Linkage analysis: A Bayesian approach

Martijn A.R. Leisink, Hilbert J. Kappen and Han G. Brunner

University of Nijmegen, Department of Biophysics
Geert Grooteplein 21, 6525EZ Nijmegen, The Netherlands
{martijn,bert}@mbfys.kun.nl

Abstract. In this article we propose a method for linkage analysis that is based on Bayesian statistics. It is non-parametric in the sense that there is no need to specify disease parameters such as penetrance values. We show that the method has significantly more statistical power than existing methods on artificially created databases. Finally, the possibility to extend the method to multi-locus diseases is discussed.

1 Introduction

One of the major questions in genetics is to link an observed disease to a certain region on the chromosome. When a disease is thought to be (partially) genetic, scientists collect a large data base of pedigree data from families in which one or more affected individuals occur. Most or all of the family members are genotyped with markers, a technique which can be used to reconstruct the way genetic information has passed along the generations (very similar to tests for fathership). Although the information it provides is never complete, it is in practice a method that is good enough to localize genes within a certain region. Over the last decades, several approaches were made either on the method of extracting the inheritance information, or on the test statistics, which indicate how likely the genetic information at a certain location is the cause of the disease.

For reconstructing the inheritance pattern in a pedigree, there exist algorithms, which are exact and make use of all available marker data. They are either exponential in the number of markers taken into account [3] or the size of the pedigree [1]. In this article we use pedigrees for which both methods are tractable.

The second part of linkage analysis (how well is the disease pattern in the family explained by the reconstructed inheritance) is the main focus of this article. Several statistics were proposed for this purpose, among which the traditional LOD-score, the maximum LOD-score (an extension), the APM-score and the NPL-score [1]. The last one is the de facto standard in the field of linkage analysis (from the program called 'Genehunter'), although the (maximum) LOD-score is also frequenly seen.

In the next section we describe the problem in terms of a Bayesian network and propose to use the likelihood of the observed data as a test statistic. It will be shown in section 3 that this statistic is more powerful than the NPL-statistic. In the last section we discuss two other important properties of the proposed

method: 1) It is straight-forward to extend its definition to diseases caused by more than one gene and 2) it is possible to incorporate prior knowledge about the disease either given by experts or by large population studies.

2 The method of linkage analysis

In linkage analysis, the objective is to link an observed affection status (phenotype) of all individuals in a pedigree to the inheritance pattern observed at a certain location on the chromosome. On the one hand there is information about the affection status of each individual (affected or unaffected) and the disease mechanism (dominant, recessive, etc). The latter might be unspecified. On the other hand information about the inheritance pattern—which genes (paternal or maternal) are inherited by the child—is available.

For each locus on the chromosome, the inheritance pattern within a pedigree is denoted by the inheritance vector, v_i, which contains two binary valued variables for each non-founder[1]. The subscript i denoted the location at the chromosomes. The binary variables indicate which one of the chromosomal pair is copied from the father and which one from the mother. Unfortunately, the inheritance vector cannot be obtained directly, but an indirect measurement using 'markers' (genotyping) is used to reconstruct v_i. Since in most cases this v_i cannot be determined uniquely, one computes a probability distribution over all states v_i. Although this procedure is far from trivial, since it is not the main object of this article, we assume $p(v_i)$ as given and refer to [3] and [1] for the standard procedures. Note that, due to cross-overs, $p(v_i)$ can vary along a single chromosome.

2.1 How likely is my observation

At this point many authors define some kind of scoring function $S(A, v_i)$ depending on the affection status, A, of all individuals and the inheritance pattern. Multiplying this by $p(v_i)$ and summing out v_i, gives a value which indicates how well the observed affection status is explained by the observed inheritance information (markers). Several scoring functions are proposed, among which the LOD-score (log-odds ratio) and a so called NPL-score (non-parametric linkage). Although the LOD-score has a long history in the field of genetics, due to space limitations we restrict the comparison made in this article to the more recent NPL-score and refer to [1] for a comparison between LOD and NPL.

In figure 1 we propose a graphical model which directly leads to a logical scoring function: the likelihood of the affection status, given v_i, thus $S(A, v_i) = p(A|v_i)$. The figure shows the model belonging to a very small pedigree, but it is straight-forward to extend it too larger pedigrees. The node 'd' determines the probability that a mutated gene[2] occurs in the population. The probability

[1] Non-founders are individuals in the pedigree whos parents are also in the pedigree

[2] The genotype, G, consists of two copies of a single gene. Each of them can be mutated or not. Depending on the disease model, one or more of these combinations can cause the disease.

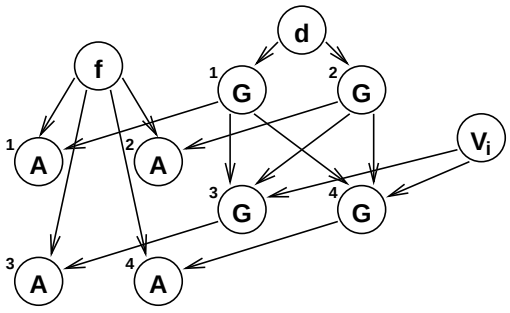

Fig. 1. The graphical model to link a genetic disease to a certain inheritance pattern for a very small pedigree of four individuals. The nodes labeled with 'A' are affection states (affected or unaffected), which are directly caused by 'G', the genotype of the individual and 'f', the disease model. The genotype of the founders (the parents or individual 1 and 2) is determined by population statistics 'd', whereas the genotype of the children (individual 3 and 4) can be determined from that of their parents given 'v_i'. This 'v_i', also known as the inheritance vector, consists of four binary states which determine which paternal genes are copied.

that an individual has precisely one mutated gene, for instance, is then given by $2d(1-d)$. Once the genotype 'G' of the founders in the pedigree is set, the genotype of the non-founders is completely determined by 'v_i'.

The genotype, G, can be either '$--$', '$-+$', '$+-$' or '$++$', where a plus denotes the presence of a mutated gene. Since there is no biological difference between '$-+$' and '$+-$' one only needs the probability of the affection status given the number of mutated genes an individual has. For instance, $p(A|\#\text{mutated genes}) = \{0, 0, 1\}$ for a purely recessive disease. In principle all probability tables are possible. The exact three values are known as 'penetrance values' and are represented in the figure with the symbol 'f'. It is immediately clear from the graphical model, that (although the penetrance values are often not known) all individuals share the same disease model.

Once this graphical model is specified, it is rather trivial to compute the total score:

$$S_i(A) = \int_{v_i} p(A|v_i)\, p(v_i) = \int_{v_i} \int_f \int_d \int_G p(A|Gf)\, p(G|v_i d)\, p(f)\, p(d)\, p(v_i) \tag{1}$$

In practical situations, the score is averaged over a few to several hundred pedigrees to make it more robust. This total score is computed at many locations i on all the chromosomes. Obviously, the regions with the highest scores are the most probable regions responsible for causing the disease. Note how easily we can deal with the missing information about the disease model. Since it is not specified, we simply integrate over all possible values. In case we had (partial) information about these parameters, we could simply incorporate that as prior knowledge in $p(f)$ or $p(d)$. Otherwise these are flat.

3 Statistical power

To assess the quality of the proposed method compared to Genehunter's NPL-score, we created artificial data sets each consisting of ten pedigrees. Each pedigree was a family with three children of which one was married and had two children. Two disease models were used to generate the data: A recessive model with $d = 0.1$, $f_0 = 0.05$, $f_1 = 0.05$ and $f_2 = 0.9$ and a dominant model with $d = 0.02$, $f_0 = 0.05$, $f_1 = 0.9$ and $f_2 = 0.9$. For all the data the likelihood as presented in this article was computed and, using Genehunter, their NPL-score.

We assumed that for every pedigree only the marker data at the disease locus is available. Since the procedure to handle missing marker data and to make use of adjacent markers does not differ between the methods presented here, there is no need to investigate this regime. To make the single marker informative enough, we assume that it is very polymorphic: ten alleles with equal population occurance. This situation is comparable to chromosome wide available data with less polymorphic markers. After randomly assigning alleles to the founders, the marker data for the non-founders was computed according to a randomly chosen v, which is, of course, not presented to any of the methods. The index i is dropped, since we only investigate a single locus, which can be either linked or unlinked to the disease. Using the same v the number of mutated genes is computed for every individual and the affection status is set randomly using the appropriate penetrance value. A pedigree is included in the data set if at least two non-founders are affected. In this way, we construct 3,000 data sets for the recessive as well as for the dominant case.

We compare the scores found in these data sets with the scores from 3,000 artificially created null data sets. These null data sets are simply copies from the other data sets, but the marker data is generated again using a new random inheritance vector. In this way, there is no linkage between the affection status and the marker data. Therefore, we expect the scores in this data base (the null scores) to be (significantly) less than the others. As with all methods, true positives (there was linkage and we detected it) compete with true negatives (there was no linkage and we did not find anything). The higher the percentage of true possitives, the lower the percentage of true negatives.

For all data sets the NPL- and likelihood-score are computed. This results in 3,000 linked scores and 3,000 null scores for the two methods. Depending on where the threshold between 'linkage' and 'no linkage' is set, there is a higher fraction of true positives or true negatives. This is shown in figure 2a for the dominant and in figure 2b for the recessive model. The statistical power of a particular score is expressed by how close the curve lies in the upper right corner. A rectangular shape is a perfect classifier, whereas a diagonal line from the upper left to the lower right corner is the worst one can get. When we set the model parameters to the ones used to create the artificial data, the so obtained likelihood score gives an idea what the maximum achievable curve looks like. This is the solid curve in both figures. It is clear that the likelihood curve is quite close to the maximum one. In both figures, there is a clear improvement compared to the NPL-score.

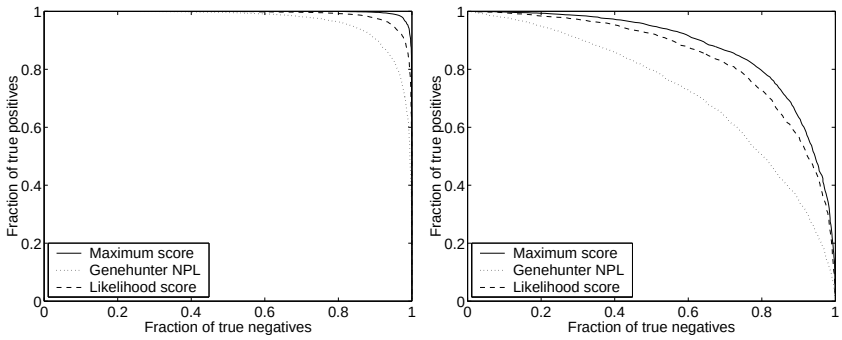

Fig. 2. The statistical power of the various methods can be seen in the plots. The solid line shows the maximum possible power, which is the likelihood score where the correct disease model is specified. The dotted line is Genehunter's NPL score, the dashed line is the non-parametric likelihood score defined in equation 1. In both cases, the latter had the most statistical power. Left is the dominant disease model, right the recessive one.

3.1 Significant cases for linkage

Although the previous subsection shows that we can expect an improved statistical power on the whole range, in practical situations one usually is only interested in the right most part of figure 2. This is the region where we can expect with a very high certainty that every positive sample found is really a case of linkage. In other words: the probability that a case labeled as 'linkage' is incorrect, the p-value, is very small.

We can not find the threshold needed for such small p-values by simply counting scores in the null data base, since this would require at least about $1/p$ samples to estimate the threshold reliably enough. For small p-values, this number can be unfeasibly high. Therefore, we use our 3,000 null samples to

	Dominant	
p-value	Likelihood	Genehunter
$< 10^{-2}$	2449 (82%)	1594 (53%)
$< 10^{-3}$	1713 (57%)	632 (21%)
$< 10^{-4}$	990 (33%)	153 (5.1%)
$< 10^{-5}$	495 (17%)	27 (0.9%)
$< 10^{-6}$	188 (6.3%)	3 (0.1%)
$< 10^{-7}$	55 (1.8%)	0 (0%)

	Recessive	
p-value	Likelihood	Genehunter
$< 10^{-2}$	661 (22%)	171 (5.7%)
$< 10^{-3}$	282 (9.4%)	21 (0.7%)
$< 10^{-4}$	114 (3.8%)	2 (0.1%)
$< 10^{-5}$	39 (1.3%)	1 (0.0%)
$< 10^{-6}$	18 (0.6%)	1 (0.0%)

Table 1. The number of linked cases out of 3,000 which could be found with a associated p-value less than indicated. The left table is the dominant, the right one the recessive case. The numbers can be read as: 'Given a disease with the dominant properties and a data set as described, we have a probability of 33% to detect linkage if it is present with a p-value less than 10^{-4} (and 5% for Genehunter).'

get an estimate of the tail of the null distribution and make a linear fit on the logarithm of the probability. Numerical studies (not shown here) indicate that this linear approximation is a perfect match within error boundaries.

Table 1 shows how many linked cases out of the 3,000 could be detected significantly for several p-values. This is compared to Genehunter, which assumes a Gaussian null distribution for the NPL score as reported by the program. It is clear from the table that our non-parametric likelihood finds far more significant cases in the linked data set than Genehunter does. We conclude that also in the practical region of small p-values the non-parametric likelihood is the best score one can use.

4 Future research

One of the major advantages of describing the problem in terms of a Bayesian network, is the trivial extension to diseases caused by more than one locus; an aspect of linkage analysis which is more and more desired, but still in its early stages of development. In most approaches, one simply performs a single locus scan and hopes that all loci have enough effect on their own to result in a signicant score. Very few approaches were recently made to treat two loci simultaneously without neglecting the correlation between them, thus leading to more powerful tests. But these approaches are still based on comparisons between pairs of individuals instead of taking the whole pedigree.

In the Bayesian framework, one simply copies a large part of the network shown in figure 1. One copy is pointing to locus i and one to locus j. The nodes determining the affection status, A, however, are shared by both networks. Obviously, the disease model, specified by the parameters f, is more complicated now, since the conditional probability for the affection status is given by $p(A|G_iG_jf)$. In this case, a completely free to choose f contains $3^2 = 9$ parameters. In these cases, it is useful to think whether certain priors on f are appropriate. For instance, AND- or OR-like mechanisms. Also known statistics (such as the a priori probability to have the disease or the probability to have the disease given that some relative is affected) can be used to restrict the number of free parameters. These numbers are usually based on a large population (e.g. all inhabitants in a country) such that they are known very precisely.

References

1. L. Kruglyak, M.J. Daly, M.P Reeve-Daly, and E.S. Lander. Parametric and nonparametric linkage analysis: A unified multipoints approach. *American Journal of Human Genetics*, 58:1347–1363, 1996.
2. N. Friedman, D. Geiger, and N. Lotner. Likelihood computations using value abstraction. In *Proc. Sixteenth Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2000.
3. R.C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Human Heredity*, 21:523–542, 1971.