TreeDT: Gene Mapping by Tree Disequilibrium Test

Petteri Sevon
Dept. of Computer Science & Finnish Genome Center, Univ. of Helsinki
P.O. Box 63
FIN-00014 Univ. of Helsinki, Finland
+358 9 19125482
petteri.sevon@cs.helsinki.fi

Hannu T.T. Toivonen
Nokia Research Center & Rolf Nevanlinna Institute, Univ. of Helsinki
P.O. Box 407
FIN-00045 Nokia Group, Finland
+358 50 4837669
hannu.tt.toivonen@nokia.com

Vesa Ollikainen
Finnish Genome Center, Univ. of Helsinki
P.O. Box 63
FIN-00014 Univ. of Helsinki, Finland
+358 9 19125480
vesa.ollikainen@cs.helsinki.fi

ABSTRACT

We introduce and evaluate TreeDT, a novel gene mapping method which is based on discovering and assessing tree-like patterns in genetic marker data. Gene mapping aims at discovering a statistical connection from a particular disease or trait to a narrow region in the genome. In a typical case-control setting, data consists of genetic markers typed for a set of disease-associated chromosomes and a set of control chromosomes. A computer scientist would view this data as a set of strings.

TreeDT extracts, essentially in the form of substrings and prefix trees, information about the historical recombinations in the population. This information is used to locate fragments potentially inherited from a common diseased founder, and to map the disease gene into the most likely such fragment. The method measures for each chromosomal location the disequilibrium of the prefix tree of marker strings starting from the location, to assess the distribution of disease-associated chromosomes.

We evaluate experimentally the performance of TreeDT on realistic, simulated data sets. We also compare the results to those obtained using TDT (transmission/disequilibrium test), an established method for gene mapping, and Haplotype Pattern Mining (HPM), an earlier data mining method.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *data mining, scientific databases*
J.3 [**Life and Medical Sciences**]: Biology and Genetics
G.3 [**Probability and Statistics**]: Statistical Computing

General Terms

Algorithms, Experimentation.

Keywords

Gene mapping, algorithms, permutation tests, prefix trees.

1. INTRODUCTION

Gene mapping aims at discovering a statistical connection from a given trait or disease to a narrow region in the genome, which can then be further investigated by laboratory methods for a gene that affects the trait. In particular, the discovery of new disease susceptibility genes can have an immense importance for human health care. The gene and the proteins it produces can be analyzed to understand the disease causing mechanisms and to design new medicines. Further, gene tests on patients can be used to assess individual risks and for preventive and individually tailored medications. Obviously, gene mapping is receiving increasing interest among medical industry.

Genetic markers along chromosomes provide data that can be used to discover associations between patient phenotypes (e.g., diseased vs. healthy) and chromosomal regions (i.e., potential disease gene loci). The growing number of available genetic markers, anticipated to reach hundreds of thousands in the next few years, offers new opportunities but also amplifies the computational complexity of the task.

We introduce TreeDT, a novel method for gene mapping. It analyses the observed strings of markers by constructing tree-structured patterns that reflect the possible genetic history of a disease susceptibility (DS) gene. The gene is then predicted to be where the strongest genetic contribution is visible in the trees. The contributions of TreeDT are:

(1) a novel approach to gene mapping using tree patterns,
(2) an efficient algorithm for generating and testing tree patterns,
(3) a method for estimating the statistical significance of findings.

For reasons of brevity we focus on the first two in this paper. We evaluate the method experimentally with realistic, simulated data, and compare it to previous state of the art methods in gene mapping.

2. PROBLEM BACKGROUND

Let us assume the goal is to locate a disease-susceptibility gene for a given disease. We next briefly review the genetic background; without loss of generality, we restrict the discussion in this paper to one chromosome.

Marker Data A genetic *marker* is a short polymorphic region in the DNA, denoted here by M1, M2, …. The different variants of DNA that different people have at the marker are called *alleles*, denoted in our examples by 1, 2, 3, …. The number of alleles per marker is small: typically less than ten (for so called microsatellite markers) or exactly two (for so called SNPs). The collection of markers used in a particular study is its *marker map*, and the corresponding alleles in a given chromosome constitute its *haplotype* (Figure 1). It is a major task of a gene mapping study to design the marker map and to obtain the haplotype data. That is, however, where we start, and for the purposes of this paper the

Conference '00, Month 1-2, 2000, City, State.

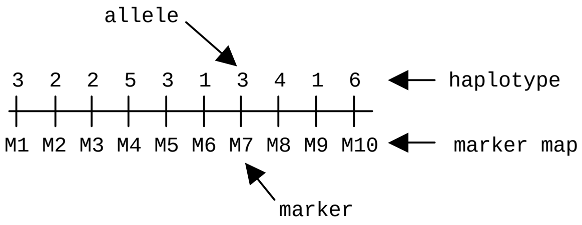


Figure 1. A marker map of ten markers and a sample haplotype consisting of alleles in adjacent markers.

input data consists of haplotypes of diseased and control persons – or, in computer science terms, aligned allele strings, classified to positive and negative examples.

Linkage disequilibrium All the current carriers of a disease-susceptibility gene have inherited it from a founder who introduced the gene mutation to the population (Figure 2). If there has been only one such founder, then current carriers are related and segments from the mutation carrying founder chromosome are over-represented among the affected at mutation locus. It can then be possible to observe *linkage disequilibrium* (LD), non-random association between nearby markers.

Gene Mapping In diseases with a reasonable genetic contribution, and especially in population isolates where few founders have introduced the mutation, affected individuals are likely to have higher frequencies of founder alleles and haplotype patterns near the DS gene than control individuals. This is the starting point of LD-based mapping methods: where does the set of affected chromosomes show linkage disequilibrium? The problem is far from trivial, however. The coalescence process is stochastic; mutation carriers often only have a higher risk of being diseased than non-carriers, and in a case–control study both groups are usually mixes of carriers and non-carriers; finally, there is missing information.

Summary of Background and Problem Genetic markers provide an economical, sparse view of chromosomes. Even sparsely located markers can be very informative: given an ancestor with a mutated gene, the descendants that inherit the gene are also likely to inherit alleles of nearby markers. The exact probability of inheriting any combination of markers depends on the gene location with respect to the markers, the population history or the coalescence history, and marker mutations; all of these are unknown.

Our framework consists of a case–control setting, where the input consists of haplotypes. Each individual contributes a chromosome pair, so the number of chromosomes is twice the number of individuals. We ignore the fact that chromosomes come in pairs and simply consider the input data as consisting of a set of disease-associated haplotypes (from the cases) and a set of control haplotypes.

The LD-based gene mapping problem is now the following. The input consists of a marker map, and a set of disease-associated haplotypes and a set of control haplotypes on the given map. The task is to predict the location of a disease susceptibility gene on the map.

3. Method

For any pair of chromosomes in the sample there has been a common origin in the population history, an ancestral chromosome at which their paths have diverged. Due to recombinations different parts of chromosomes have different histories, and at any given location the chromosomes in the sample and their most recent common origins form a coalescence tree. In the coalescence tree for the DS gene location, all the chromosomes in one or more subtrees carry the DS mutation, and we should observe excess of disease-associated haplotypes as the leaves of these subtrees. Looking at coalescence trees for various locations, the closer the location is to the DS gene the more and larger subtrees are identical to those in the tree at the DS gene location.

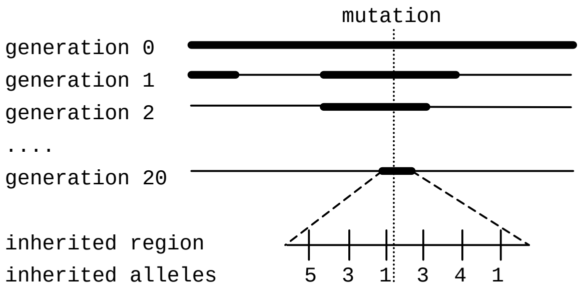


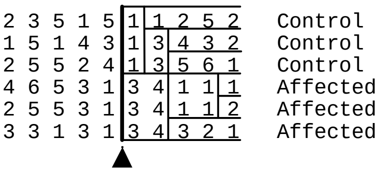
Figure 2. A carrier of the mutation in generation 20 has inherited alleles from the ancestral chromosome in generation 0 around the gene locus.

Based on the observed haplotypes, TreeDT evaluates the most likely coalescence tree at a number of locations along the analyzed chromosome, and then assesses the subtree clustering of disease-associated haplotypes in these trees. For the latter task we introduce a novel tree disequilibrium test, intended for predicting DS gene locations. The vicinity of the location for which the test gives the lowest p value is the most likely candidate area for the DS gene location. The method also computes the corrected overall p value for the best finding. It can be used for predicting whether the chromosome carries a DS gene at all or not.

Haplotype Prefix Trees Given a location in the chromosome – a potential gene locus – the haplotypes to the right (or to the left) of the location can be organized into a prefix tree (Figures 3 and 4). The tree structure may be considered as a possible coalescence tree for the location, with the following exceptions: 1) The order of nodes may differ from that in the true coalescence tree, e.g., in Figure 4, 34--- might actually be a more recent node than 3411-. However, because the expected length of the shared region of two chromosomes decreases monotonically as the time from their divergence increases, it is easy to see that the order given by subsumption is the most likely one. 2) Because haplotypes may also share a substring by chance, the internal nodes may represent a combination of nodes in true coalescence tree.

Instead of considering alternative coalescence trees leading to the same observed haplotypes, TreeDT uses the unique haplotype prefix tree as a canonical representation of such a set of coalescence trees. TreeDT builds two prefix trees, one to the left and one to the right, between each pair of consecutive markers and tests their disequilibrium.

Tree Disequilibrium Test The tree disequilibrium test for a haplotype (prefix) tree T tests the alternative hypothesis *The distribution of the disease-association statuses deviates in some subtrees of T from the overall distribution of statuses* against the null hypothesis *The disease-association statuses are randomly*



```
2 3 5 1 5 1 1 2 5 2   Control
1 5 1 4 3 1 3 4 3 2   Control
2 5 5 2 4 1 3 5 6 1   Control
4 6 5 3 1 3 4 1 1 1   Affected
2 5 5 3 1 3 4 1 1 2   Affected
3 3 1 3 1 3 4 3 2 1   Affected
```

Figure 3. String-sorted set of haplotypes to the right from the location pointed by the arrow.

distributed in the leaves of T. TreeDT identifies the set S of subtrees in which the observed status distribution deviates most from the expectation under the null hypothesis. In the next subsection we discuss how to estimate the significance of the deviation as a p value and to use in gene mapping.

For measuring the disequilibrium, we use a variant of the Z test. The test statistic Z_k for a tree with k deviant subtrees $T_1, \ldots, T_k$ is

$$Z_k = \sum_{i=1}^{k} \frac{a_i - n_i p}{\sqrt{n_i p(1-p)}},$$

where a_i is the number of disease-associated haplotypes and n_i the total number of haplotypes in subtree $T_i \in S$, and p is the proportion of disease-associated haplotypes in the sample. The score measures the distance of the observed number of disease-associated chromosomes (a_i) from the expectation ($n_i\ p$) in standard deviations (square root of $n_i\ p\ (1-p)$), under the assumption of binomial distribution with parameters n_i and p. We use a one-tailed test, since we are interested only in subtrees in which the proportion of disease-associated haplotypes is greater than expected.

We could use a $2\times(k+1)$ χ^2-statistic as a measure of deviation for a given subtree set S. The χ^2-statistic, however, is not easily maximized in the space of all possible subtree sets and is therefore not a very practical choice.

Significance Tests Z_k is a measure for the disequilibrium of a given tree, corresponding to a certain location in the chromosome, with given k deviant subtrees. Given a tree, TreeDT finds for each k the set S of subtrees that maximizes Z_k. (Z_k can be efficiently maximized simultaneously for all k using a recursive algorithm, as shown in the Algorithms section.) However, in order to find the best value of k for the given tree, simple maximization is not possible. Since the statistics for different degrees of freedom k are not comparable, TreeDT estimates the p value for each maximized Z_k (under the null hypothesis of random distribution of disease status). Because the distribution of the maximized Z_k is very complex and dependent on the tree structure, p values are estimated by a permutation test.

In order to get a single p value for the disequilibrium at a given location, we need to combine the information from the trees to the left and to the right of the location. As a combined measure we use the product of the lowest p value over all k from each side. Again, since the measures are not necessarily directly comparable, a new p value for the combination is estimated. The results are now comparable between different locations.

The output of TreeDT is essentially the p value ranked list of locations. A point prediction for the gene location is obtained by taking the best location; a (potentially fragmented) region of length l is obtained by taking best locations until a length of l is covered.

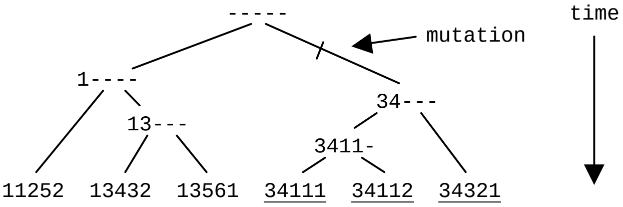


Figure 4. The prefix tree, and also a possible coalescence tree, for the haplotypes of Figure 3.

Since multiple locations are tested for a p value, and also since the p values at nearby locations are not independent, a direct link between the p value and the probability that the gene is indeed close to the location can not be established. The p values are used simply as a method of ranking the locations.

However, a single corrected p value for the best finding can be obtained with a third test using the lowest local p value as the test statistic. This p value can also be used to answer the question whether there is a gene in the investigated area in the first place or not. All these three nested p value tests (for each tree and k, for each location, for the best location) can be carried out efficiently at the cost of a single test [8].

4. Algorithms

Constructing Haplotype Prefix-Trees The haplotype prefix-trees to the left and right from each analyzed location can be efficiently identified using a string-sorting algorithm. The algorithm produces as intermediate results for each marker the sorted list of the partial haplotypes to the right from the marker. All the right-side trees can be easily derived from these intermediate lists, because the haplotypes belonging to a single node form a continuous block in the sorted list. The left-side trees can be identified similarly by sorting the inverted haplotypes. The computational cost of constructing the trees is linear both in the number of markers and the number of haplotypes and it is negligible compared to the cost of the permutation test procedure.

An Algorithm for Maximizing the Tree Disequilibrium Statistic It is essential that the time-complexity of the algorithm for maximizing the Z values is as low as possible, because it must be executed for each tree location and permutation in turn. The key observation is that if subtree set S maximizes $Z_{|S|}$ in a tree T, then the restriction S' of S to any subtree T' of T maximizes $Z_{|S'|}$ for T'. Also, if $S_1 \cap S_2 = \emptyset$ then $Z(S_1 \cup S_2) = Z(S_1) + Z(S_2)$. These observations lead us to the following recursive algorithm that propagates the locally maximized Z values upwards in the tree.

```
Input: A haplotype prefix tree $T$
Output: Maximum values of $Z_k$ in the tree $T$ for each $k$
Call Maximize($T$)

Maximize($T$):
If $T$ is not a leaf:
1.  For each immediate subtree $T_i$ of $T$: Recursively call
    Maximize($T_i$).
2.  For each $k$: calculate the maximum value $Z_{MAX,\ k}(T)$ for $Z_k$
    over all $S$ that can be obtained by combining subtree sets
    from each subtree $T_i$ of $T$.
3.  Calculate $Z_1$ for $T$. If $Z_1 > Z_{MAX,\ 1}(T)$ then set $Z_{MAX,\ 1}(T) := Z_1$.
If $T$ is a leaf, then set $Z_{MAX,\ 1}(T) := 0$.
```

Step 2 can be further refined:

```
2.1   Set $Y_k := 0$ and $Z_{\mathrm{MAX},\,k}(T) := 0$ for all $k$, $1 \le k \le n$, where $n$ is the number of leaves in $T$.
2.2   For each subtree $T'$ of $T$ :
2.2.1   For each pair $(i,j)$, $1 \le i \le p$ and $1 \le j \le q$, where $p$ is the number of leaves in $T'$ and $q$ is the total number of leaves in all the subtrees processed prior to $T'$:
2.2.2   If $Z_{\mathrm{MAX},\,i}(T') + Y_j > Z_{\mathrm{MAX},\,i+j}(T)$, then set $Z_{\mathrm{MAX},\,i+j}(T) := Z_{\mathrm{MAX},\,i}(T') + Y_j$.
2.2.3   For each $k$, $1 \le k \le p$: If $Z_{\mathrm{MAX},\,k}(T') > Z_{\mathrm{MAX},\,k}(T)$, then set $Z_{\mathrm{MAX},\,k}(T) := Z_{\mathrm{MAX},\,k}(T')$.
2.2.4   For each $k$, $1 \le k \le p+q$: If $Z_{\mathrm{MAX},\,k}(T) > Y_k(T)$, then set $Y_k(T) : = Z_{\mathrm{MAX},\,k}(T)$.
```

The time complexity of the algorithm is $O(n^2)$, where n is the number of leaves in the tree i.e. the number of haplotypes in the data set. By setting an upper limit k for the size of the subtree sets, the average time complexity can be reduced to $O(n)$ with a constant coefficient proportional to k^2, k being typically small, ≤ 10. In principle there is no need to set an upper limit – the number of leafs, i.e., the number of chromosomes is the maximum number of subtrees – but whenever LD-mapping is applicable, the majority of the mutation carriers is concentrated in only few such subtrees in which the shared region is long enough to identify a deviant substring. In the experiments for this paper we use an upper limit of 6 subtrees.

Multiple Nested Permutation Tests The straightforward algorithm for a three-level nested permutation test using nested loops would have time complexity proportional to n^3, where n is the number of permutations at each level. The test would be intractable already with rather low permutation counts. However, the time complexity can be drastically reduced using the same set of permutations at each level of the test and thus only maximizing the Z_k-values n instead of n^3 times for each location [8].

5. Related Work

Most current gene mapping methods based on linkage disequilibrium look just at individual markers or neighboring markers, measure their association to the disease status, and predict the gene locus to be co-located with the strongest association. However, since different mutation carriers share different segments, there is no single marker or pattern that is representative of the shared segments.

In the recent years, several statistical methods have been proposed to detect LD [1][3][4][6][10]. The emphasis has been on fairly involved statistical models of LD around a DS gene. The methods tend to be computationally heavy and therefore better suited for fine mapping than genome screening.

Haplotype Pattern Mining or HPM [11] is based on analyzing the LD of sets of haplotype patterns, essentially strings with wildcard characters. The method first finds all haplotype patterns that are strongly associated with the disease status, using ideas similar to the discovery of association rules. In the second step, each marker is ranked by the number of patterns that contain it. Either this score is used as a basis for the prediction or, preferably, a permutation test is used to obtain marker-wise p values. HPM has been extended for detecting multiple genes simultaneously [12] and to handle quantitative phenotypes and covariates [7].

Nakaya et al investigate the effect of multiple separate markers, each one thought to correspond to one gene, on quantitative phenotypes [5]. They do not handle haplotype patterns.

An alternative approach for LD-based mapping is linkage analysis. The idea is to analyze family trees, and to find out which markers tend to be inherited to offspring in conjunction with the disease. Transmission/disequilibrium tests (TDT) [9] are an established way of testing association and linkage in a sample where linkage disequilibrium exists between the mutation locus and nearby marker loci. TDT detects deviations between observed and expected counts for each allele, or, in its multipoint variant, haplotype of several alleles, transmitted from heterozygous parents to affected offspring. We performed the TDT analysis using GENEHUNTER2 software package [2].

6. Experiments

We compare TreeDT empirically to TDT, to multipoint TDT (m-TDT) using haplotypes of up to four alleles, and to HPM, our recent proposal based on pattern discovery. We evaluate the methods on difficult data collections carefully simulated to resemble a realistic population isolate.

Simulation of Data We designed several different test settings, with variation in the fraction (A) of mutation carriers in the disease-associated chromosomes, in the number of founders who introduced the mutation to the population, and in the amount of missing information. For statistical analyses, we created 100 independent artificial data sets in each test setting. Great care was taken to generate realistic data by a simulation procedure that included four steps: pedigree generation, simulation of inheritance, diagnosing, and sampling [8]. Here we only give an outline of the nature of the resulting data sets.

For a baseline test setting we selected a challenging disease model where only a small proportion (A=10%) of the disease-associated chromosomes carries the disease-predisposing mutation, a complication that often is encountered in the analysis of common diseases. In the baseline setting there is one founder, and on average 3.7% of alleles are missing, making the mapping task more difficult but also more realistic.

The location of the mutation was selected randomly and independently for each of the 100 data sets produced in every setting. Each data set was in turn collected from 100 affected individuals. The length of the region to be analyzed was 100 cM[1], and allelic data were created using a map of 101 equidistantly spaced markers, each having 5 alleles. Both chromosomes of each affected individual in each sample were labeled disease-associated whereas the control chromosomes were constructed from the non-transmitted alleles in the parental chromosomes. Each data set thus consisted of 200 disease-associated and 200 control chromosomes

Analysis of TreeDT First we assess the prediction accuracy of TreeDT with different values of A, the proportion of disease-associated chromosomes that actually carry the mutation (Figure 5A). The results are reported as curves that show the percentage of 100 data sets where the gene is within the predicted region, as a function of the length of the predicted region. Or, in other words, the x coordinate tells the cost a geneticist is willing to pay, in terms of the length of the region to be further analyzed, and the y coordinate gives the probability that the gene is within the region.

[1] Morgan is a unit of genetic length. 1 cM is the distance at which recombination occurs 1 out of 100 times, on average about 10^6 base pairs. Human chromosomes are about 50–300 cM.

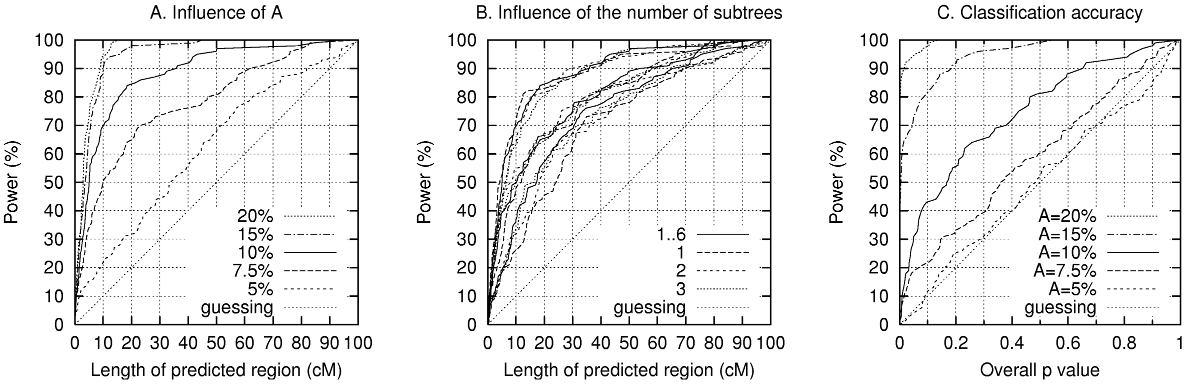


Figure 5. Analysis of the performance of TreeDT. A: Gene localization power with different values of *A*, the proportion of disease-associated chromosomes that actually carry the mutation. B: Gene localization power with different numbers of subtrees (method parameter, given in the legend) and different numbers of founders (population parameter; 1 for the highest set of curves, 2 for the curves in the middle, and 3 for the lowest set of curves). C: Classification accuracy for the existence of a disease susceptibility gene.

For *A*=20% or 15% the accuracy is very good, and with lower values of *A* the accuracy decreases until with *A*=5% only in 20-30% of data sets can the gene be localized within a reasonable accuracy of 10-20 cM. We remind the reader that the test settings have been designed to be challenging, and to test the limits of the approach.

Next we evaluate the effect of the only parameter of TreeDT, the number of deviant subtrees that are searched for in each tree. An upper limit of 6 subtrees, used in the previous test, is evaluated against fixed amounts of 1, 2, or 3 subtrees, with a varying number of founders that introduced the mutation (Figure 5B). As we increase the number of founders, evidence about the gene location becomes more fragmented, and accordingly the performance degrades. While the differences between different numbers of subtrees are not large, it is interesting to note that for each number of founders, the same number of subtrees gives marginally the best result. The upper limit of 6 subtrees gives consistently competitive results, so we continue using it in the following experiments.

Gene mapping studies like the ones imitated in the above tests assume, based on some other analyses, that a disease susceptibility gene is indeed present in the analyzed area. TreeDT has the important advantage over plain gene localization methods that it can also be used to predict whether the analyzed region contains a disease susceptibility gene at all or not. The overall *p* value TreeDT produces indicates the corrected significance of the best single finding, and by setting an upper limit for its value TreeDT can be used to classify data sets to ones that do or do not contain a gene. For data sets with no gene, TreeDT correctly produces overall *p* values that are uniformly distributed in [0,1]. So, smaller thresholds for *p* result in less false positives, but also in less true positives. Figure 5C shows the experimental relationship between power (ratio true positives/all positives) and overall *p* (ratio false positives/all negatives). For higher values of *A* the classification accuracy is extremely good. However, for *A*=5% the classification accuracy is no better than random guessing, although the localization accuracy for an existing gene is still adequate in 20-30% of the cases (Figure 5A).

Comparison to other methods TreeDT, HPM, and m-TDT have practically identical performance in localizing the DS gene in the baseline setting (Figure 6A). TDT is clearly inferior compared to the other methods. Tests with other values of *A* give similar results.

In a test setting with three founders who introduced the mutation to the population, differences between the three best methods start to appear (Figure 6B). TreeDT has an edge over HPM, which in turn has an edge over m-TDT. TDT barely beats random guessing.

Finally, we compare the methods with a large amount of missing data (Figure 6C). Expectedly, HPM is most robust with respect to missing data since it allows gaps in its haplotype patterns. Surprisingly, TreeDT is not much weaker than HPM, although no actions have been taken in it to account for missing or erroneous data. Performance of m-TDT degrades much more clearly.

Method to method comparisons (not shown) indicate that the prediction errors are mostly caused by random effects in population history – since different methods tend to make mistakes in the same data sets – rather than by systematic differences between the methods. However, those cases where one method succeeds and another fails will give useful input for further improvements of the methods.

The execution time of TreeDT for a single data set is about ten minutes using 1,000 permutations on a 450 MHz Pentium II. The respective time for HPM with permutations is over 20 minutes.

7. Discussion and Future Work

We have introduced TreeDT, a novel method for gene mapping. It is based on detecting linkage disequilibrium in the haplotype prefix trees to the right and left of the disease susceptibility gene location. We showed how tree disequilibrium can be efficiently evaluated between every pair of consecutive markers, and be subsequently tested for statistical significance using permutation tests. Empirical evaluation on a realistic, simulated data shows that the method is competitive with other recent data mining based methods, and clearly outperforms more traditional methods.

Our experiments show that TreeDT is effective in extreme conditions typical for current mapping problems: with lots of noise (only 10-20% of affected chromosomes carry the mutation, lots of missing data) and with small sample sizes (200 affected and 200 control chromosomes). However, the highest potential of the method lies in the data intensive tasks of future – such as

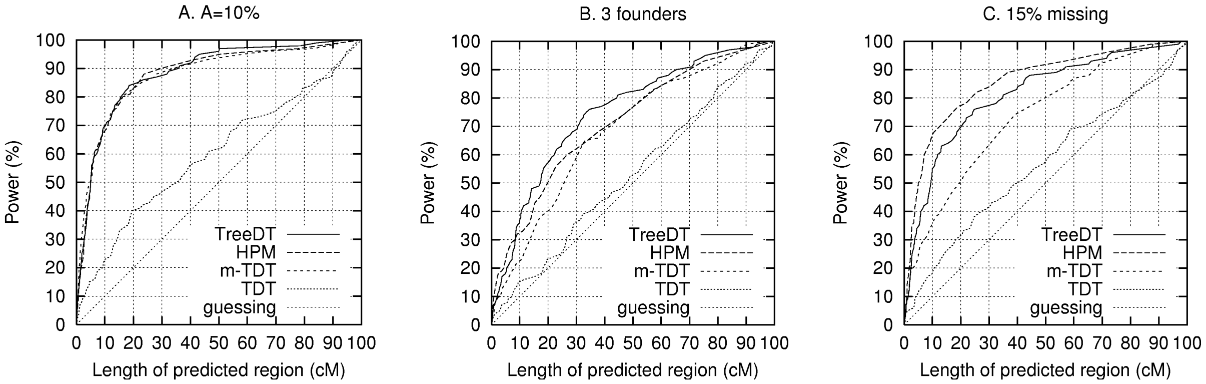


Figure 6. Comparison of the gene localization performance of TreeDT, HPM, multipoint TDT (m-TDT), and TDT. A: The baseline test setting. B: The baseline setting with three founders. C: The baseline setting with 15% missing data.

genome scanning with larger samples and larger number of markers – due to its low computational complexity.

In comparison to state of the art methods, TreeDT is most competitive. In terms of gene localization accuracy, it gave best results in the case of multiple founders and demonstrated good robustness with respect to missing data. Unlike the compared methods, TreeDT can be used to predict whether a gene is present at all or not. Finally, in comparison to its closest competitor, HPM, TreeDT has much smaller computational cost. An additional advantage of TreeDT is that it has only one input parameter, the (maximum) number of deviant subtrees, whereas for HPM one has to set several more or less arbitrary thresholds.

Our future work will address several issues. One is more complex haplotype data: robustness towards missing information, errors, and marker mutations is important with noisy, real-life data sets.

A whole set of issues concerns improving tests and models for the tree disequilibrium. Now we combine the left and right trees at a locus without considering how the haplotype strings actually extend over the locus; obviously we miss some information. Another way of improving the model performance is to average the disequilibrium test over all different tree structures. The test statistic itself will be improved to better account for the genetic processes that produces the data.

8. Acknowledgements

Thanks to Päivi Onkamo and Juha Kere for support and co-operation during this research. This research has been funded by Helsinki Graduate School in Computer Science and Engineering (HeCSE) and Graduate School in Computational Biology, Bioinformatics, and Biometry (ComBi).

9. References

[1] B. Devlin, N. Risch, and K. Roeder. Disequilibrium Mapping: Composite Likelihood for Pairwise Disequilibrium. *Genomics*, 36:1-16, 1996.

[2] L. Kruglyak, M. Daly, M. Reeve-Daly, E. Lander. Parametric and Nonparametric Linkage Analysis: a Unified Multipoint Approach. *Am J Hum Genet*, 58:1347-1363, 1996.

[3] L. Lazzeroni. Linkage Disequilibrium and Gene Mapping: an Empirical Least-Squares Approach. *Am J Hum Genet*, 62:159-170, 1998.

[4] M. McPeek and A. Strahs. Assessment of Linkage Disequilibrium by the Decay of Haplotype Sharing, with Application to Fine-scale Genetic Mapping. *Am J Hum Genet*, 65:858-875, 1999.

[5] A. Nakaya, H. Hishigaki, and S. Morishita. Mining the Quantitative Trait Loci Associated with Oral Glucose Tolerance in the Oletf Rat. *Proc. of Pacific Symposium on Biocomputing*, pp 367-379, January 4-9, 2000.

[6] S. Service, D. Temple Lang, N. Freimer, and L. Sandkuijl. Linkage-Disequilibrium Mapping of Disease Genes by Reconstruction of Ancestral Haplotypes in Founder Populations. *Am J Hum Genet*, 64:1728-1738, 1999.

[7] P. Sevon, V. Ollikainen, P. Onkamo, H. Toivonen, H. Mannila, and J. Kere. Mining Associations Between Genetic Markers, Phenotypes and Covariates. *Genetic Analysis Workshop 12, Genetic Epidemiology*, 21 (Suppl. 1), 2001. In press.

[8] P. Sevon, H. Toivonen, V. Ollikainen. TreeDT: gene mapping by tree disequilibrium test (extended version). Report C-2001-32, Department of Computer Science, University of Helsinki, Finland, 2001.

[9] R. Spielman, R. McGinnis, W. Ewens. Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-Dependent Diabetes Mellitus (IDDM). *Am J Hum Genet*, 52:506-516, 1993.

[10] J. Terwilliger. A Powerful Likelihood Method for the Analysis of Linkage Disequilibrium Between Trait Loci and One ore More Polymorphic Marker Loci. *Am J Hum Genet*, 56:777-787, 1995.

[11] H. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, M. Herr, and J. Kere. Data Mining Applied to Linkage Disequilibrium Mapping. *Am J Hum Genet*, 67:133-145, 2000.

[12] H. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, and J. Kere. Gene Mapping by Haplotype Pattern Mining. *Proc. Bio-Informatics and Biomedical Engineering*, pp 99-108, Arlington, VA, November 8-10, 2000.