A survey of data mining methods for linkage disequilibrium mapping

Päivi Onkamo[1] and Hannu Toivonen[2]*

[1] Department of Biological and Environmental Sciences, FI-00014, University of Helsinki, Finland
[2] HIIT-BRU, Department of Computer Science, FI-00014, University of Helsinki, Finland
* *Correspondence to*: Tel: +358 9 191 59111; Fax: +358 9 191 58754; E-mail: paivi.onkamo@helsinki.fi

Date received (in revised form): 22nd December 2005

Abstract

Data mining methods are gaining more interest as potential tools in mapping and identification of complex disease loci. The methods are well suited to large numbers of genetic marker loci produced by high-throughput laboratory analyses, but also might be useful for clarifying the phenotype definitions prior to more traditional mapping analyses. Here, the current data mining-based methods for linkage disequilibrium mapping and phenotype analyses are reviewed.

Keywords: *data mining, linkage disequilibrium mapping, association analysis*

Introduction

During recent years, there has been growing interest in using data mining methods in gene mapping, motivated by the lack of success of the more traditional approaches for complex diseases, and also by the intriguing possibility of simultaneous detection of multiple loci.[1,2] Although a wide spectrum of computational approaches is used for data mining, they tend to share certain attractive characteristics for genetic association analysis.

First, the methods are usually computationally efficient and scale to high numbers of markers and individuals, such as those expected in the near future in genome-wide association scans. Obviously, this efficiency comes with a price: the models considered tend to be simpler than those usually used in statistical genetics.

Secondly, data mining methods are often aimed at exploration or discovery — for example, by generating plausible models (or hypotheses) for further analysis rather than considering one given model in great detail. This aim coincides with a general trend in data analysis to move from hypothesis-driven- to hypothesis-generating research. The results of such exploration must often be complemented with more traditional statistical analysis.

Thirdly, data mining methods typically handle discrete data and use symbolic structures, giving results and explanations that may be easier to understand and utilise for users but are less suitable for statistical analysis.

'Data mining' is often loosely defined as 'non-trivial extraction of implicit, previously unknown and potentially useful information from data'. For this review of data mining methods for linkage disequilibrium (LD) mapping, the authors have chosen, at their own discretion, methods which reflect the three above-mentioned characteristics.

The data mining approaches of this review can be roughly categorised into three groups: (1) classification methods that directly aim to find markers and other features that help to predict the disease status; (2) clustering techniques for finding subgroups of subjects, based on their genotypic and phenotypic similarity, and analysis of their disease association; and (3) methods based on the discovery of typical haplotypes (or haplotype patterns) and analysis of their associations with the disease (Table 1).

In addition to gene mapping, data mining approaches have been applied to related areas, such as disease-susceptibility gene identification using literature databases.[3,4] This review focuses on LD mapping approaches. For more complete coverage, work from workshop proceedings and unpublished articles are also included. Some of the methods are available as software, for these a web page or e-mail address is provided.

Classification methods

Classification methods aim at finding rules or regularities that predict the value of a target variable from the independent variables. When applied to gene mapping, the goal is to find markers or haplotypes (and potentially other variables) that together are good predictors of the phenotype and then, more as a side-effect, predict a disease-susceptibility gene to be close to these markers. Regression analysis is a well-known

Table 1. Main classes of data mining approaches to gene mapping, characterised by three criteria: 1) Descriptive methods primarily aim to recognise the ancestral, shared chromosomal segments identical by descent, whereas predictive methods directly associate with the disease status. 2) Some approaches try to partition the set of subjects into homgeneous groups, some emphasise local similarities in haplotypes, and some are compromises between these extremes. 3) The suitability for describing and computing interactions varies between approaches.

Approach	Methods		Characteristics	
Classification	RP,[5,6] SDA,[7] DICE,[10] MDR,[11] SVMs[13], Association rules[9]	Predictive	Haplotype and subject-oriented	Models interactions
Haplotype clustering	HapMiner,[16] CLADHC,[17] Spatial clustering[18]	Descriptive	Haplotype and subject-oriented	No interactions
Phenotype clustering	MCA[19]	Predictive	Subject-oriented	No interactions
Haplotype patterns	HPM[20] and derivatives,[21,22,26–28] TreeDT[30]	Descriptive	Haplotype-oriented	Can model few interactions

prediction method for quantitative traits; this review focuses on classification methods for categorical traits.

Recursive partitioning (RP) methods (also known as decision/classification/regression trees) have been used for this purpose — for example by Young and Ge[5] and Cook *et al.*[6] RP produces a tree which can be described as a series of carefully crafted questions about the attributes of the test record, where each question splits the data into two parts and the next question is always conditional on the previous one(s). The gene finding method is, consequently, conditional: once a split is made based upon a single gene (or marker or haplotype), then the subsequent analysis is conditional on the results of that split — which is a very natural assumption for genetic effects. Young and Ge[5] present a successful application of RP, carried out with HelixTree® (www.goldenhelix.com) on simulated clinical trial data, where the aim is to find (out of 80 genetic polymorphisms) those markers that have the highest impact on the efficacy and safety of a blood pressure medication.

Symbolic discriminant analysis (SDA) was utilised in integrated analysis of multiple data types (genetic markers, genomic and proteomic data) in a review by Reif *et al.* in 2004.[7] SDA is a supervised pattern-mining approach that carries out variable selection and model selection simultaneously and automatically. SDA builds discriminate functions from a list of mathematical operators (eg $+$, $-$, x, /) and explanatory variables that can distinguish between disease classes in the data. In an integrative analysis of simulated multiple-type data, the authors showed that, in particular, when the aetiology of the disease is complex, the integrated analysis can be highly advantageous. The SDA approach implemented by Reif *et al.* can be obtained from jason.h.moore@dartmouth.edu.

Association rules have been applied to genetic problems — for example, by Rova *et al.*[8] in a candidate gene analysis for bronchopulmonary dysplasia in newborns, where a number of non-genetic risk factors had also been measured and best combinations of covariates and genetic markers were sought. Association rules describe co-occurrences of sets of features and can be computed very efficiently. In this case, the presence of two different, but sometimes co-occurring, syndromes were set as targets and the significance of association of conjunctions of several genetic and non-genetic risk factors to either syndrome was measured from the association rules. Two separate polymorphisms were proposed to have a phenotypic effect via separate molecular mechanisms.[8] Although the implementation used by this group is not available, a general purpose Apriori algorithm for finding association rules is given by Agrawal *et al.*;[9] freely available implementations are numerous (eg http://www.kdnuggets.com/software/).

The DICE algorithm[10] identifies a subset of genetic and non-genetic covariates that are, either individually or in combination, associated with a phenotype. The relationship between the phenotype and the covariates is modelled using a logistic, linear or Cox regression model. The algorithm explores, by means of a forward procedure, a set of competing models and selects the most parsimonious and informative approximating model(s) that minimise(s) the information criterion. Thus, the method combines the advantages of regressive approaches in terms of modelling and interpretation of effects with those of data exploration tools. It should be well suited to detecting interactions between genetic and non-genetic factors within the framework of association studies. DICE has been successfully applied to a handful of datasets[10] (http://ecgene.net/genecanvas/modules/news/article.php?storyid = 7). DICE is available upon request from laurence.tiret@chups.jussieu.fr.

Multifactor dimensionality reduction (MDR)[11] is a non-parametric approach to detecting and characterising non-linear interactions among discrete genetic and environmental attributes. Multilocus genotypes are pooled into high-risk and low-risk groups, reducing the numbers of genotype predictors. The reduced-dimension variable is used to classify and predict disease status through cross-validation and permutation testing. MDR has been shown to be capable of revealing significant high-order interactions in real datasets[12] (http://www.epistasis.org/mdr.html).

Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyperplane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features. Waddell *et al.*[13] used SVM to predict the age at diagnosis of multiple myeloma, based on 3,000 single nucleotide polymorphisms (SNPs) genotyped in 40 young age-at-onset and 40 old age-at-onset patients. Although the authors do not refer to their method as being LD mapping, their search for best predictor SNPs for the trait is based on the hypothesis that if there is a genetic factor to the trait, then a SNP in the haplotype block (ie in strong LD) containing that gene will be discovered. In fact, the trained SVM produced a model with a reasonable accuracy (71 per cent by cross-validation), but the model was not easily interpretable: it consisted of 150 SNPs. A general-purpose SVM algorithm, SVM^{light}, is publicly available at http://svmlight.joachims.org.

Clustering

Clustering aims to locate relatively homogeneous subgroups in the given data. In the context of LD mapping, clustering of study subjects has been suggested as an approach for finding subgroups of individuals who potentially share genetic factors. Such clustering can be based on haplotypes of the individuals, or on their phenotypes. After successful clustering, it should be easier to locate the genetic factors within the clusters, improving statistical power; however, power may be reduced if the effective sample size decreases.

A crucial factor here seems to be that genetically motivated similarity measures are used, based on haplotype sharing between individuals. 'Length measure' — the length spanned by the longest continuous interval of matching alleles — is one typical option — and 'count measure' — the number of alleles in common in a window — is another.[14,15] With such measures, the clustering method is directed to search for clusters with shared genetic aetiology. The association of clusters to the phenotype can then be measured — for example, using the χ^2 statistic, and the disease gene can be predicted to be where the best cluster shows similarity of haplotypes. An instance of this approach is implemented in the HapMiner[16] software (http://vorlon.cwru.edu/~jxl175/HapMiner.html). Durrant *et al.*[17] use hierarchical clustering to produce approximations of genealogical trees and map genes based on these trees. The method has been coded in the CLADHC algorithm, available as a linux executable, with accompanying documentation, on request from amorris@well.ox.ac.uk. Molitor *et al.*[18] perform fine mapping by spatial clustering of haplotypes based on a similarity metric that measures the length of the shared region and by estimating the risk that each haplotype 'cluster' has for the trait. The implementation of the method is available from jmolitor@usc.edu.

A good example of a slightly different approach to LD mapping, also based on measuring haplotype similarities but not on clustering, is given by Tzeng *et al.*[14] They investigated the hypothesis that the average similarity between case haplotypes tends to be higher than between control haplotypes. Under this assumption, disease-susceptibility genes can be localised directly by measuring the statistical significance of haplotype similarity in the cases without explicit clustering or goodness of fit tests, such as χ^2. The authors concluded that similarity measures are actually more powerful than goodness of fit tests when the mutation occurs on a common haplotype, but that goodness of fit tests are superior for rare haplotypes. Haplotype similarity and clustering were proposed as exploratory haplotype analysis methods by Toivonen *et al.*[15]

The other major variant of the clustering theme is to cluster subjects by their phenotypes, rather than haplotypes. Again, the aim is to find (phenotypic) subgroups that potentially have more homogeneous genetic aetiologies, but now utilising rich phenotypic datasets, where they exist. Wilcox *et al.*[19] clustered subjects from the Framingham Heart Study for this purpose. Different phenotypic measurements can have very different ranges and distributions, and these have to be handled to avoid unintended bias. Wilcox and others used multiple correspondence analysis (MCA), a non-parametric analogue of principal component analysis, to produce a reduced number of dimensions in which clustering was then performed. They subsequently used linkage analysis for mapping; there do not appear to be any publications on phenotype clustering for LD mapping, even though the approach should be equally feasible there.

Discovery of frequent patterns

The most popular data mining method applied to gene mapping has been the discovery of typical haplotypes (or haplotype patterns) and analysis of their associations with the disease. In simple terms the goal is firstly to discover sites and haplotypes potentially identical by descent, and then to test their disease associations.

Haplotype pattern mining (HPM) was the first such method (http://www.cs.helsinki.fi/group/genetics/).[20] The algorithm finds all haplotype fragments (patterns) of arbitrary length — possibly up to some limit and possibly with gaps — that show statistical association with the disease. The set of associated fragments is used as a whole to evaluate association across the chromosomal area studied. The area that shows the most significantly elevated number of patterns is the most likely for a disease-susceptibility locus.

The significance of the finding is evaluated by permutation tests, where both marker-wise nominal significances — as well as a corrected significance for the best finding — are computed.[21,22]

The HPM method is fast, especially with respect to the number of markers, and it is sensitive to small genetic effects. The results are rough, however, and more elaborate (and computationally more expensive) statistical models are expected to predict the disease mutation locale better than HPM. In conclusion, HPM seems to work excellently as the first-stage analysis tool of genome-wide association and has been successfully applied in various circumstances — for example, for asthma-related traits,[23] glucocorticoid sensitivity[24] and familial glioma.[25]

Variants of HPM include a method for finding two (interacting) loci at the same time[26] and QHPM for analysis of quantitative traits.[27] F-HPM developed by Zhang *et al.*[28] is a further development of HPM, in which the strength of association is tested in pedigrees using the quantitative pedigree disequilibrium test.[29]

The tree disequilibrium test (TreeDT; http://www.cs.helsinki.fi/group/genetics/)[30] is a more elaborate attempt to model the unknown coalescence, rather than just haplotyping fragments potentially identical by descent. TreeDT constructs, at each locus, trees that approximate the genealogy of the haplotypes at that locus, much like the method of Durrant *et al.*[17] These trees can be obtained efficiently using known algorithms for strings, making the method computationally fast. After trees are built for all locations, a disequilibrium test is performed on each of them to test if there is a small set of subtrees with relatively high proportions of disease-associated chromosomes, suggesting shared genetic history for those and a likely disease-gene location. Again, permutation tests are used to measure significances. TreeDT is fast and has been shown to be relatively accurate, especially when allelic heterogeneity is present in a disease locus.

Conclusions

Notably, the methods presented here are mostly intended for exploratory analysis and not so much for final stages of identifying a causative variant in genotype data. The user's expertise and insight play a key role: they are needed in choosing the methods and parameter values and are crucial in interpreting the results. Also, there is no universally optimal method for all purposes; it can be useful to try several different approaches for the same problem.

As pointed out by Hoh and Ott,[2] what is most needed for future large-scale genetic and genomic data analysis are 'methods for discovering sets of susceptibility genes and environmental factors, as well as systematic verifications of the gene–environment–disease network'. According to the present review, there already exist a number of data mining approaches to gene mapping or identification purposes (Table 1); however, they are still rather scattered, consisting of somewhat solitary attempts to use different machine-learning or data mining approaches.

Classification methods are typically strong in modelling interactions, unlike most other approaches in this review. Several of the classification methods produce a set of interacting loci that best predict the phenotype. However, a straightforward application of classification methods to large numbers of markers has a potential risk picking up randomly associated markers.

Approaches based on haplotype sharing, such as most of the reviewed clustering and pattern discovery methods, explicitly aim to reduce this problem by considering loci that are more likely to be identical by descent. Of course, combinations are possible; for instance, all frequent haplotype patterns could first be found and a classifier used to choose a subset of those and to model the interactions of their loci.

In the more distant future, one might expect to gain most from integrated large-scale analyses: data mining of high-throughput SNP data for LD mapping combined with phenotype subgroup analysis; expression analysis results — information about co-regulated enzymes in normal and trait-carrying individuals — integrated with the information on known metabolic pathways; and linking of the new experimental information to existing public data by mining literature and biological databases.

References

1. Hoh, J. and Ott, J. (2003), 'Mathematical multi-locus approaches to localizing complex human trait genes', *Nat. Rev. Genet.* Vol. 4, pp. 701–709.
2. Hoh, J. and Ott, J. (2004), 'Genetic dissection of diseases: Design and methods', *Curr. Opin. Genet. Dev.* Vol. 14, pp. 229–232.
3. Perez-Iratxeta, C., Bork, P. and Andrade, M.A. (2003), 'Association of genes to genetically inherited diseases using data mining', *Nat. Genet.* Vol. 31, pp. 316–319.
4. Perez-Iratxeta, C., Wjst, M., Bork, P. *et al.* (2005), 'G2D: A tool for mining genes associated with disease', *BMC Genet.* Vol. 6, p. 45.
5. Young, S.S. and Ge, N. (2005), 'Recursive partitioning analysis of complex disease pharmacogenetic studies I. Motivation and overview', *Pharmacogenetics* Vol. 6, pp. 65–75.
6. Cook, N.R., Zee, R.Y.L. and Ridker, P.M. (2004), 'Tree and spline based association analysis of gene-gene interaction models for ischemic stroke', *Stat. Med.* Vol. 23, pp. 1439–1453.
7. Reif, D.M., White, B.C. and Moore, J.H. (2004), 'Integrated analysis of genetic, genomic and proteomic data', *Expert Rev. Proteomics* Vol. 1, pp. 67–75.
8. Rova, M., Haataja, R., Marttila, R. *et al.* (2004), 'Data mining and multiparameter analysis of lung surfactant protein genes in bronchopulmonary dysplasia', *Hum. Mol. Genet.* Vol. 13, pp. 1095–1104.
9. Agrawal, R., Mannila, H., Srikant, H. *et al.* (1996), 'Fast discovery of association rules', in Fayyad, U.M., Piatetsky-Shapiro, G., Smy, P. *et al.* (eds), '*Advances in Knowledge Discovery and Data Mining*', AAAI Press/The MIT Press, Menlo Park, CA, pp. 307–328.
10. Tahri-Daizadeh, N., Tregouet, D.A., Nicaud, V. *et al.* (2003), 'Automated detection of informative combined effects in genetics association studies of complex traits', *Genome Res.* Vol. 13, pp. 1952–1960.

11. Ritchie, M.D., Hahn, L.W., Roodi, N. *et al.* (2001), 'Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer', *Am. J. Hum. Genet.* Vol. 69, pp. 138–147.
12. Moore, J.H. (2004), 'Computational analysis of gene–gene interactions using multifactor dimensionality reduction', *Expert Rev. Mol. Diagn.* Vol. 4, pp. 795–803.
13. Waddell, M., Page, D., Zhan, F. *et al.*, (2005), 'Predicting cancer susceptibility from single-nucleotide polymorphism data: A case study in multiple myeloma', in: '*Proceedings of the 5th ACM SIGKDD Workshop on Data Mining in Bioinformatics*', Chicago, IL.
14. Tzeng, J.Y., Devlin, B., Wasserman, L. *et al.* (2003), 'On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit', *Am. J. Hum. Genet.* Vol. 72, pp. 891–902.
15. Toivonen, H., Onkamo, P., Hintsanen, P. *et al.* (2005), 'Data mining for gene mapping', in Kantardzic, M.M. and Zurada, J. (eds), '*New Generation of Data Mining Applications*', IEEE Press, Hoboken, NJ, pp. 263–293.
16. Li, J. and Jiang, T. (2005), 'Haplotype-based linkage disequilibrium mapping via direct data mining', *Bioinformatics* Vol. 21, pp. 4384–4393.
17. Durrant, C., Zondervan, K.T., Cardon, L.R. *et al.* (2004), 'Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes', *Am. J. Hum. Genet.* Vol. 75, pp. 35–43.
18. Molitor, J., Marjoram, P. and Thomas, D. (2003), 'Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques', *Am. J. Hum. Genet.* Vol. 73, pp. 1368–1384.
19. Wilcox, M.A., Wyszynski, D.F., Panhuysen, C.I. *et al.* (2003), 'Empirically derived phenotypic subgroups — Qualitative and quantitative trait analyses', *BMC Genet.* Vol. 4(Suppl. 1), p. S15.
20. Toivonen, H.T.T., Onkamo, P., Vasko, K. *et al.* (2000), 'Data mining applied to linkage disequilibrium mapping', *Am. J. Hum. Genet.* Vol. 67, pp. 133–145.
21. Sevon, P., Toivonen, H.T.T. and Onkamo, P. (2005), 'Gene mapping by pattern discovery', in Wang, J., Zaki, M., Toivonen, H. *et al.* (eds), '*Data Mining in Bioinformatics*', Springer-Verlag, London, UK, pp. 105–126.
22. Sevon, P. (2004), 'Algorithms for association-based gene mapping'. Academic dissertation, Helsinki University Printing House, Helsinki, Finland.
23. Laitinen, T., Polvi, A., Rydman, P. *et al.* (2004), 'Characterization of a common susceptibility locus for asthma-related traits', *Science* Vol. 304, pp. 300–304.
24. Stevens, A., Ray, D.W., Zeggini, E. *et al.* (2004), 'Glucocorticoid sensitivity is determined by a specific glucocorticoid receptor haplotype', *J. Clin. Endocrinol. Metab.* Vol. 89, pp. 892–897.
25. Paunu, N., Lahermo, P., Onkamo, P. *et al.* (2002), 'A novel low-penetrance susceptibility locus for familial glioma at 15q23-q26.3', *Cancer Res.* Vol. 62, pp. 3798–3802.
26. Toivonen, H.T.T., Onkamo, P., Vasko, K. *et al.* (2000), 'Gene mapping by haplotype pattern mining', in: '*Proceedings of IEEE International Symposium on Bio-Informatics and Biomedical Engineering*', IEEE, Arlington, VA, USA, pp. 99–108.
27. Onkamo, P., Ollikainen, V., Sevon, P. *et al.* (2002), 'Association analysis for quantitative traits by data mining: QHPM', *Ann. Hum. Genet.* Vol. 66, pp. 419–429.
28. Zhang, S., Zhang, K., Li, J. *et al.* (2002), 'On a family-based haplotype pattern mining method for linkage disequilibrium mapping', in: Altman, R.B., Dunker, A.K., Hunter, L. *et al.* (eds), Pacific Symposium on Biocomputing, Vol. 7, World Scientific Press, Singapore, pp. 100–111.
29. Zhang, S., Zhang, K., Li, J. *et al.* (2001), 'Test of linkage and association for quantitative traits in general pedigree: The quantitative pedigree disequilibrium test', *Genet. Epidemiol.* Vol. 21(Suppl. 1), pp. S370–S375.
30. Sevon, P., Toivonen, H. and Ollikainen, V. (2006), 'TreeDT: Tree pattern mining for gene mapping', in '*IEEE/ACM Transactions on Computational Biology and Bioinformatics*' (in press).