no claim that the algorithms proposed by us will outperform other probabilistic search algorithms.

We have developed a computer program, MOTIFIND, implementing these algorithms. This program is written in C, and can be obtained by writing to the authors. This program can handle both haploid and diploid genotype data.

Acknowledgments

This work was partially supported by grants from the Department of Biotechnology and Council for Scientific and Industrial Research, Government of India. We thank Dr. A. Chowdhury for allowing us to include the unpublished data on Gilbert's syndrome. We also thank two anonymous reviewers for comments that have helped to substantially improve an earlier version of this work.

References

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., et al. 1981. Sequence and organization of the human mitochondrial genome. *Nature* **290:** 457–465.

Balasubramanian, S., Harrison, P., Hegyi, H., Bertone, P., Luscombe, N., Echoles, N., McGarvey, P., Zhang, Z.L., and Gerstein, M. 2002. SNPs on chromosomes 21 and 22 analysis in terms of protein features and pseudogenes. *Pharmacogenomics* **3:** 1–10.

Bamshad, M.J., Kivisild, T., Watkins, W.S., Dixon, M.P., Ricker, L.E., Rao, B.B., Naidu, M., Prasad, B.V.R., Reddy, P.G., Rasanayagam, A., et al. 2001. Genetic evidence on the origins of Indian caste populations. *Genome Res.* **11:** 994–1004.

Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., Dey, B., Roy, M., Roy, B., Bhattacharyya, N.P., et al. 2003. Ethnic India: A genomic view, with special reference to peopling and structure. *Genome Res.* **13:** 2277–2290.

Bosma, P.J., Chowdhury, J.R., Bakker, C., Gantla, S., deBoer, A., Oostra, B.A., Lindhout, D., Tytgat, G.N., Jansen, P.L., Oude Elferink, R.P., et al. 1995. The genetic basis of the reduced expression of UDP-glucuronosyltransferase 1 in Gilbert's syndrome. *New Engl. J. Med.* **333:** 1171–1175.

Breslow, N.E. and Day, N.E. 1993. *Statistical methods in cancer research: The analysis of case-control studies*. International Agency for Research on Cancer, Lyon.

Cardon, L.R. and Bell, J. 2001. Association study designs for complex diseases. *Nat. Genet.* **2:** 91–99.

Collins, F.S., Brooks, L.D., and Chakravarti, A. 1998. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8:** 1229–1231.

Couzin, J. 2002. Human genome. HapMap launched with pledges of $100 million. *Science* **298:** 941–942.

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

Goldberg, D.E. 1989. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley Publishing Co., Boston, MA.

Gupta, M. and Liu, J.S. 2003. Discovery of conserved sequence patterns using a stochastic dictionary model. *J. Amer. Stat. Assoc.* **98:** 55–66.

Handt, O., Meyer, S., and Haeseler, A., 1998. Compilation of human mtDNA control region sequences. *Nucleic Acids Res.* **26:** 126–129.

Keiler, K.C. and Shapiro, L. 2001. Conserved promoter motif is required for cell cycle timing of dnaX transcription in Caulobacter. *J. Bacteriol.* **183:** 4860–4865.

Khani-Hanjani, A., Lacaille, D., Horne, C., Chalmers, A., Hoar, D.I., Balshaw, R., and Keown, P.A. 2002. Expression of QK/QR/RRRAA or DERAA motifs at the third hypervariable region of HLA-DRB1 and disease severity in rheumatoid arthritis. *J. Rheumatol.* **29:** 1358–1365.

Kivisild, T., Bamshad, M.J., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., Laos, S., Parik, J., Watkins, W.S., Dixon, M.E., et al. 1999. Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Curr. Biol.* **9:** 1331–1334.

Liang, F. and Wong, W. 2001. Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *J. Amer. Stat. Assoc.* **96:** 653–666.

Liu, J.S. 2001. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics, Springer, Heidelberg, Germany.

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonne-Tamir, B., Sykes, B., and Torroni, A. 1999. The emerging tree of West Eurasian mtDNAs: A synthesis of control-region sequences and RFLPs. *Am. J. Hum. Genet.* **64:** 232–249.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21:** 1087–1091.

Quintana-Murci, L., Semino, O., Bandelt, H.J., Passarino, G., McElreavey, K., and Santachiara-Benerecetti, A.S. 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat. Genet.* **23:** 437–441.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419:** 832–837.

Schwaiger, F.W. and Epplen, J.T. 1995. Exonic MHC-DRB polymorphisms and intronic simple repeat sequences: Janus' faces of DNA sequence evolution. *Immunol. Rev.* **143:** 199–224.

Stephens, M., Smith, N.J., and Donnelly, P. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **69:** 906–914.

Tateno, Y., Ikeo, K., Imanishi, T., Watanabe, H., Endo, T., Yamaguchi, Y., Suzuki, Y., Takahashi, K., Tsunoyama, K., Kawai, M., et al. 1997. Evolutionary motif and its biological and structural significance. *J. Mol. Evol.* **44:** S38–S43.

Wallace, D.C. 1995. Mitochondrial DNA variation in human evolution, degenerative disease and aging. *Am. J. Hum. Genet.* **57:** 201–223.

Winkler, G. and Lutz, G.F.H. 2003. *Image analysis, random fields and Markov chain Monte Carlo methods: A mathematical introduction*. Applications of Mathematics Series. Springer, Heidelberg, Germany.

Web site references

http://www.hvrbase.org/; The URL of the mtDNA database.

http://droog.gs.washington.edu/parc/data/ldlr/welcome.htm; URL of the LDL receptor.

Received January 15, 2004; accepted in revised form October 21, 2004.

SUPPLEMENTAL RESEARCH DATA

Supplementary Text – 1

Details of Implementation of the Algorithm for Finding "Variant" Motifs, with Special Reference to Case-Control Data

Consider data matrices $((a_{ij}))_{N\times L}$ and $((b_{ij}))_{N\times L}$, where a_{ij} denotes a nucleotide (A,T,G or C) at the j^{th} polymorphic site ($j = 1, 2, \ldots, L$) for the i^{th} individual ($i = 1, 2, \ldots, N$) among the cases and b_{ij} denotes the same for the i^{th} individual ($i = 1, 2, \ldots, N$) among normal controls. These data matrices are generated from aligned DNA sequences of a specific homologous genomic segment of $2N$ individuals (N cases and N controls), from which all monomorphic sites have been removed. For simplicity we have considered the sample size (N) to be same for cases and controls, but these can be different in practice. For genotype data, we can numerically recode genotypes, e.g., AA as 0, AG as 1 and GG as 2 (assuming that there are two variant nucleotides A and G at the locus under consideration) or we can estimate haplotypes and carry out analyses where a_{ij} denotes the nucleotide (A,T,G or C) at the j^{th} polymorphic site ($j = 1, 2, \ldots, L$) for the i^{th} haplotype ($i = 1, 2, \ldots, N$) among the cases or among controls. We also note that if disjoint segments of DNA are to be simultaneously examined for motif finding, then appropriate segments may be separately aligned and the aligned segments concatenated in the data matrix.

As before, let $V = \{1, 2, \ldots, L\}$ denote the set of all L polymorphic sites in the data. Let Π_p denote the set of all possible combinations of p sites in V. In general, $\Pi_p = \{V_p^k\}$, $k = 1, 2, \ldots, \binom{L}{p}$ and $V_p^k = \{x_1^k, x_2^k, \ldots, x_p^k : x_i^k \in V\}$. For a fixed k, we define the *modal sequence among cases* on V_p^k as that particular combination of nucleotides at the sites $\{x_1^k, x_2^k, \ldots, x_p^k\}$ included in V_p^k, $k = 1, 2, \ldots, \binom{L}{p}$, which has the highest frequency (f_{case}) among the cases. In the data matrix of Supplementary Table 1, the modal sequence, for example, on $V_2^1 = \{1, 2\}$ is AG with frequency 2, on $V_2^2 = \{1, 3\}$ is AT with frequency 3, etc.

Since our goal is to simultaneously look for a sequence that occurs with a high frequency among controls and is variant to that occurring at a large frequency among

the cases, we evaluate the following:

Given a specific string, S_p, of length p, we first find the *modal sequence among cases* (ξ_{case}) and its frequency f_{case}. Let us denote that sequence by ξ_{case}. Then we enumerate from the control data all possible sequences $\xi_{l,p}$ $(l = 1, 2, \ldots)$ of nucleotides, at the sites included in S_p. For each such sequence $\xi_{l,p}$, we calculate its frequency $f_{l,p}$. We then calculate the number of mismatches, $m_{l,p}$, of each of these sequences $\xi_{l,p}$ $(l = 1, 2, \ldots)$ with the reference sequence. Since we are interested in identifying a motif that is different from the reference sequence (in this case ξ_{case}), we need to take these $m_{l,p}$ values into account. We provide a greater weightage to a sequence that has a larger number of mismatches with the reference sequence. In this problem, therefore, we have used a modified objective function of the form $G(S_p) = g(f_1, f_2, m)$, where f_1, f_2 and m denote, respectively, the frequencies of the string of nucleotides at the sites in S_p among cases and controls, respectively, and m denotes the number of mismatches between these two nucleotide sequences among cases and controls. The specifics of the use of such an objective function are explained below with the help of an example.

In the data matrix of Supplementary Table 1, the modal sequence among cases (ξ_{case}), on $V_2^2 = \{1, 3\}$ is AT with frequency 3. For this set of sites $V_2^2 = \{1, 3\}$ there are 2 distinct sequences, AT and GC, in the control data set with frequencies 3 and 1, respectively. The number of mismatches of these two sequences with ξ_{case}, on $V_2^2 = \{1, 3\}$ are, respectively, 0 and 2. Hence the value of the objective function for this choice of two candidate sites is: maximum of $((0.75+1)(0.75+1)(0)$ and $(0.75+1)(0.25+1)(2))$. Therefore, although the sequence AT among controls occur at a high frequency, it is not chosen as a candidate sequence because of its smaller number of mismatches with ξ_{case}. AT among cases and GC among the controls are the preferred candidates for variant motifs if these two sites are chosen. We then use the Metropolis-Hastings algorithm to choose the set of sites which globally maximises the objective function.

Supplementary Table 1 An Example of a Case-Control Data Matrix

	Sequence/ Individual No.	Variant Site No. 1	2	3	4	5	6	7
Cases	1	A	A	T	T	G	C	C
	2	A	G	T	C	G	C	T
	3	A	G	T	T	A	C	T
	4	G	G	C	C	A	T	T
Controls	1	A	G	T	T	G	C	C
	2	A	T	T	C	G	C	T
	3	A	T	T	T	A	C	T
	4	G	G	C	T	A	T	T

Supplementary Text 2

Details on the Method of Generating Synthetic Data Set 1

Our synthetic Data Set 1, comprises a data-matrix of size $N \times L$, corresponding to data on N individuals at L binary polymorphic sites. At each site, for each individual, we assigned a binary digit (0 or 1) with probability 0.5. (However, as noted in the Introduction, the assumption of each polymorphic site being binary is not crucial to this algorithm.) A motif of length p was planted in a fraction u of the N individuals. To do this, we selected p sites randomly from the L sites; that is, we chose p columns of the $N \times L$ data-matrix randomly. Then, $[N \times u]$ rows were randomly chosen (where $[x]$ denotes the largest integer contained in x), and the elements of each of the p chosen columns corresponding to each of these chosen rows were replaced by 1. For a given set of values of (N, L, and p), 1000 independent synthetic data matrices were thus generated.

Supplementary Table 2. Performance of the algorithm on Synthetic Data Set 1 with N=200 for different values of the variables L (number of segregating sites) and u (proportion of the planted motif among N), and for different values of the control parameter c. (The mean and s.d. of the number of sweeps to convergence and the % of simulation runs in which the planted motif was correctly identified are based on 1000 independent simulation runs for each combination of values of the variables and the parameter. The minimum and maximum values of the significance level as the motif length was increased from 9 to 10 and from 10 to 11 were also calculated over 1000 simulation runs.)

L	u	c=50					c=100					c=200				
		No. of Sweeps		% correct	Sig. Level* (min,max)		No. of Sweeps		% correct	Sig. Level* (min,max)		No. of Sweeps		% correct	Sig. Level* (min,max)	
		Mean	s.d.		p: 9→10	p: 10→11	Mean	s.d.		p: 9→10	p: 10→11	Mean	s.d.		p: 9→10	p: 10→11
50	0.3	130.13	57.34	100	.024, .027	529, 532	72.52	48.71	100	.024, .027	529, 532	66.26	49.29	100	.024, .027	529, 532
	0.4	95.74	43.44	100	.15, .17	527, 530	72.42	35.69	100	.15, .17	527, 530	63.92	41.62	100	.15, .17	527, 530
	0.5	81.41	34.88	100	.20, .22	526, 528	70.68	39.15	100	.20, .22	526, 528	63.08	39.06	100	.20, .22	526, 528
	0.6	77.10	47.42	100	.41, .45	526, 527	70.88	42.77	100	.41, .45	526, 527	62.73	43.35	100	.41, .45	526, 527
	0.7	71.85	35.55	100	1.1, 1.5	527, 527	62.72	39.23	100	1.1, 1.5	527, 527	62.08	43.78	100	1.1, 1.5	527, 527
100	0.3	728.88	352.03	100	.024, .027	529, 532	216.63	99.06	100	.024, .027	529, 532	127.69	100.62	53	.024, .027	529, 532
	0.4	363.97	149.04	100	.16, .17	527, 530	200.23	123.80	100	.16, .17	527, 530	143.26	87.27	83	.16, .17	527, 530
	0.5	288.40	124.22	100	.20, .22	526, 528	157.20	96.93	100	.20, .22	526, 528	143.27	91.69	95	.20, .22	526, 528
	0.6	246.02	120.07	100	.41, .45	526, 527	161.04	93.10	100	.41, .45	526, 527	141.54	88.79	98	.41, .45	526, 527
	0.7	213.92	104.83	100	1.1, 1.5	527, 527	161.30	90.30	100	1.1, 1.5	527, 527	139.05	90.39	100	1.1, 1.5	527, 527
150	0.3	190.86	97.07	100	.024, .027	529, 532	435.49	202.93	100	.024, .027	529, 532	1736.32	900.54	32	.024, .027	529, 532
	0.4	149.52	67.88	100	.15, .17	527, 530	331.85	173.77	100	.15, .17	527, 530	954.16	497.42	40	.15, .17	527, 530
	0.5	189.40	117.54	100	.20, .22	526, 528	287.73	133.55	100	.20, .22	526, 528	620.00	250.47	55	.20, .22	526, 528
	0.6	228.16	130.53	100	.41, .45	526, 527	259.56	163.36	100	.41, .45	526, 527	461.06	190.07	74	.41, .45	526, 527
	0.7	217.61	138.81	100	1.1, 1.5	527, 527	239.62	130.98	100	1.1, 1.5	527, 527	436.14	227.85	88	1.1, 1.5	527, 527
200	0.3	369.50	200.10	100	.024, .027	529, 532	720.90	294.71	100	.024, .027	529, 532	2932.18	1218.12	4	.024, .027	529, 532
	0.4	301.17	140.12	100	.15, .17	527, 530	610.35	264.93	100	.15, .17	527, 530	2028.86	986.43	14	.15, .17	527, 530
	0.5	206.58	135.76	100	.20, .22	526, 528	416.3	231.80	100	.20, .22	526, 528	1252.38	474.05	22	.20, .22	526, 528
	0.6	224.83	126.04	100	.41, .45	526, 527	404.4	194.50	100	.41, .45	526, 527	817.95	203.28	46	.41, .45	526, 527
	0.7	283.26	183.31	100	1.1, 1.5	527, 527	415.52	242.10	100	1.1, 1.5	527, 527	721.78	263.08	62	1.1, 1.5	527, 527

* All values are multiplied by 10^{-3}

Supplementary Text 3

Validation and Performance of the Proposed Method of Assessment of Statistical Significance of Motif Discovery

Before proceeding further, we provide some general results pertaining to the proposed method of assessing the statistical significance of a motif discovered by our algorithm. To estimate the statistical significance of a motif discovered by our algorithm in relation to a random data set of "similar" structure, we created a data set with N=20, L=10 and planted a motif of length p=5 (=1,1,1,1,1) at 5 randomly-chosen sites from among the 10 sites with a frequency u (among N). Three values of u were used; these were 0.3, 0.5 and 0.7. The remaining cells in the $N \times L$ data matrix were filled with 1 or 0, each with probability 0.5. We shall refer to this data set as the "real" data set. We then used our algorithm to discover a motif in this "real" data set. Next, we created 10,000 $N \times L$ replicate data matrices in which the 1's and the 0's in cells in column i were randomly permuted so that the total number of 1's and 0's occurring in the column remain same as that in the "real" data. In each of the 10,000 replicate random data matrices, we searched for the "best" motif of length 5 *by complete enumeration*. The large-deviation probability, that is, the probability that the best or the discovered motif occurs in a random data set with a frequency that is greater than or equal to that of the motif discovered by our algorithm in the "real" data set, is ≤ 0.0001, for all values of u (Supplementary Table 3). These findings further indicate that our algorithm performs well.

Since for a large data set, it is not possible to search for the "best" motif by complete enumeration, we additionally sought to evaluate our algorithm by the above statistical-significance criteria using an approximate method. In this approximate method, the only change that was made is that instead of searching for the "best" motif by complete enumeration in a random data set, we applied our algorithm to discover the "best" motif. The "real" data sets were generated in the same way as the Synthetic Data Set 1, with N=200, L=50 and 200, and p=10. Three values of u were used 0.3, 0.5 and

0.7. Motif search was performed using c=100 in both the "real" and the random data sets. In all cases, the estimated probability of existence of a motif in a random data set with a higher frequency than the motif discovered in the real data set was $< 10^{-7}$.

Supplementary Table 3. Probabilities that the best motif discovered in a random data set has a frequency (f^*) greater than the frequency (f) of the motif discovered in the real data set of size 20 × 10, for different values of u[1,2]

u	Prob ($f^* \geq f$)
0.3	< 0.00001
0.5	0.0001
0.7	<0.00001

[1] Results are based on 1000 real data sets for each value of u and 10000 random data sets for each real data set.

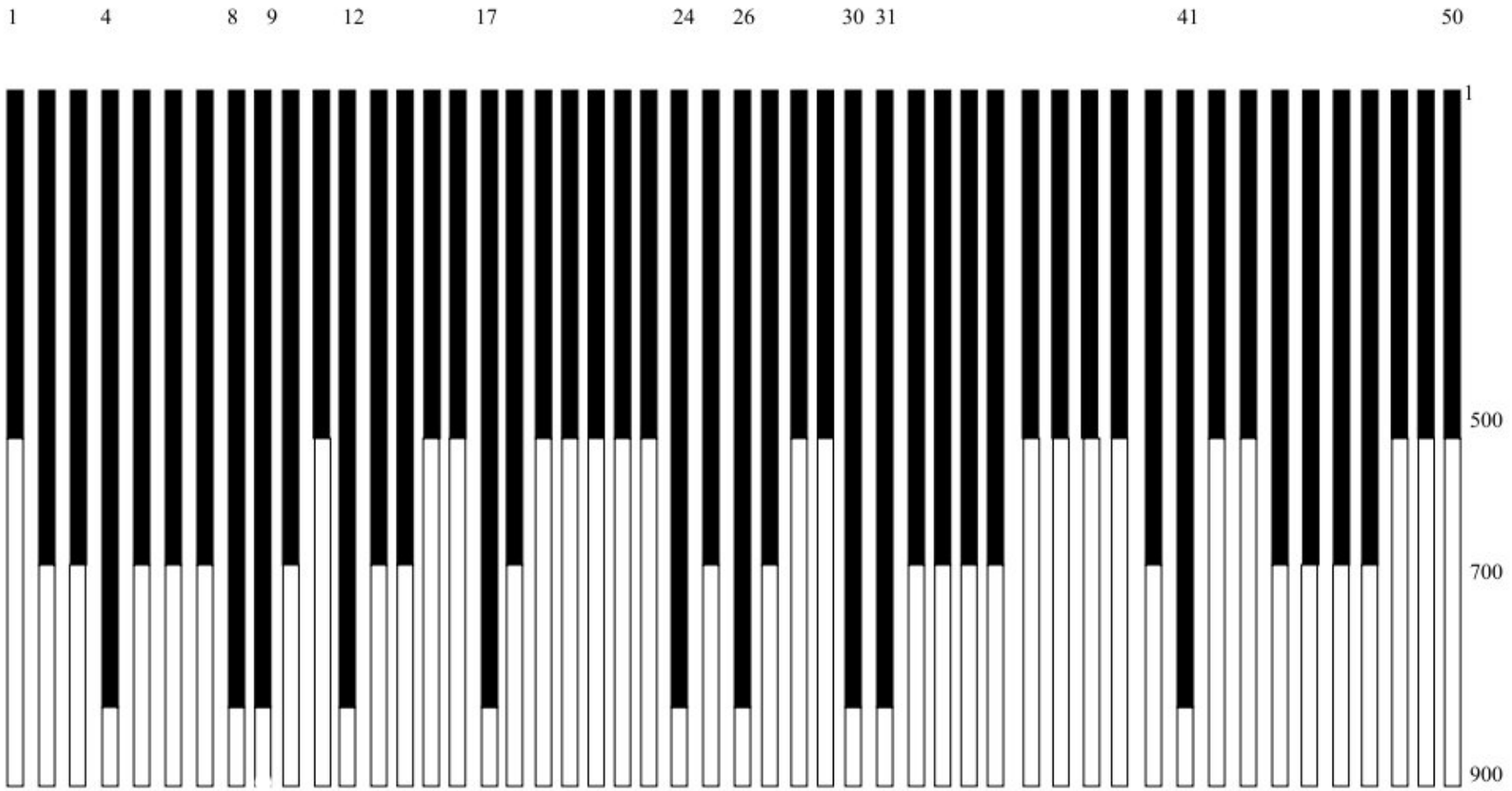
[2] In each case, the motif discovered in the real data set coincided with the planted motif, and its proporation was also close to u.

Supplementary Text 4

Performance of the Algorithm in the Presence of a Large Number of Local Optima

We generated synthetic data matrices of structures similar to that given in Supplementary Figure 1. The motivation for analyzing these synthetic data sets was to assess the performance of our algorithm when the search space comprises a large number of local optima. In this data matrix of size 1000×50, 10 columns were randomly chosen. Each chosen column, was filled with 900 1s and 100 0s. In other words, the first 900 elements of each chosen column were 1, and the remaining 100 were 0. Of the remaining 40 columns of the data matrix, 20 were randomly chosen, and each chosen column was filled with 700 1s and 300 0s. Each of the remaining 20 columns, was filled with 500 1s and 500 0s. For the example data matrix presented in Supplementary Figure 1, the set of columns, S, filled with 900 1s and 100 0s is: $S = \{4, 8, 9, 12, 17, 24, 26, 30, 31, 41\}$. Thus, this set of sites comprises the motif $(1, 1, \ldots, 1)$ of length $p = 10$, with $G(S) = 900$. However, this motif is clearly almost impossible to find, because there is exactly one such among $\binom{50}{10}$ possibilities. Among these $\binom{50}{10}$ points in the search space, there are $\left[\binom{50}{10} - \binom{30}{10}\right]$ 10-site combinations at which the sequence will be $((1))_{1\times 10}$ with a frequency of 500, and $\left[\binom{30}{10} \quad 1\right]$ 10-site combinations at which the sequence will be $((1))_{1\times 10}$ with a frequency of 700. Thus, the set of 10 out of 50 sites versus the frequency distribution of individuals has a very discrete structure: there is only one element in this set with 900 individuals, a very large number of elements with 700 individuals and a similarly large number of elements with 500 individuals. Clearly, therefore, to find the element with 900 individuals is nearly impossible. In 1000 runs of our algorithm with μ=10 and different initial values of c, we were never able to discover the correct motif. Invariably, the convergence was to a string with frequency either 500 or 700. However, when $(m_{500}, m_{700}, m_{900})$, where m_i denotes the number of columns in each of which there are i 1s and $(1000 - i)$ 0s $(i = 500, 700, 900; m_{500} + m_{700} + m_{900} = 50)$, was changed from (20,20,10) to other sets

of values, the proportions of runs in which the correct motif of length 10 was discovered increased. In other words, when the structure of the search space was slightly changed so that there were multiple - not just one - elements in the space with 900 individuals, the algorithm converged correctly and identified the motif. Simulation experiments were performed with various values of the control parameter c, the results of which are presented in Supplementary Table 4. Best results were obtained with c=200.



Supplementary Figure 1. Model structure of the synthetic data sets with multiple local optima. [Dark boxes are filled with 1; white boxes are filled with 0. Ten columns have 900 1s and 100 0s. Expected motif is (1,1,1,1,1,1,1,1,1,1) at sites (4, 8, 9, 12, 17, 24, 26, 30, 31, 41).]

Supplementary Table 4. Results of 1000 Simulation Runs for Different Structures of Synthetic Data Set 2 as Specified by $(m_{500}, m_{700}, m_{900})$ for Different Values of the Control Parameter c[1]

$(m_{500}, m_{700}, m_{900})$	c	% runs in which the motif was correctly identified	Mean±s.d. of the number of sweeps to convergence
(17,17,16)	10	36.8	1101.44±540.73
	30	50.8	986.43±574.89
	50	51.0	976.62±538.11
	100	52.0	926.69±589.32
	200	55.0	896.77±571.40
(16,16,18)	10	84.5	823.33±504.37
	30	92.0	627.87±489.17
	50	93.9	590.74±489.17
	100	92.5	638.92±503.21
	200	93.6	621.96±489.63
(15,15,20)	10	99.4	458.32±354.85
	30	99.8	270.20±237.87
	50	99.9	295.04±268.50
	100	100	293.07±268.81
	200	100	288.20±267.71

[1] The significance level of the motif, when correctly identified, was $\leq 10^{-15}$ when assessed by the "drop" procedure, and was = 0 when assessed by the bootstrap procedure (indicating that no motif "better" than that identified by the algorithm existed in the data).

Supplementary Text 5

Performance of the Algorithm when the Motif Proportion or Sample Size is Small

We generated multiple synthetic Data Sets 1 with u=0.1; that is, only 10% of individuals carry a known motif of size p=10. Further, we genenerated multiple synthetic Data Sets 1 with N=50; that is, data sets with small sample sizes. Although, both these scenarios are somewhat unrealistic, we carried out these simulation experiments to examine the limits to which our algorithm can be pushed. The results are given in Supplementary Table 5(a) and (b). We have used relatively small values of c, which is what he prescribe should be used when the motif frequency or the sample size is small. We find that even in these extreme cases, our algorithm performs well.

Supplementary Table 5. Performance of the algorithm on Synthetic Data Set 1 with (a) N=200 for different values of the variables L (number of segregating sites) and u (proportion of the planted motif among N) = 0.1; and, (b) N=50, L=200 and various values of $u. (The mean and s.d. of the number of sweeps to convergence and the % of simulation runs in which the planted motif was correctly identified are based on 1000 independent simulation runs for each combination of values of the variables and the parameter.)

(a)

L	c=50			c=100		
	No. of Sweeps			No. of Sweeps		
	Mean	s.d.	% correct	Mean	s.d.	% correct
50	1203.67	575.49	100	208.36	94.30	100
100	3660.30	1157.10	100	990.04	494.44	87

(b)

u	c=50			c=100		
	No. of Sweeps			No. of Sweeps		
	Mean	s.d.	% correct	Mean	s.d.	% correct
0.3	2067.20	1847.40	74	1983.21	1541.60	85
0.5	551.36	275.80	100	657.20	301.60	100
0.7	398.60	223.80	100	382.10	212.50	100

Supplementary Text 6

Method of Creating Synthetic Data Set 2

Two separate data matrices, each of size $N \times L$, corresponding to the cases and controls, were created. Elements of each column of each matrix were randomly filled with 1 or 0; the proportion of 1s occurring in any column was taken to be 0.5. Then p columns (polymorphic sites) were chosen at random. In the first data matrix corresponding to the cases, a set of $[N \times u_1]$ rows were randomly chosen, where $0 < u_1 < 1$. In each of these rows, the elements corresponding to the p chosen columns were replaced with 1. Thus, we planted, in the case data matrix, a motif (1,1,...,1) of length p in a proportion of u_1 individuals. Under the common-disease, common-variant model (Collins et al. 1998), each of the p sites (SNPs) carries a small relative risk, RR, to the disease, that collectively results in a large haplotype (motif) relative risk. If v_i and w_i denote, respectively, the number of 1s (that is, the specific nucleotide that confers a higher risk) at the i^{th} site ($1 \leq i \leq p$), among cases and controls, then $\mathrm{RR} = v_i/w_i$. (We have assumed that the site-specific relative risk is the same for each of the p sites.) Hence, in the data matrix corresponding to the controls, for the i^{th} of the p sites (that is, the i^{th} column), we placed 1s in $[N \times w_i]$, where $w_i = v_i/RR$, randomly chosen rows, and filled the remaining elements in that (i^{th}) column with 0s.

Supplementary Text 7

Method of Creating Synthetic Data Set 3

For constructing the data set, we first created a data matrix of size 1000×50, and assigned a value of 0 or 1 to each cell with probability 0.5. Then, we randomly selected 10 sites (that is, 10 columns of the data matrix) from the set (Π) of 50 sites, and changed the 0s to 1s at each site (column) so that at each of these sites the proportion of 1s among the 1000 individuals was ≥ 0.8. This resulted in the data matrix, D_1, of the ancestral population which expectedly has a motif of length 10 comprising the set of the 10 randomly selected sites, which we shall denote as Π_1. We then created two daughter populations of this ancestral populations. The data matrix, of size 1000×50, corresponding to the first daughter population was initially created by sampling 1000 rows (each with 50 columns), with replacement, from the data matrix of the ancestral population. We then selected a set (Π_2) of 5 sites randomly from $\Pi \backslash \Pi_1$, and at these selected sites we randomly replaced 0s by 1s in the initial data matrix of the first daughter population such that the proportions of 1s among the 1000 individuals at each of these 5 sites was ≥ 0.8. This yielded the final data matrix, D_2, corresponding to the first daughter population in which the motif expectedly comprises sites belonging to $\Pi_1 \cup \Pi_2$ of length 15. For the second daughter population, the initial data matrix was similarly created. A set (Π_3) of 5 sites were chosen from $\Pi \backslash (\Pi_1 \cup \Pi_2)$ and the final data matrix, D_3, was similarly created. In the second daughter population, the expected motif of length 15 has sites belonging to $\Pi_1 \cup \Pi_3$.

Supplementary Table 6. Mean + s.d. of the Number of Sweeps to Convergence in 1000 Independent Simulation Runs for Synthetic Data Set 3

Population (Data Matrix)	Mean ± s.d. of the number of sweeps to convergence
Population 1 (D_1)	45.76 ⊥ 32.15
Population 2 (D_2)	33.25 ⊥ 12.96
Population 3 (D_3)	32.86 ± 13.07