FORESTS OF LATENT TREE MODELS FOR THE DETECTION OF GENETIC ASSOCIATIONS

Christine Sinoquet[1*], Raphaël Mourad[2*] and Philippe Leray[2]
[*] *Joint first authors*
[1]LINA, UMR CNRS *6241, Université de Nantes, 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex, France*
[2]LINA, UMR CNRS *6241, Ecole Polytechnique de l'Université de Nantes*
rue Christian Pauc, BP 50609, 44306 Nantes Cedex 3, France

Keywords: Probabilistic graphical model, Bayesian network, Latent tree model, Detection of genetic association, Latent variable, Data dimension reduction.

Abstract: Together with the population aging concern, increasing health care costs require understanding the causal basis for common genetic diseases. The high dimensionality and complexity of genetic data hamper the detection of genetic associations. To alleviate the core risks (missing of the causal factor, spurious discoveries), machine learning offers an appealing alternative framework to standard statistical approaches. A novel class of probabilistic graphical models has recently been proposed - the forest of latent tree models - , to obtain a trade-off between faithful modeling of data dependences and tractability. In this paper, we evaluate the soundness of this modeling approach in an association genetics context. We have performed intensive tests, in various controlled conditions, on realistic simulated data. We have also tested the model on real data. Beside guaranteeing data dimension reduction through latent variables, the model is empirically proven able to capture indirect genetic associations with the disease, both on simulated and real data. Strong associations are evidenced between the disease and the ancestor nodes of the causal genetic marker node, in the forest. In contrast, very weak associations are obtained for other nodes.

1 INTRODUCTION

Thanks to their ability to capture (conditional) independences and dependences between variables, probabilistic graphical models (PGMs) offer an adapted framework for a fine modeling of relationships between variables in an uncertain data framework. A PGM is a probabilistic model relying on a graph encoding conditional dependences within a set of random variables. A PGM provides a compact and natural representation of the joint distribution of the set of variables. Bayesian networks (BNs) are a commonly used branch of PGMs.

Despite the fact that the observed variables are often sufficient to describe their joint distribution, sometimes, additional unobserved variables, also named latent variables, have a role to play. In this context, hierarchical Bayesian networks such as latent tree models (LTMs), formerly named hierarchical latent class models, were proposed. LTMs are tree-shaped BNs where leaf nodes are observed while internal nodes are not. LTMs generalize latent class models (LCMs), defined as containing a unique latent variable and edges only connecting the latent variable to all the observed variables. In LTMs, multiple latent variables organized in a hierarchical structure allow to depict a large variety of relations encompassing local to higher-order dependences (see Figure 1). LCMs enforce observed variables to be independent, conditionally on the latent variable. In contrast, LTMs relax this local independence assumption which is often violated for observed data.

Few algorithms have been developed to learn such models and still fewer for applications in association genetics (Zhang and Ji, 2009). Forests of LTMs have been recently proposed as *potentially* useful for association studies (Mourad et al., 2010; Mourad et al., 2011). In the biomedical research domain, association studies rely on the description of DNA variants at characterized genome loci - or genetic markers - for all subjects in case and control cohorts. Such studies attempt to identify any putative dependence - or association - between one or possibly some genetic markers and the affected/unaffected status. In the case of a

Sinoquet C., Mourad R. and Leray P..
FORESTS OF LATENT TREE MODELS FOR THE DETECTION OF GENETIC ASSOCIATIONS.
DOI: 10.5220/0003703400050014
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2012), pages 5-14
ISBN: 978-989-8425-90-4

single causal locus, a putative association is revealed if the distribution of variants between cases and controls shows an accumulation of the former with respect to some variant(s). From now on, we will refer to the most popular genetic markers, that is, Single Nucleotide Polymorphisms (SNPs).

One of the first motivations to propose this novel model - the forest of LTMs (FLTM) - is to take account of linkage disequilibrium (LD) in the most possible faithful way. Linkage disequilibrium occurs because DNA variants close on the chromosome are scarcely separated by the shuffling of chromosomes (recombination) that takes place during sex cell formation. Such variants are therefore transmitted together (as an haplotype) from parent to child. Such patterns are at the basis of the so-called haplotype block structure (Daly et al., 2001): "blocks" where statistical dependences between loci are high alternate with shorter regions characterized by low statistical dependences, the recombination hotspots. LD is crucial for association studies since a causal locus not sharply coinciding with a SNP is nevertheless expected to be flanked by SNPs highly likely to be shown (indirectly) associated with the phenotype. Besides, benefitting from high correlations is appealing to implement data dimension reduction.

Data dimension reduction exploiting LD is not new to genetics. However, tackling this issue through adapted Bayesian networks has but recently been proposed (Mourad et al., 2011). Notably, these authors have successfully compared the FLTM-based method to other methods with respect to faithfulness in LD modeling and data dimension reduction. Besides, FLTM models seem appealing to enhance association studies: due to their hierarchical structure, FLTM models would help pointing out a region containing a genetic factor associated with a studied disease. However, bottom-up information fading is likely to be observed in the hierarchical structure. The impact on downstream analyses such as association studies remains questionable. The very point is to check whether latent variables covering a causal region are found associated with the disease. In this paper, we have conducted a systematic and comprehensive evaluation of the ability of the FLTM model to help evidence genetic associations through latent variables.

This paper is organized as follows: after the Section "Definitions", the motivation for the FLTM model proposal is provided, together with the context of this proposal. The next Section describes the FLTM learning algorithm used in this paper and highlights the differences with the initial version. In Section "Study Protocol", we define the notion of "indirect association"; then we detail the protocol implemented to evaluate the ability of FLTM's latent variables to capture indirect associations. The Section Results and Discussion describes and discusses intensive tests on realistic simulated data and real genotypic data.

2 DEFINITIONS

BNs are defined by a directed acyclic graph $G(X,E)$ and a set of parameters θ. The set of nodes $X = \{X_1,...,X_p\}$ represents p random variables and the set of edges E captures the conditional dependences between these variables (*i.e.* the structure). The set of parameters θ describes a conditional probability distributions $\theta_i = [\mathbb{P}(X_i/Pa_{X_i})]$ where Pa_{X_i} denotes node i's parents. If a node has no parent, then it is described by an *a priori* probability distribution. The variables are described for n observations. For further understanding, we now briefly recall some definitions.

Definition 1 (conditional independence)**.** *Given a subset of variables $S \subseteq X\backslash\{X_i,X_j\}$, conditional independence between X_i and X_j is defined as: $\mathbb{P}(X_i,X_j|S) = \mathbb{P}(X_i|S)\ \mathbb{P}(X_j|S)$. The non-equality entails that both variables are conditionally dependent given S.*

Definition 2 (entropy, mutual information)**.** *The entropy of variable X writes as: $\mathcal{H}(X) = -\sum_{i=1}^{n}\mathbb{P}(x_i)\log\mathbb{P}(x_i)$ where $\mathbb{P}(x_i)$ is the probability mass function of outcome x_i.*
Given two variables X_1 and X_2, the mutual information measures the dependence of the two variables, expressing the difference of entropies between the independent model $\mathbb{P}(X_1)\ \mathbb{P}(X_2)$ and the dependent model $\mathbb{P}(X_1|X_2)\ \mathbb{P}(X_2)$: $I(X_1,X_2) = \Big(\mathcal{H}(X_1) + \mathcal{H}(X_2)\Big) - \Big(\mathcal{H}(X_1|X_2) + \mathcal{H}(X_2)\Big) = \mathcal{H}(X_1) - \mathcal{H}(X_1|X_2)$. The larger the difference between entropies, the higher is the dependence.

Due to the presence of pairs of chromosomes in the human genome, the DNA at a given chromosome locus (SNP) may either be described through a pair of variants (alleles or phased data) at the finer description level or through a unique variant (unphased data). As SNPs are *biallelic*, only two alleles are encountered at the corresponding loci (instead of the 4 possible nucleotides A,T,C,G). Thus, SNPs are discrete variables whose three possible values may be coded as, say 0, 1 and 2, to respectively account for aa, $\{Aa,\ aA\}$ (usually not distinguishable) and AA, where A and a are the two alleles. In the context of this paper, we restrain our concern to discrete and finite variables (either observed or latent). In this work, we address the case of the single causal genetic factor.

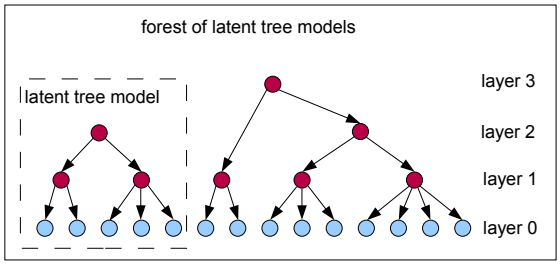


Figure 1: Latent tree model and forest of latent tree models. The light shade indicates the observed variables whereas the dark shade points out the latent variables.

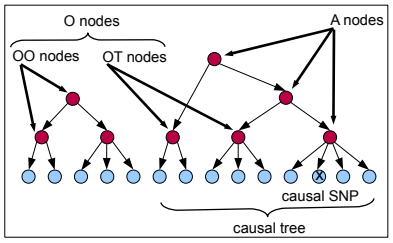


Figure 2: Illustration of key terms specific to our approach. A nodes: ancestor nodes of the causal SNP; O nodes: other latent nodes categorized in OT nodes (in causal tree) and OO nodes (outside the causal tree). See Figure 1 for node nomenclature.

3 MOTIVATION AND RELATED WORK

3.1 Motivation

To tackle the difficult problem of disease association detection, several algorithms coming from the machine learning domain have been proposed. Some of them use PGMs (Verzilli et al., 2006; Han et al., 2010). Recently, forests of latent tree models have been investigated for LD modeling purpose (Mourad et al., 2011). A forest of latent tree models (FLTM) is a forest whose trees are LTMs (see Figure 1). FLTMs generalize LTMs, since the variables are not constrained to be dependent upon one another, either directly or indirectly. Thus, FLTMs can describe a larger set of configurations than LTMs.

When modeling such highly correlated variables as those in genotypic data, the challenge is all the more crucial for downstream analyses such as study and visualization of linkage disequilibrium, mapping of disease susceptibility genetic patterns and study of population structure. Most notably, the benefits of using FLTMs to model LD rely on their ability to account for multiple degrees of SNP dependences and to naturally deal with the fuzzy nature of LD block boundaries. As we will further emphasize, this latter advantage results from the FLTM learning algorithm, which does not impose that the SNPs subsumed by the same latent variable be neighbouring SNPs (along the genome).

3.2 Related Work

As for general BNs, besides learning of parameters (θ), *i.e. a priori* and conditional probabilities, one of the tasks in LTM learning is structure inference. This task generally remains the most challenging due to the complexity of the search space. Regarding LTM learning, the proposals published in the literature fall into two categories. The first category relies on standard Bayesian network learning techniques. The second category is based on the clustering of the variables. To learn θ, both categories rely on the expectation maximization (EM) algorithm or an EM variant.

In the first category, the algorithms explore the search space through a local search strategy and optimize a score, such as the BIC score (Schwartz, 1978). Zhang proposed a greedy algorithm, to navigate in the structure search space (Zhang, 2004). This algorithm is coupled with a hill climbing procedure, to adjust the cardinality of the latent variables. A more efficient variant of this algorithm has recently been proposed (Chen et al., 2011). In this first category, structural expectation maximization (SEM) has also been adapted to the case of Bayesian networks with latent variables. SEM successively optimizes θ, conditionally on the structure $\mathcal{S}$, then optimizes $\mathcal{S}$ conditionally on θ. Parameter learning being a time-consuming step, in this framework, Zhang and Kocka have adapted a procedure, called local EM, to optimize the variables whose connexion or cardinality have been modified in the transition from former to current model (Zhang and Kocka, 2004). In one of the two LTM-based approaches dedicated to LD modeling, Zhang and Ji use a set of LCMs and apply a SEM strategy (Zhang and Ji, 2009). The number of LCMs has to be specified. To avoid getting trapped in local optima while running the EM algorithm to learn a set of latent models, these authors have adapted a simulated annealing approach.

The above score-based approaches require the computation of the maximum likelihood in presence of latent variables, a prohibitive task regarding computational burden. Thus, various methods based on the clustering of variables have been implemented. They all construct the model following an ascending strategy; they all rely on the mutual information (MI) criterion, to identify clusters of dependent variables. In their turn, these methods may be sub-categorized into binary- and non binary-based approaches.

Hwang and co-workers' learning algorithm is dedicated to binary trees and binary latent variables (Hwang et al., 2006). It has to be noted that the trees are possibly augmented with connexions between siblings, that is nodes sharing the same parent into the immediate upper layer. Also confining themselves to binary trees, Harmeling and Williams have proposed two learning algorithms (Harmeling and Williams, 2011). One of them approximates MI between a latent variable H and any other variable X, based on a linkage criterion (single, complete or average) applied for X and the variables in the cluster subsumed by H. A variant of this first algorithm locally infers the data corresponding to any latent variable; therefore it is possible to achieve an exact computation of the MI criterion between a latent variable and any other variable.

Two approaches have been proposed to circumvent binary tree-based structures. Wang and coworkers first build a binary tree; then they apply regularization and simplification transformations which may result in subsuming more than two nodes through a latent variable (Wang et al., 2008). In their approach devoted to LD modeling, Mourad and collaborators implement the clustering of variables through a partitioning algorithm (Mourad et al., 2011); the latter yields cliques of pairwise dependent variables. Besides, this method imposes the control of information fading as the level increases in the hierarchy, which generally results in the production of a forest of LTMs (instead of a single LTM).

Two of the above cited methods have been shown to be tractable. For these two non binary-based approaches, some reports are available: Hwang and coworkers' approach was able to handle 6000 variables and around 60 observations. The scalability of the FLTM construction by Mourad *et al.* has been shown for benchmarks describing 10^5 variables and 2000 individuals.

4 THE LEARNING ALGORITHM

We now describe the algorithm we have used to test the FLTM model in the context of association genetics. We have adapted the initial version of Mourad and co-workers. We will highlight the differences between the two implementations.

4.1 Sketch of the Algorithm

The learning is performed through an adapted agglomerative hierarchical clustering procedure. At each iteration, a partitioning method is used to assign variables into non-overlapping clusters. The partitioning is based on the identification of cliques of strongly dependent variables in the complete graph of pairwise dependences. Amongst the clusters, each cluster of size at least two is a candidate for subsumption into a latent variable H. To acknowledge or reject the creation of H, a prior task considers the LCM rooted in this latent variable candidate and whose leaves are all the variables of the cluster. Parameter learning using the EM algorithm is performed for this LCM. Then probabilistic inference allows missing data imputation for the latent variable. Once all the data are known for this LCM, a validation step checks whether the latent variable captures enough information from its children. If a latent variable is validated, its child variables are then replaced with the latent variable. In contrast, the nodes in unvalidated clusters are kept isolated for the next iteration. Iterating these steps yields a hierarchical structure. In other words, latent variables capture the information borne by underlying observed variables (*e.g.* genetic markers). In their turn, these latent variables, now playing the role of observed variables, are synthesized through additional latent variables, and so on.

For a better understanding, we now detail five points of this algorithm. We start with the two points establishing the difference between the initial version in (Mourad et al., 2011) and our novel version.

4.1.1 Window-based Data Scan versus Straightforward Data Scan

First, we remind the reader that the initial observed variables are SNPs, which are located along the genome in a sequence of "neighbouring" (but generally non contiguous) genetic markers. To meet the scalability criterion, a divide-and-conquer procedure has been implemented in (Mourad et al., 2011): the data is scanned through contiguous windows of identical fixed sizes. However, such splitting is questionable. It entails a bias in the processing of the variables located in the neighbourhood of the artificial window frontiers. Managing overlapping windows would not have lead to a practicable algorithm. Therefore, a first notable difference with the algorithm in (Mourad et al., 2011) lies in that our novel version does not require data splitting. Instead, a simple principle is implemented: not all pairs of variables are processed by the partitioning algorithm. Beyond a physical distance on the chromosome, δ, specified by the geneticist, variables are not allowed in the same cluster. Setting the δ constraint actually corresponds to implementing a sliding window approach.

4.1.2 Partitioning of Variables into Cliques

Standard agglomerative hierarchical clustering considers a similarity matrix. As a latent variable is intended to connect pairwise dependent variables, the standard agglomerative approach was adapted accordingly. Within each window, our previous version run a clique partitioning algorithm on the complete graph of pairwise dependences. In the novel version, no complete matrix is required anymore. The physical constraint δ leads to calculate a sparse matrix of pairwise dependences, where only computed values are stored.

In our former version, we used the clique partitioning algorithm CAST devoted to the clustering of variables (Ben-Dor et al., 1999). The dependence between two variables, evaluated through pairwise mutual information, is used to derive a binary similarity measure (requested by CAST), depending on a threshold $\tau_{pairwise}$. Our algorithm automatically adjusts this threshold, based on a given quantile value of the mutual information values in the whole matrix (*e.g.* the median value). We have rewritten the CAST algorithm to take account of the physical constraint δ. Through the management of a sparse matrix of pairwise dependences, we actually allow the modulation of the useful matrix bandwidth, depending on the physical constraint imposed by the sliding window size δ.

However, unlike SNPs, latent variables are not characterized by a physical location on the chromosome. In this specific case, we average the locations of the SNPs subsumed by the latent variable.

4.1.3 Data Imputation for Latent Variables

Data imputation is processed locally, that is considering the LCM rooted in the latent variable and whose leaves are the variables in the cluster. For simplification, the cardinality of the latent variable is estimated as an affine function of the number of leaves. Parameter learning is first performed in this LCM, through the EM algorithm. This step yields the marginal distribution of the latent variable and the conditional distributions of the child variables. Therefore, (linear) probabilistic inference can be carried on, based on the following principle:

$$\mathbb{P}(H=c|\mathbf{x}^j)=\frac{\Pi_{i=1}^{p}\,\mathbb{P}(x_i^j|H=c)\,\mathbb{P}(H=c)}{\sum_{c=1}^{k}\,\Pi_{i=1}^{p}\,\mathbb{P}(x_i^j|H=c)\,\mathbb{P}(H=c)},$$

with k the cardinality of latent variable H, c a possible value for H, j an observation, *i.e.* an individual in our case, and $\mathbf{x}^j$ the vector of values $\{x_1^j,...,x_p^j\}$ corresponding to the variables in the cluster $\{X_1,...,X_p\}$.

4.1.4 Local Parameter Learning

In parallel with the structure growing, the parameters of the forest of LTMs are learned locally (see Subsection 4.1.3). At a given iteration, for any variable shown to be a leaf node in an LCM (corresponding to a cluster), the current marginal distribution is replaced with the conditional distribution learned in the LCM. Thus, during the bottom-up construction of the FLTM, marginal distributions are successively replaced with conditional distributions.

4.1.5 Validation of Latent Variables

The subsumption of the candidate cluster into the latent variable H is validated through a criterion averaging a normalized dependence measure between H and each of H's child nodes:

$$Criter=\frac{1}{\mid C_H \mid}\sum_{i\,\in\, C_H}\frac{I(X_i,H)}{\min\,(\mathcal{H}(X_i),\ \mathcal{H}(H))}\geq\tau_{latent},$$

with $\mid C_H \mid$ the size of cluster C_H.

4.2 Recapitulation

In the forest of LTMs, the subsumption process is controlled through thresholds $\tau_{pairwise}$ and τ_{latent} and constraint δ. No latent variable is allowed to subsume variables which are not highly pairwise dependent ($\tau_{pairwise}$) or which refer to regions which are too far from one another (δ); τ_{latent} controls bottom-up information fading through the hierarchy. $\tau_{pairwise}$, τ_{latent} and δ thus monitor the number of connected components (trees) and the number of layers in the forest. These three parameters rule the trade-off between faithfulness to the underlying reality and tractability of the modeling.

We have tested the behaviour of our algorithm - in particular its handling of large sparse matrices - on datasets describing 10^5 SNPs for 2000 individuals. In (Mourad et al., 2011), the running time was around 15 hours for an arbitrary window size of 100 SNPs. When setting the sliding window size δ to 0.5 *Mb*, a reasonable choice to capture LD, our novel algorithm now runs in less than 12 hours. It has to be emphasized that as our algorithm runs EM with 10 restarts, a significant improvement has been brought with respect to the initial version. Finally, we have checked that our algorithm is quasi linear with the number of SNPs and linear with the sliding window size. The corresponding experimentations are not shown in this paper. Neither are our examinations of the robustness with respect to parameter adjustment. The application software is available at http://sites.google.com/site/raphaelmourad/Home/programmes.

5 STUDY PROTOCOL

Our purpose in this paper is to investigate how information about causality fades from bottom to top in the hierarchy and what are the trends regarding the ratios of latent variables erroneously associated with the disease. Therefore, we used realistic simulated data or real data designed or known to harbour a causal SNP. We name *indirect genetic association* any dependence between a causal SNP ancestor node (abbreviated as A) in the FLTM and the disease. This dependence is due to the fact that an A node is likely to capture the information of the causal SNP. If indirect genetic association may be evidenced for A nodes, the identification of A nodes will allow pointing out potentially causal markers since the latter are leaf nodes of the trees rooted in A nodes (see Figure 2 which clarifies the meaning of specific key terms further used). We will examine the difference between causal SNP ancestors (As) and other latent nodes (abbreviated as Os). We will also examine the behaviour of Os in causal trees (OTs) and of Os outside causal trees (OOs).

5.1 Simulation of Realistic Genotypic Data

Conducting a systematic analysis under controlled conditions requires that we are able to simulate both realistic SNP data and an association between one of these SNPs and the disease status (affected/unaffected). For this purpose, we have chosen one of the most widely used software applications, namely HAPGEN (http://www.stats.ox.ac.uk/∼marchini/software/gwas/hapgen.html) (Spencer et al., 2009). The reader well acquainted with such HAPGEN simulations may skip the two following paragraph, where we describe the simulation in the case of a single causal SNP.

Generating realistic genotypic simulation lies in the ability to mimic linkage disequilibrium (see Introduction, fourth paragraph). HAPGEN relies on the haplotypes (sequence of alleles, see Introduction, last paragraph) of a population of reference, to generate new haplotypes as mosaics of the known haplotypes, for a user-specified number of cases and controls. The genotype of any individual is generated based on the two haplotypes simulated for this individual.

HAPGEN selects at random the causal SNP, checking for the minor allele frequency to be within a user-specified range. Assuming causality under a specific disease model and effect sizes, it is straightforward to calculate the genotype frequencies in cases at that locus. On this basis, any case individual is simulated by first simulating the alleles at the causal locus and then working outwards in each direction to construct the two haplotypes. Note that the same mechanism governs the construction of haplotypes, whatever the status of the individual (case or control). The only distinction lies in that the locus from which the extension is started is chosen at random, for controls. For cases, the extension is initiated from the causal locus. The extension processes conditionally on reference haplotypes and is ruled by the fine-scale knowledge of recombination rates and the physical distance between loci, to calculate the probability of breaks in the mosaic pattern as one moves along the region. Moreover, partial copies (of haplotype subregions) are blurred by simulated mutations.

To control the simulation conditions, three ingredients have been combined: *minor allele frequency* (MAF) of the causal SNP, severity of the disease expressed as *genotype relative risks* (GRRs) for various *disease models*. The range of the MAF at the causal SNP has been specified to be 0.1-0.2, 0.2-0.3 or 0.3-0.4. Various genotype relative risks have been considered and the disease model has been specified amongst additive, dominant, multiplicative or recessive (*add*, *dom*, *mul*, *rec*). These choices are justified as standards used for simulations in association genetics.

For short, together with GRRs, the disease models allow specifying the probability to be affected, depending on the genotype at the causal locus: $GRR = \frac{\mathbb{P}(affected|Aa)}{\mathbb{P}(affected|aa)}$, where A is the disease allele. The specification of the disease model amongst *add*, *dom*, *mul* and *rec* allows the adjustment of the probability to be affected when carrying the two disease alleles *AA*, with respect to the probability to be affected when carrying *Aa* (or *aA*). Thus various effect sizes may be simulated. If 1 stands for the effect when no disease allele is present at the causal locus (*aa*), the effect sizes for the *Aa* and *AA* carriers are respectively: $1+\frac{\alpha}{2}$,$1+\alpha$ (*add*); $1+\alpha$, $1+\alpha$ (*dom*); $1+\alpha$, $1+\alpha^2$ (*mul*); 1, $1+\alpha$ (*rec*).

To run HAPGEN, we have chosen the widely used reference haplotypes of the HapMap phase II coming from U.S. residents of northern and western European ancestry (CEU) (http://hapmap.ncbi. nlm.nih.gov/). The disease prevalence (percentage of cases observed in a population) specified to HAPGEN has been set to 0.01, a standard value used for disease locus simulation. The simulated data have been generated for 1000 unaffected subjects and 1000 affected subjects and consist of unphased genotypes relative to a 1.5 *Mb* region containing around 100 SNPs. Combining all previous conditions led to testing 36 scenarii ($3 \times 3 \times 4$). To derive significant trends, we replicated each sce-

nario 100 times. Together with our aim of a comprensive study, the necessity to replicate explains our choice of the number of variables (100 SNPs). Standard quality control for genotypic data has been carried out: SNPs with MAF less than 0.05 and SNPs deviant from the so-called Hardy-Weinberg Equilibrium (not detailed) with a p-value below 0.001 have been removed.

5.2 Detection of Genetic Associations

We have used the G^2 standard test of independence rather than the well-known Chi^2 test: for relatively small sample sizes (below 300 subjects) as is the case for the real dataset analyzed, G^2 is recommended. We have compared the p-values obtained, successively testing the phenotype Y against the causal SNP, the causal SNP ancestor nodes (A nodes) and other nodes (abbreviated as Os) in the FLTM's graph. We display the $-log_{10}$(p-value) values. Values near 0 point out independence and the previous indicator increases with the strength of the dependence.

To measure the significance of associations, we have implemented a permutation procedure dedicated to the computation of the per-test error rate α' (type I error), in order to control the family-wise error rate α (type I error) at 5%. Namely, α' defines the significance threshold for each association test. α controls the probability to make one or more false discoveries among all hypotheses when performing multiple association tests. The procedure implemented to obtain α' is the following: (i) for each permutation, and each FLTM's layer L, we perform independence tests between the variables in L and the phenotype. For each permutation, the minimum of the p-values over all variables belonging to L is identified (*pmin*). Given the threshold α, the distribution of *pmins* over all permutations allows to extract a pointwise threshold α'.

An advantage of the FLTM strategy relies on the fact that there are less variables in the highest layers than in the lowest ones. Thus, we expect an increase of α' with the layer level. That is the reason why our permutation procedure was adapted to the calculation of layer-specific thresholds α'.

6 RESULTS AND DISCUSSION

6.1 Simulated Data

In the following, the data analysis has entailed the generation of up to 7 layers in the FLTMs. We will not report results obtained for layers with numbers above 3: such layers do not provide sufficient data to compute representative medians or draw informative boxplots. On average, over all 3600 FLTMs (36 scenarii $\times$ 100 replicates), the percentages of nodes are distributed as follows: 89.1% in layer 0, 9.5% in layer 1, 1.2% in layer 2 and 0.2% in layer 3.

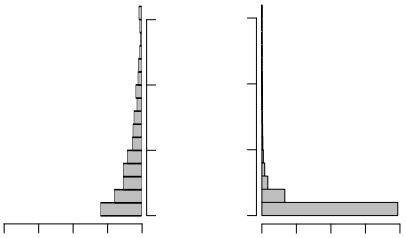


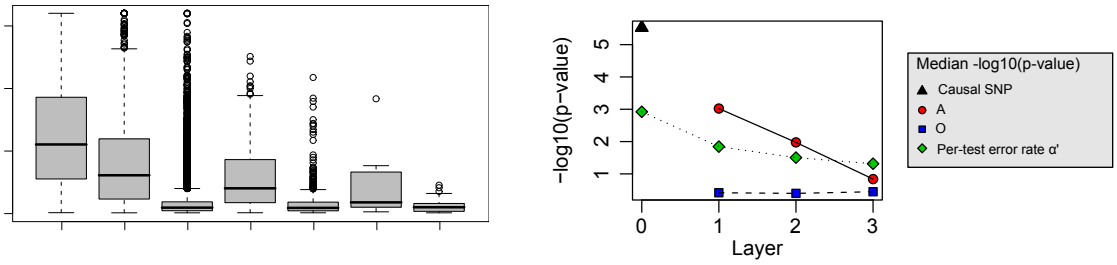
Figure 3: Histograms of $-log_{10}$(p-value) values resulting from association tests of the phenotype with the causal SNP ancestor nodes (As) and the other latent nodes (Os). General results compiling all simulated scenarii (see 5.1): MAF (0.1-0.2, 0.2-0.3 and 0.3-0.4), heterozygous GRR (1.4, 1.6 and 1.8) and disease model (dominant, recessive, additive and multiplicative).

6.1.1 General Trends

Figure 3 compares the histograms of $-log_{10}$(p-value) values resulting from association tests of Y with A nodes and O nodes, respectively. The comparison of these two histograms reveals a large dissimilarity between the two distributions. The majority (70%) of $-log_{10}$(p-value) values relative to A nodes is greater than 1, whereas it is the case for only 19% of O nodes. Indeed, we observe that large $-log_{10}$(p-value) values (*e.g.*, greater than 5) are common for the former and are very rare for the latter. A non-parametric test, the Wilcoxon rank-sum test, shows a p-value less than 10^{-16}, thus confirming that A and O p-values follow two different distributions.

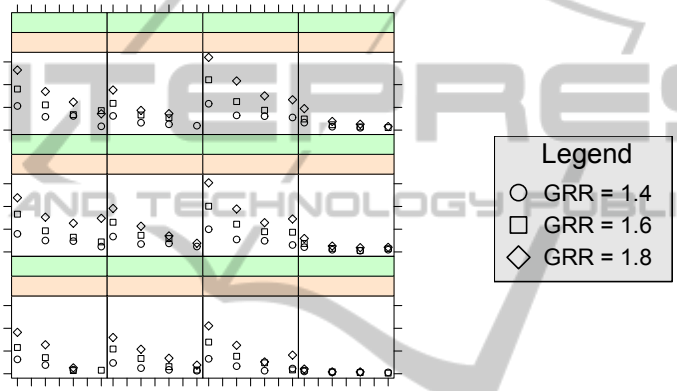
Figure 4(a) more thoroughly describes the $-log_{10}$(p-value) values observed for the different layers of the FLTM in the cases of tests relative to A and O nodes. The layer 0 refers to the association tests between the phenotype and the causal SNP and serves as the reference value. In this figure, we observe that the association strength for A nodes slowly decreases when the layer number increases, whereas the association for O nodes sticks to $-log_{10}$(p-value) values below 0.4, corresponding to p-values greater than 0.4. Although O nodes reveal false positive associations (less than 10% have a p-value below 0.01), these results clearly highlight a general trend: indirect associations are captured by the A nodes while it is not the case for a large majority of O nodes.

Figure 4(b) emphasizes the general trend of

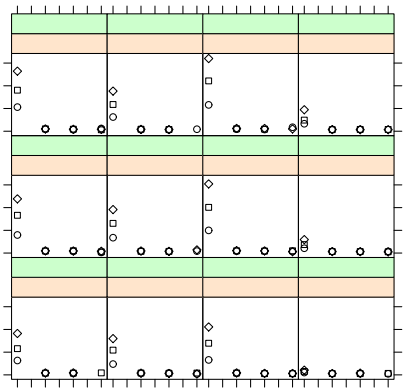


(a) (b)

Figure 4: $-log_{10}$(p-value) values for the different layers of the FLTM, resulting from association tests of the phenotype with the causal SNP ancestor nodes (As) and with the other latent nodes (Os) - simulated data. (a) Boxplots. (b). Median values. The layer 0 shows the results of the association tests between the phenotype and the causal SNP (over all simulated scenarii). See Figure 3 for details about the scenarii, see last paragraph of 5.2 for the definition of error rate α'.



(a)



(b)

Figure 5: Median $-log_{10}$(p-value) values for the different layers of the FLTM, resulting from association tests between the phenotype and latent nodes - simulations under thirty-six conditions. (a) Causal SNP ancestor nodes (As). (b) Other latent nodes (Os). The different windows represent possible genetic scenarii. At the top of each window, the range of the simulated causal SNP's minor allele frequency and the disease model assumption (additive, dominant, multiplicative or recessive) are indicated. The three different symbols used refer to as many genotype relative risks considered for the simulated causal SNP (see Legend and 5.1). The layer 0 refers to the association tests between the phenotype and the causal SNP (over all 100 replications).

$-log_{10}$(p-value) observed for A and O nodes, and compares the median $-log_{10}$(p-value) value obtained for each layer to the corresponding value associated with the significance threshold α' specific to this layer (see Subsection 5.2, second paragraph). This figure reveals that up to the second layer, significant associations are identified for A nodes. In contrast, regarding O nodes, for all layers, median $-log_{10}$(p-value) values are smaller than the corresponding $-log_{10}(\alpha')$ values. Focusing on the O distribution, we observe that the percentage of p-values lower than α' (false positives) is 4.7%.

The existence of the false positives (FPs) can partly be explained by the presence of indirect dependences between the causal SNP and the OTs, that is the O nodes located in the causal tree. The causal tree is the tree containing the causal SNP (see Figure 2). At the opposite, no FPs are expected for O nodes outside the causal tree (OOs). The three key relations are: Os = OTs(21%) + OOs(79%); True Positives = As; False Positives = FP Os = FP OTs (73%) + FP OOs (27%). Thus 73% of FPs are in the causal tree, representing only 21% of O nodes (in causal tree and other trees). In the causal tree, the FP rate is 16%; over all non causal trees, the FP rate is 1.6%. In conclusion, the major part of FPs is confined in the causal tree.

6.1.2 Behaviour under Thirty-six Genetic Scenarii

We now compare association test results between A and O nodes for each scenario described in Subsection 5.1 (see Figure 5). Globally, similar tendencies are observed over all scenarii: the association strength drops continuously from bottom to layer number 3; in the case of O nodes, an overwhelming majority of results points out the absence of association, whichever the FLTM's layer concerned.

When considering the easiest case (MAF range = 0.3-0.4, GRR = 1.8 and multiplicative model), over all layers, the A nodes present strong associations ($-log_{10}$(p-value) > 7). Regarding a less ideal but more plausible configuration (MAF range = 0.2-0.3, GRR = 1.6 and additive model), the median $-log_{10}$(p-value) value computed for A nodes decreases from 8.3 at layer 0, to reach 4.6, 3.2 and 2.2 at layers 1, 2 and 3, respectively. On the contrary, when the model is recessive, the association with the causal SNP is low and the A nodes cannot capture anything (similar results are obtained with most of the methods dedicated to association studies). As regards the O nodes, null associations are reported in all configurations.

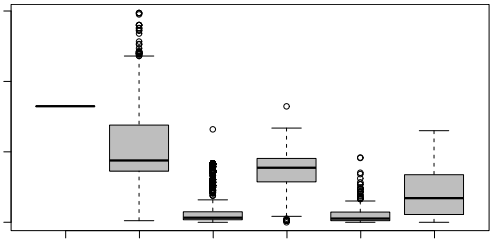


Figure 6: Boxplot of $-log_{10}$(p-value) values for the different layers of the FLTM, resulting from association tests of the phenotype with the causal SNP ancestor nodes (As) or with the causal SNP non-ancestor nodes (Os) - real data. Layer 0 refers to the association test between the phenotype and the causal SNP (marker 19). In layer 3, no O nodes are observed in the FLTMs.

6.2 Real Data

We have also evaluated on real data the ability of FLTM models to capture the indirect associations with the phenotype. We have used a reference dataset relative to a 890 *kb* region flanking the *CYP2D6* gene on human chromosome 22q13. This gene has a confirmed role in drug metabolism (Hosking et al., 2002). The dataset consists of 32 SNP markers genotyped for 268 individuals and was downloaded from the R package graphminer developed by Verzilli and collaborators (Verzilli et al., 2006). The SNP 19 at the position 550 *kb* is the closest marker to *CYP2D6* gene (at 525.3 *kb*). For this reason, for our experiment, we have considered the SNP 19 as the causal marker.

To take into account the stochastic nature of our algorithm (random initialization of parameters during the EM algorithm), we present the results relative to 1000 runs. Each run takes on average 5.4 *s* on a standard PC computer (3 *GHz*, 2 *GB* RAM). On average, over all 1000 FLTMs (1000 replicates), the percentages of nodes are distributed as follows: 82.62% in layer 0, 16.89% in layer 1, 0.39% in layer 2 and 0.10% in layer 3. Figure 6 shows the $-log_{10}$(p-value) values of association tests relative to As and Os. As expected in view of experiments led on simulated data, the A nodes succeed in capturing indirect association, in particular in layer 1, with a median value of 5.5, corresponding to p-values lower than 5.10^{-6}. In the other layers, the strength of associations is lower but remains relatively high as in the layer 2 showing a median value of 4, equivalent to a p-value of 10^{-4}. As previously seen, when we focus on O nodes, we observe very few strong associations. The majority of p-values (over 80%) is greater than 0.01.

7 CONCLUSIONS AND PERSPECTIVES

Based on both simulated and real data analyses, this paper promotes the use of FLTMs as a simple and useful framework for disease association detection in human genetics. Efficient capture of indirect genetic association is achieved through two major reasons: (i) the causal SNP ancestor nodes succeed in capturing indirect associations with the phenotype; (ii) at the opposite, the other latent nodes globally show very weak associations. In other words, this property allows to distinguish between true and false indirect genetic associations.

The numbers of SNPs in the benchmarks were limited. Nonetheless, this limitation is not a bias to the sound characterization of the fading of information in the FLTM hierarchies: bottom-up information decays does concern the forest depth and does not interfere with the forest width. It must be underlined that our tests were not designed to meet the small n, large p condition (many more variables (SNPs) than subjects) as in genome-wide association studies (GWASs). Again, this is not a bias to our study: over thirty-six various scenarii, we have shown that the overwhelming part (about three quarters) of false positives confines in a unique tree, namely the one harbouring the causal SNP (causal tree). In the conditions of a GWAS, the forest width may well be far larger than those observed in our tests, the false positives are expected to remain confined in the causal tree, for the major part.

In a previous work, we have developped a scalable FLTM learning algorithm, thus reaching orders of magnitude consistent with GWAS demands (10^5 variables, 2000 individuals). In addition to scalability, data dimension reduction advocates the use of FLTM-based modeling in GWASs: the issue of multiple hypothesis testing in GWASs would be resolved by testing a low number of latent variables instead of a large number of observed variables. However, before envisaging an FLTM-based GWAS, an inescapable prerequisite was testing whether the bottom-up information fading through the forest would nevertheless allow reliable association detection. No less unavoidable was the close examination of ratios of latent variables erroneously associated with the disease.

A precursory work to the GWAS concern, the present contribution assets the soundness of the FLTM model for association detection. Besides, we have conceived a procedure to guarantee a given family-wise (type I) error rate through the computation of layer-specific per-test error rates. The successful test of our algorithm under a large spectrum of conditions allows its integration in a GWAS tool.

REFERENCES

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). Clustering gene expression patterns. In *Proc. of the 3rd annual int. con. on Computational molecular biology*, pages 33–42.

Chen, T., Zhang, N., Liu, T., Poon, K., and Wang, Y. (2011). Model-based multidimensional clustering of categorical data. In *Artificial intelligence, in press*.

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(2):229–232.

Han, B., Park, M., and Chen, X. W. (2010). A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinformatics*, 11(Suppl 3):S5+.

Harmeling, S. and Williams, C. K. I. (2011). Greedy learning of binary latent trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1087–1097.

Hosking, L. K., Boyd, P. R., and Xu, C. F. e. a. (2002). Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity. *Pharmacogenomics J.*, 2(3):165–175.

Hwang, K.-B., Kim, B.-H., and Zhang, B.-T. (2006). Learning hierarchical bayesian networks for large-scale data analysis. In *ICONIP*, pages 670–679.

Mourad, R., Sinoquet, C., and Leray, P. (2010). Learning hierarchical Bayesian networks for genome-wide association studies. In *COMPSTAT*, pages 549–556.

Mourad, R., Sinoquet, C., and Leray, P. (2011). A hierarchical Bayesian network approach for linkage disequilibrium modeling and data-dimensionality reduction prior to genome-wide association studies. *BMC Bioinformatics*, 12:16+.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Spencer, C. C., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, 5(5):e1000477+.

Verzilli, C. J., Stallard, N., and Whittaker, J. C. (2006). Bayesian graphical models for genome-wide association studies. *The American Journal of Human Genetics*, 79:100–112.

Wang, Y., Zhang, N. L., and Chen, T. (2008). Latent tree models and approximate inference in Bayesian networks. *Machine Learning*, 32:879–900.

Zhang, N. L. (2004). Hierarchical latent class models for cluster analysis. *JMLR*, 5:697–723.

Zhang, N. L. and Kocka, T. (2004). Efficient learning of hierarchical latent class models. In *ICTAI*, pages 585–593.

Zhang, Y. and Ji, L. (2009). Clustering of SNPs by a structural EM algorithm. In *Int. Joint Conf. on Bioinformatics, Systems Biology and Intelligent Computing*, pages 147–150.