Bayesian multi-locus pattern selection and computation through reversible jump MCMC

Christine Sinoquet

Bayesian multi-locus pattern selection and computation through reversible jump MCMC

Christine Sinoquet

LINA, UMR CNRS 6241, Université de Nantes, 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex, France

— KnOwledge and Decision (KOD) —

LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

RESEARCH REPORT

N° hal-00524885

October 2010

UNIVERSITÉ DE NANTES

ECOLE DES MINES DE NANTES

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Christine Sinoquet

Bayesian multi-locus pattern selection and computation through reversible jump MCMC

32 p.

Bayesian multi-locus pattern selection and computation through reversible jump MCMC

Christine Sinoquet

christine.sinoquet@univ-nantes.fr

Abstract

In the human genome, susceptibility to common diseases is likely to be determined by interactions between multiple genetic variants. We propose an innovative Bayesian method to tackle the challenging problem of multi-locus pattern selection in the case of quantitative phenotypes. For the first time, in this domain, a whole Bayesian theoretical framework has been defined to incorporate additional transcriptomic knowledge. Thus we fully integrate the relationships between phenotypes, transcripts (messenger RNAs) and genotypes. Within this framework, the relationship between the genetic variants and the quantitative phenotype is modeled through a multivariate linear model. The posterior distribution on the parameter space can not be estimated through direct calculus. Therefore we design an algorithm based on Markov Chain Monte Carlo (MCMC) methods. In our case, the number of putative transcripts involved in the disease is unknown. Moreover, this dimension parameter is not fixed. To cope with trans-dimensional moves, our sampler is designed as a reversible jump MCMC (RJMCMC). In this document, we establish the whole theoretical background necessary to design this specific RJMCMC.

Introduction

In the hunt for genes affecting our health and wellbeing, association studies look for associations between genetic features and phenotypes such as health / illness. Many common diseases in humans are suspected to be caused by complex *epistatic* interactions among multiple genes. In the literature, different acceptations are encountered for the term *epistasis* [4]. In this paper, epistasis is defined for a set of genetic loci as the situation arising when not all loci, and possibly none at all, have a main effect on the disease, whereas the combination of the loci causes the disease. Marginal epistatic interactions have been identified for diseases such as coronary heart disease [21], breast cancer [24], Alzheimer's disease [38] and Crohn's disease [30].

The last decade has witnessed an explosion in the number of research works aiming at tackling epistasis identification. Amongst deterministic approaches, we mention SNPRuler, a branch and bound algorithm devoted to the expansion of sets of SNPs in the binary phenotype case [33]. This method describes the relationship between the epistatic SNPs and the phenotype through a predictive rule. A measurement of rule relevance is deduced from the χ-square statistic. A rule is grown if the added SNP increases the relevance. Specific properties of this measure, as well as the possibility to calculate an upper bound, allow to traverse the space of predictive rules without exhaustive search. Central to software TEAM [36] is the speeding up of contingency table (CT) computation through a true structure. This method restrains to two-SNP epistasis and tests such as the χ-square test. Given the CTs of two single SNPs and the CT for genotype relation between these two SNPs, only little effort is required to compute the two-locus CT. The CT for genotype relation between the two SNPs is inferred from a minimum spanning tree built on the SNPs. Therein, each edge represents the genotype difference between the two connected SNPs.

Supervised learning algorithms include standard regression methods as well as stepwise approaches. Least square or maximal likelihood estimations are the rule for quantitative (continuous) phenotypes. Logistic regression (or binomial regression) is devoted to binary phenotypes (affected/unaffected status). In logistic regression, parameter estimation maximizes the likelihood and usually relies on Markov Chain Monte Carlo (MCMC) sampling strategies for this purpose. Some approaches combine forward stepwise procedure and logistic regression [16, 22]. Logic regression attempts to identify boolean combinations of SNPs for the prediction of the affected/unaffected status of an observation. The logic expressions are represented by logic trees. Permissible moves in the tree-growing process are alternating an operator or a variable, pruning or growing a branch, and adding or removing variables. To find the best models, stochastic algorithms are applied, such as simulated annealing [26] or MCMC [25]. As a matter of fact, in the former work, simulated annealing is applied to different subsets of the data. Dealing with binary phenotypes, Symbolic Discriminant Analysis (SDA) may be viewed as an extension of both linear and logic regression approaches [18]. In SDA, the data dictates the size, shape and complexity of a symbolic function, discriminant for case/control status. The symbolic function combines mathematical functions from a list provided by the user. Genetic programming is used to optimize the discriminant power of the models [17].

Non-parametric data mining strategies have been investigated. Besides standard forest-based approaches [3], random forests combine bagging with random selection of features (see MegaSNPHunter [32], for example). In *bagging* - or bootstrap aggregating - a few hundred to a few thousand classification or decision trees are generated from as many bootstrap samples drawn from the available data. Multifactor dimensionality reduction (MDR) applies an exhaustive search to pool genotypes from combinations of SNPs [9]. Thus, data dimension is reduced to one, with two genotype pools accumulating either affected or unaffected subjects. When the phenotype is continuous, pooling is achieved through a combinatorial partition of the genotypes [35]. In the previous Combinatorial Partitioning Method (CPM), for each com-

bination of SNPs, the partitions are exhaustively enumerated and tested for discriminating power. The Restricted Partition Method (RPM) is a heuristic which guides the straightforward construction of the best possible partition, per each combination of SNPs [6]. RPM implements ascending hierarchical clustering for this purpose.

Probabilistic graphical models were also used to search for causal SNP combinations. In an approach based on Markov random field models [31], the graph structure connects cliques of (pairwise) dependent SNPs with the phenotype node. An MCMC strategy samples over the space of possible graphs, with a restriction on physical distance between any two markers in a clique. The MCMC strategy samples from the posterior distribution of graphs conditional on the data. Only MCMC moves towards decomposable graphs are allowed, to allow an easy computation of marginal likelihood. A novel framework, forests of hierarchical latent class models, was introduced to handle high-dimensional data [19]. To learn the model, an ascending hierarchical clustering first discovers cliques of dependent SNPs, subsume them through additional (latent) variables if possible, then iterates the previous two steps on the latent variables and remaining SNPs.

Bayesian approaches relying on MCMC strategies have been investigated to search the space of SNP combinations. The BEAM [37] and *epi*MODE [29] programs implement a Bayesian marker partitioning model to identify candidate combinations, together with MCMC computation of the posterior probability that each candidate combination is associated with the disease. The BAMSE algorithm explores sets of effects (SNPs and environmental factors) that increase the risk (binary phenotype) or the phenotypic value (quantitative phenotype), for individuals who fulfill the criterion defined by the set [1].

Several reviews provide coverage of recent algorithm developments in the research domain around epistasy (see for instance [11, 20, 12, 27, 15]. The subject is hot topic and advanced methods are constantly proposed to attempt to tease associations out of datasets. For instance, some leads are incorporating data imputation to an association study (AS) process [10] or integrating gene expression data (GED) [13]. In particular, there was still room for investigating a Bayesian method based on GED integration. We propose a novel approach based on transcriptomic and genetic data integration, to tackle AS under the multigenic hypothesis. The genetic data considered are Single Nucleotide Polymorphisms (SNPs) and we only address continuous phenotypes in the present work. Downstream specific analyses, whose purpose is relating phenotypes to transcripts and genetic markers to gene expression data (and thus to transcripts), our procedure explores the search space consisting of SNP sets - or multi-locus patterns (MLPs) -. Such MLPs are as many candidates for phenotype explanation. Crossing phenotype/GED associations and GED/MLP associations into phenotype/GED/MLP associations is the final objective or our approach. However, since we address multigenic etiology, any such MLP may be covered by a set of transcripts - a transcript pattern (TP) -, on the genome. Thus, we can replace the previous scheme phenotype/GED/MLP with phenotype/TP/MLPs, where each transcript in the TP is co-located with one of the MLPs. Thus, we avoid a fine-grained description of MLPs, and escape the expensive search in the space of MLPs. The core idea of our proposal lies in that the SNP search space is connected to TP subspaces, which allows a coarse-grained MLP description. As we do not constrain the number of transcripts potentially involved in the disease etiology, we have to explore TP subspaces of various dimensions.

Besides, the linear regression model has often proven useful to describe the relationships between SNPs and continuous phenotype. Indeed, regression-based tests are current tools offered by the software suites dedicated to genome-wide association studies, such as the PLINK software toolbox ([23], http://pngu.mgh.harvard.edu/ purcell/plink/), the Golden Helix SNP & variation suite (http://www.golden helix.com), the snpMatrix R package [7] distributed as part of the BioCond uctor project (http://www.bioconductor.org). Mixing three ingredients - Bayesian framework, transcriptomic / genetic data integration, linear model, -

we have designed an innovative approach. Within this framework, the relationship between the genetic variants and the quantitative phenotype is modeled through a multivariate linear model. Then, to only focus on parts of the posterior distribution that are of interest on the large parameter space, we have conceived an algorithm based on Markov Chain Monte Carlo (MCMC) methods. In our case, the number of putative transcripts involved in the disease is unknown. Furthermore, this dimension parameter is not fixed. Therefore, to cope with trans-dimensional moves in the MCMC, our sampler is designed as a reversible jump MCMC (RJMCMC).

Our contribution in this report is twofold. We describe a whole Bayesian theoretical framework meant to integrate transcriptomic and genetic data for genetic association purpose. We describe the RJMCMC designed to perform the Bayesian computation. In particular, we provide here the theoretical background and derive the corresponding calculuses necessary to the implementation of our algorithm.

The first Section states the problem and gives the nomenclature necessary to describe the search space, for our specific case. Section 2 provides a gentle introduction to readers not familiar with MCMCs and RJMCMCs. The third Section introduces our framework. It first presents the moves allowed in our MCMC approach. Then it shows how transcriptomic and genotypic data are integrated through a multivariate linear model. This section ends with a sketch of the algorithm. The next section is devoted to the derivation of the posterior parameter distribution.

1 Premiminaries

1.1 Statement of the problem

We consider τ quantitative phenotypes - or targets -. Our aim is to identify potentially causal epistatic SNPs, in order to explain each target. Since we address multigenic etiology, we consider that any such causal set of SNPs may be covered by a *set* of transcripts - a transcript pattern -, on the genome. The problem we tackle arises downstream two series of studies: through a previous approach, relations between genetic markers and transcripts have been derived; besides, for any such transcript, associations with MLPs have been investigated. A solution to our problem assigns a TP to each target and an MLP to each transcript in the TP. This assignment has to best explain the determinism of the MLPs assigned to each target, on this target. No *a priori* is provided, regarding the sizes of TPs.

1.2 Notations and definitions

In the following, since we deal with transcripts that are co-located with SNPs, these transcripts will be referred to as co-mRNAs. The search space S to be explored is a set of solutions each assigning a set of active co-mRNAs to each target, together with an active MLP per each such active co-mRNA. A solution - or a state of the RJMCMC - is described through its parameters, η. Most of the constituents of η, together with their domains of variation, are described in Table 1), and illustrated by Figure 1.

2 A short introduction to Markov Chain Monte Carlo methods and reversible jump MCMCs

Here, we first provide a brief introduction to MCMCs. Then, we justify the construction of RJMCMCs and show how the theoretical framework is adapted to take into account trans-dimensional moves.

An ergodic (aperiodic and irreducible) Markov chain will converge towards a unique stationary distribution, π. Markov Chain Monte Carlo (MCMC) methods are a class of algorithms designed to sample

$\mathcal{T}, \mid \mathcal{T} \mid = \tau$	set of τ targets (or quantitative phenotypes)
$C^i, \mid C^i \mid = q^i, \sum_{i \in \mathcal{T}} q^i = q$	set of all q^i possible co-mRNAs known to potentially exert an impact on target i
$\mathcal{C}^i, \mid \mathcal{C}^i \mid = s^i$	the set of s^i active co-mRNAs for target i, in current JRMCMC state
$k = \sum_{i \in \mathcal{T}} s^i$	the number of active co-mRNAs over all targets
$M^j, \mid M^j \mid = m^j$	the set of all possible multi-locus patterns for co-mRNA j
$\mathcal{M}^i$	the set of all active multi-locus patterns for target i
	$\mathcal{M}^i = [u_{j_1}, u_{j_2}, \cdots, u_{j_{s^i}}]$
$z^{i\ell} = \mid \mathcal{M}^{i\ell} \mid = \mid u_{j_\ell} \mid$	size of the multi-locus pattern $\mathcal{M}^{i\ell}$ of target i, corresponding to active co-mRNA ℓ

Table 1: Nomenclature for parameter space description

from a desired probability distribution: their principle consists in constructing a Markov chain that has the desired distribution as its stationary distribution. Given an ergodic Markov chain, and p, the probabilities of transition from state to state in search space S (transition kernel), the property of reversibility between states x and y holds: $\pi(s)\ p(s\prime \mid s) = \pi(s\prime)\ p(s \mid s\prime)$ (detailed balance equation). Though reversibility is not necessary to guarantee convergence of the posterior to π, it is sufficient. Then, the key to MCMC consists in expressing the transition kernel $p(s\prime \mid s)$ as the product of an arbitrary proposal distribution, q, and an associated acceptance distribution, a: $p(s\prime \mid s) = q(s\prime \mid s)\ a(s, s\prime)$. To explain the intuition behind these concepts, suppose, without loss of generality, that for states s and $s\prime$, some given transition kernel p verifies $\pi(s)\ p(s\prime \mid s) > \pi(s\prime)\ p(s \mid s\prime)$. Artificial coercion of the previous formula towards reversibility is straightforward, introducing two terms $a(s, s\prime)$, strictly lower than 1, and $a(s\prime, s)$, equal to 1: $\pi(s)\ q(s\prime \mid s)\ a(s, s\prime) = \pi(s\prime)\ q(s \mid s\prime)\ a(s\prime, s)$. If inequality is reversed, then $a(s\prime, s)$, strictly greater than 1, and $a(s, s\prime)$, equal to 1, will be used instead. Finally, acceptance probability is calculated as: $a(s, s\prime) = min\left(1, \frac{\pi(s\prime)\ q(s\prime, s)}{\pi(s)\ q(s, s\prime)}\right)$. The arbitrary proposal distribution q and the acceptance probability a are the two ingredients of the Metropolis-Hastings (MH) algorithm (see Algorithm 1).

Algorithm 1 Metropolis-Hastings

1: initialize state X_0 arbitrarily; $i \leftarrow 0$
2: **repeat until convergence**
3: propose next value $X_{i+1} = y$ from the proposal distribution $q(. \mid X_i = x)$.
4: sample uniformly u in interval $[0, 1]$
5: **if** $(u \leq a(x, y))$ **then** $X_{i+1} \leftarrow y$ /* acceptance of proposed move */
6: **else** $X_{i+1} \leftarrow x$ /* rejection of proposed transition */
7: $incr(i)$
8: **end repeat**

When the search space writes as $S = \left\{(k, \theta^{(k)}), k \in \mathcal{K}, \theta^{(k)}) \in S_k\right\}$, where $\mathcal{K}$ is an enumerable set, then the posterior distribution can be factorized as $\pi(\theta^{(k)}, k) = \pi(\theta^{(k)} \mid k)\ \pi(k)$. To impose reversibility for each pair $(\theta_1^{(k_1)}), \theta_2^{(k_2)})$, the core idea is to supplement each of the corresponding sub-spaces $\mathcal{S}_{k_1}$ and $\mathcal{S}_{k_2}$ with adequate artificial spaces. Namely, θ^{k_1} and θ^{k_2} will be completed into (θ^{k_1}, u_1) and (θ^{k_2}, u_2), respectively. Sampling u_1 and u_2 from adequate distributions g_1 and g_2 will guarantee the existence of a bijection f_{k_1,k_2} between the augmented sub-spaces corresponding to $\mathcal{S}_{k_1}$ and $\mathcal{S}_{k_2}$. Under this condition, the acceptance probability now involves the product of four terms:

$$a(\theta_1^{(k_1)}, \theta_2^{(k_2)}) = min\left\{1,\ \frac{\pi(\theta^{(k_2)}, k_2)}{\pi(\theta^{(k_1)}, k_1)}\ \frac{p(k_1 \mid k_2)}{p(k_2 \mid k_1)}\ \frac{q(u_2 \mid k_2, \theta^{(k_2)})}{q(u_1 \mid k_1, \theta^{(k_1)})}\ \mathcal{J}_{f_{k_1,k_2}}\right\}, \tag{1}$$

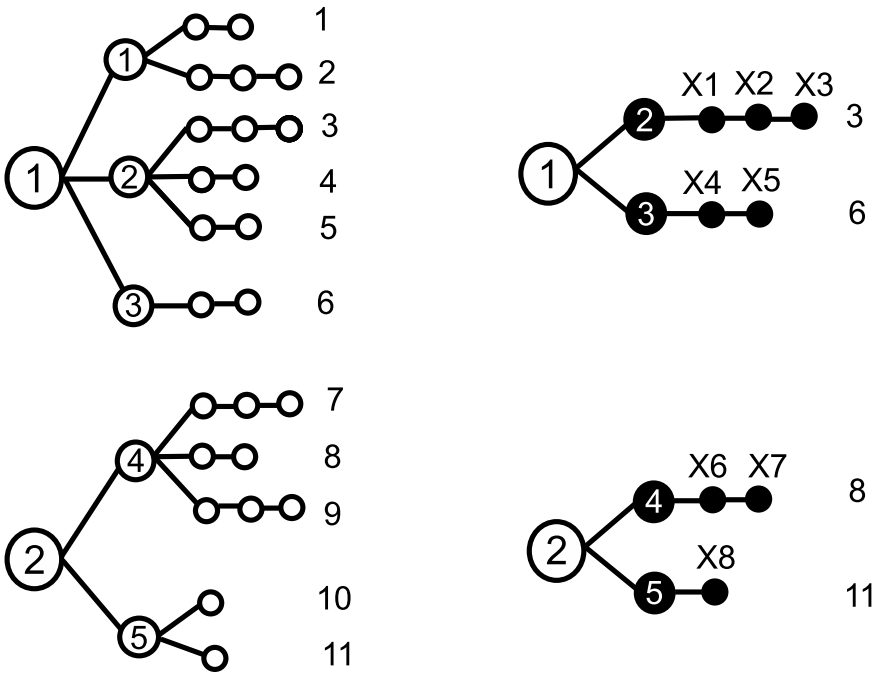

Figure 1: Illustration for definitions in Table 1. $\tau = 2$; left section of the Figure shows possible co-mRNAs and multi-locus patterns: $q = 5$; $C^1 = \{1,\ 2,\ 3\}$, $C^2 = \{4,\ 5\}$; $M^1 = \{1,\ 2\}$, $M^2 = \{3,\ 4,\ 5\}$, $M^3 = \{6\}$, $M^4 = \{7,\ 8,\ 9\}$, $M^5 = \{10,\ 11\}$ - right section displays an MCMC state with its active co-mRNAs and its active multi-locus patterns: $k = 4$; $\mathcal{C}^1 = \{2,\ 3\}$, $\mathcal{C}^2 = \{4,\ 5\}$; $\mathcal{M}^1 = \{3,\ 6\}$, $\mathcal{M}^2 = \{8,\ 11\}$; $z^{1,3} = 3$, $z^{1,6} = 2$, $z^{2,8} = 2$, $z^{2,11} = 1$. X_1 to X_8 denote random variables involved in the models describing the relationships between Y^1 and Y^2 (random variables associated with targets 1 and 2, respectively).

where $p(k_2 \mid k_1)$ denotes the probability of the dimensionality switch and $q(u_1 \mid k_1, \theta^{(k_1)})$ refers to the probability of transition $\theta^{(k_1)} \rightarrow \theta^{(k_2)}$. In some cases, including ours, the Jacobian $\mathcal{J}_{f_{k_1,k_2}}$ is equal to 1. Therefore, in these cases, the acceptance probability may be seen as the product of two terms: the posterior ratio distribution ($\frac{\pi(\theta^{(k_2)},k_2)}{\pi(\theta^{(k_1)},k_1)}$) and the sub-product $\frac{p(k_1|k_2)}{p(k_2|k_1)}\ \frac{q(u_2|k_2,\theta^{(k_2)})}{q(u_1|k_1,\theta^{(k_1)})}$, which is called the proposal ratio. In this case, the role of all these ingredients is made explicit in the generic description of an iteration of the reversible jump MCMC algorithm (RJMCMC) (see Algorithm 2).

3 The RJMCMC framework

3.1 Description of the five moves

To explore S, we allow five moves: addition of an active co-mRNA (A), dismissing of an active co-mRNA (D), substitution for an active co-mRNA (C), substitution for an active MLP (M), modification of the regression coefficients (R). Move A and move D respectively add and dismiss an entry in both lists $\mathcal{C}$ and $\mathcal{M}$. Move C updates an entry in both lists $\mathcal{C}$ and $\mathcal{M}$. Move D only updates an entry in list $\mathcal{M}$. Figure 2 starts from the MCMC state depicted in Figure 1, to describe moves A, D, S and M. Clearly, the MCMC will possibly traverse subsets of the search space characterized by different values of parameter

Algorithm 2 Generic description of an iteration in the RJMCMC algorithm

1: current state is $X_i = \theta^{(k_1)}$ in sub-space S_{k_1}
2: draw u_1 from $q(. \mid k_1, \theta^{(k_1)})$
3: calculate $\theta^{(k_2)}$ using bijection f_{k_1,k_2}: $(\theta^{(k_2)}, u_2) = f_{k_1,k_2}(\theta^{(k_1)}, u_1)$ /* $\theta^{(k_2)}$ is the value proposed for next state X_{i+1} */
4: sample uniformly u in interval $[0, 1]$ and apply lines 5 and 6 of Algorithm 1 with the acceptance probability calculated in Equation 1.

k, the number of active co-mRNAs over all targets. Thus, we have to deal with the trans-dimensional case.

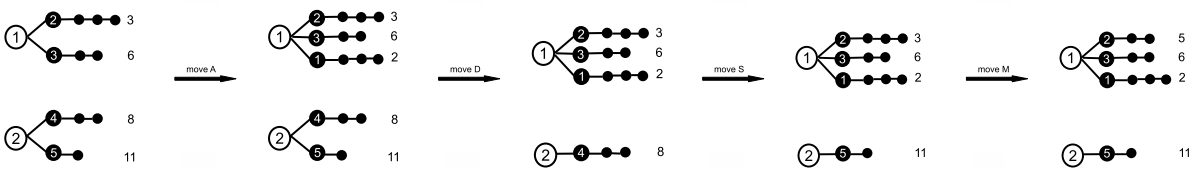

Figure 2: Moves of the MCMC

3.2 The underlying model linking phenotypes to causal multi-locus patterns *via* co-mRNA patterns

The data consist in τ quantitative phenotypes (or targets), $(y^i)_{i\in\mathcal{T}}$ and an array of genetic markers x. These date describe p individuals. Random variables are defined accordingly: Y^i, defined over $\mathbb{R}$, for target i, and $X^{i\ell t}$ (categorical, defined on domain $\{0, 1, 2\}$, $i \in \mathcal{T}, 1 \leq \ell \leq s^i, 1 \leq t \leq z^{i\ell}$; individuals in rows, SNPs in columns). Our hypothesis is that of a multivariate linear model:

$Y^i = a^{i0} + \sum_{1\leq\ell\leq s^i, 1\leq t\leq z^{i\ell}} a^{i\ell t} X^{i\ell t} + \varepsilon^i$, with ε^i following a normal distribution ($\varepsilon^i \sim \mathcal{N}(0, \sigma^i)$). In the previous formula, i denotes a target, ℓ an active co-mRNA of target i, and t the t^{th} SNP in the current active MLP of active co-mRNA ℓ. The predictors $X^{i\ell_1} \cdots X^{i\ell z^{\mathcal{M}^{i\ell}}}$ correspond to the SNPs in active MLP $\mathcal{M}^{i\ell}$. The (complete) putative causal MLP consists of MLPs $\mathcal{M}^{i1}$ $\mathcal{M}^{i2}$ ... $\mathcal{M}^{is^i}$. For example, in Figure 1, the two current linear regression models respectively describe the relationship between Y^1 and $X_1 \cdots X_5$, and between Y^2 and $X_6 \cdots X_8$.

We now introduce a convenient notation to refer to the matrix of regression coefficients associated with i^{th} target:

Notation 3.1 *Regression coefficients* a_{X^i}
The predictor set X^i *associated with target* i *has size* $\sum_{\ell=1}^{s^i} z^{i\ell}$. *The whole set of regression coefficients is then:*

$$a^i = a_{X^i} = (a^{i0}, (a^{i\ell t})_{1\leq t\leq z^{i\ell}}) \tag{2}$$

Thus, a state in our RJMCMC is fully described through parameter $\eta = (k, s, \mathcal{C}, \mathcal{M}, z, X, a, \sigma)$.

3.3 Outline of the algorithm

The sketch of the method is depicted in Algorithm 3.

Algorithm 3 RJMCMC

INPUT:
$\mathbf{x}$, a matrix describing p subjects (rows) with regard to e genetic markers (columns).
$\mathbf{x_i^j}$ is a categorical value (genotype code) ($1 \leq i \leq p$; $1 \leq j \leq e$)
$\mathbf{y}$, a matrix describing τ targets (rows), with regard to the p subjects (columns)
$\mathbf{y_i^j}$ is a quantitative continuous value
For each target i, $\mathbf{M^i}$ is the set of mRNAs hypothesized to exert an influence on target i.
For each mRNA i, $\mathbf{C^i}$ is a set of multi-loci patterns co-localized with co-mRNA i.
A multi-loci pattern is a set of genetic markers.

OUTPUT:
For each target i, $\mathcal{M}^\mathbf{i}$, the most frequent set of active multi-loci patterns encountered during stationary regime of the RJMCMC.

Initialization: $(k, s, \mathcal{C}, \mathcal{M}, z, X, a, \sigma) \leftarrow (k_0, s_0, \mathcal{C}_0, \mathcal{M}_0, z_0, X_0, a_0, \sigma_0)$
do
  $sample\ u \sim \mathcal{U}_{[0,1]}$ /* uniform draw in interval [0,1] */
  switch
    $u < a_k$: proposeMoveA /* add-active-co-mRNA */
    $a_k \leq u < a_k + d_k$: proposeMoveD /* Delete-active-co-mRNA */
    $a_k + d_k \leq u < a_k + d_k + c_k$: proposeMoveC /* Substitute-active-co-mRNA */
    $a_k + d_k + c_k \leq u < a_k + d_k + c_k + m_k$: proposeMoveM /* Substitute-active-multi-locus-pattern */
  else moveR /* change of regression coefficients for all active multi-locus patterns */
                /* associated with all active co-mRNAs */
  end switch
until (convergence)

Moves A to M occur with respective probabilities a_k, d_k, c_k and m_k, depending on k, the current number of active co-mRNAs over all targets. Probabilities c_k and m_k indirectly depend on k since a_k and d_k are evaluated as follows:

$$a_k = c\, min(1,\ \frac{p_{\bar{k}}(k+1)}{p_{\bar{k}}(k)}),\ d_k = c\, min(1,\ \frac{p_{\bar{k}}(k-1)}{p_{\bar{k}}(k)}), \tag{3}$$

where k is assumed to follow an *a priori* truncated Poisson distribution with mean λ, in the line of multiple changepoint approaches involving reversible jump MCMC [8, 28]:

$$p_{\bar{k}}(k) \propto \frac{\lambda^k}{k!} 1_{\{k \leq \bar{k}\}}. \tag{4}$$

Depending on c value adjustment and balance between c_k and m_k, we can state than some moves are more often proposed than others. Hyperparameter λ is updated at each iteration of the MCMC. Following [2], λ is sampled as follows:

$$k \sim \mathcal{N}(\frac{1}{2} + k + \varepsilon_1, 1 + \varepsilon_2), \tag{5}$$

with $\varepsilon_i << 1$, $(i = 1, 2)$.

Except for move R, the feasibility of a move is subject to the satisfaction of various constraints (see Table 2):

move	constraints
A	$a_{\bar{k}} = 0,\ a_{\bar{q}} = 0$
D	$d_0 = 0$ Each target must appear in any proposed solution with at least one active co-mRNA. Thus, the active co-mRNA proposed for dismissing must be checked to be associated with a target currently showing a number of active co-mRNAs strictly greater than 1.
C	The co-mRNA proposed for replacing an already active co-mRNA is necessarily associated with a target checking the following constraint: the number of possible co-mRNAs must be strictly greater than the number of current active co-mRNAs.
M	The active co-mRNA concerned by the replacement of its current active multi-loci pattern must be checked to possess at least one more possible multi-loci pattern.

Table 2: Constraints involved in the calculation of move probabilities. $\bar{k}$: maximal value allowed for k; q, maximal number of possible co-mRNAs, over all targets.

4 Space parameter posterior distribution

For all moves except move R, the acceptance probability must first be evaluated, to further validate or reject the move from current state η to proposed state η'. Andrieu and Doucet's works pionnered the theoretical construction of an RJMCMC based on a multivariate linear model [2]. In their founder approach, the parameter description includes, quite classically, a variance parameter, and more specifically, regression coefficients. These authors have shown that in the evaluation of the acceptance probability, the Jacobian term is equal to 1.

In the expression

$$\begin{aligned} \alpha_{\eta,\eta'} &= min(1,\ r_{\eta,\eta'}) \\ &= min(1,\ \textit{posterior distribution ratio} \times \textit{proposal ratio}) \\ &= min\left(1,\ \frac{p(\eta^{'} \mid y)}{p(\eta \mid y)} \times \textit{proposal ratio}\right), \end{aligned} \tag{6}$$

we now focus on posterior distribution ratio ($\frac{p(\eta^{'}|y)}{p(\eta|y)}$), where y represents the data, that is the phenotypes, in our case. To evaluate the posterior distribution ratio, we have to derive an algebraical expression for $p(\zeta \mid y)$.

Indeed, we will not deal with $\frac{p(\eta^{'}|y)}{p(\eta|y)}$, but instead with $\frac{p(\zeta^{'}|y)}{p(\zeta|y)}$. It is crucial to distinguish that in the full description of the MCMC state η:

$$\eta = (k, s, \mathcal{C}, \mathcal{M}, z, X, a, \sigma) = (\zeta, a, \sigma),$$

parameters a and σ are not assigned a status identical to that of other parameters. Our guidelines, the works of Andrieu and Doucet, have established that we are allowed to carry out the integration of the

so-called "nuisance parameters" a and σ in expression $p(\eta \mid y)$, to obtain an expression for $p(\zeta \mid y)$. Thus we will consider a move proposal without previously generating parameters a (and σ, in the case of move A) for the modified target. To be rigorous, we will now write $a_{\zeta,\zeta'}$ and $r_{\zeta,\zeta'}$ instead of $a_{\eta,\eta'}$ and $r_{\eta,\eta'}$.

To reach our objective, calculate $p(\zeta \mid y)$, we will implement the six steps recapitulated in Table 3.

(1)	To render explicit $p(\eta \mid y)$, use Bayes theorem and write $p(\eta \mid y) \propto p(y \mid \eta)\, p(\eta)$.
(2)	evaluate $p(\eta)$.
(3)	evaluate $p(y \mid \eta)$.
(4)	substitute the expressions obtained in steps (2) and (3) for the corresponding terms in $p(y \mid \eta) \times p(\eta)$ and obtain a first algebraic formula for $p(y \mid \eta)$.
(5)	transform $p(y \mid \eta)$ into a formula more appropriate for integration.
(6)	perform marginalization over the nuisance parameters a and σ, that is, eliminate a and σ from $p(\eta \mid y)$, through integration, to obtain the posterior distribution $p(\zeta \mid y)$.

Table 3: The six steps necessary to derive the parameter posterior distribution.

4.1 Step 1 - Use of Bayes formula to render explicit the posterior distribution

To render explicit $p(\eta \mid y)$, we use Bayes theorem, $p(\eta \mid y)\, p(y) = p(y \mid \eta)\, p(\eta)$, to write

$$p(\eta \mid y) \propto p(y \mid \eta)\, p(\eta), \tag{7}$$

that is *posterior distribution = likelihood* $\times$ *prior*.
The constant $p(y)$ will be ignored since $r_{\eta,\eta'}$ deals with a ratio of posterior probabilities.

4.2 Step 2 - Analytical formulation of prior $p(\eta)$

$$\begin{aligned} p(\eta) &= p(k, s, \mathcal{C}, \mathcal{M}, z, X, a, \sigma) \\ &= p(k)\, p(\mathcal{C} \mid k)\, p(s \mid k, \mathcal{C}) \prod_{i \in \mathcal{T}} p(z^i, X^i, a^i, \sigma^i, \mathcal{M}^i \mid \mathcal{C}^i) \end{aligned} \tag{8}$$

In Equality 8, the term $p(s \mid k, \mathcal{C})$ is fully determined. We recall that the total number k of co-mRNAs (all targets considered) is assumed to follow an *a priori* Poisson distribution with mean λ (see Expression 4). Conditional on number k, the vector $\mathcal{C}$ of (active) co-mRNAs is drawn with the following prior uniform distribution:

$$p(\mathcal{C} \mid k) = \frac{1}{\mathcal{C}^k_{\sum_{i \in \mathcal{T}} q^i}} = \frac{1}{\mathcal{C}^k_q}. \tag{9}$$

As s is fully determined conditional on k and $\mathcal{C}$, probability $p(s \mid k, \mathcal{C})$ is equal to 1.

We now concentrate on the evaluation of the term $\prod_{i \in \mathcal{T}} p(z^i, X^i, a^i, \sigma^i, \mathcal{M}^i \mid \mathcal{C}^i)$:

$$\prod_{i\in\mathcal{T}} p(z^i, X^i, a^i, \sigma^i, \mathcal{M}^i \mid \mathcal{C}^i) \tag{10}$$

$$= \prod_{i\in\mathcal{T}} \prod_{l=1}^{s^i} p(z^{i\ell}, X^{i\ell}, a^{i\ell}, \sigma^{i\ell}, \mathcal{M}^{i\ell} \mid \mathcal{C}^{i\ell})$$

$$= \prod_{i\in\mathcal{T}} \prod_{l=1}^{s^i} p(z^{i\ell}, X^{i\ell}, a^{i\ell} \mid \sigma^{i\ell}, \mathcal{M}^{i\ell}, \mathcal{C}^{i\ell})\, p(\sigma^{i\ell})$$

$$= \prod_{i\in\mathcal{T}} \prod_{l=1}^{s^i} p(X^{i\ell} \mid \mathcal{C}^{i\ell})\, p(z^{i\ell} \mid X^{i\ell})\, p(a^{i\ell} \mid z^{i\ell}, X^{i\ell}, \sigma^{i\ell})\, p(\sigma^{i\ell}). \tag{11}$$

First right-hand term $p(X^{i\ell} \mid \mathcal{C}^{i\ell})$ follows a uniform distribution ($\frac{1}{m^{\mathcal{C}^{i\ell}}}$). Second right-hand term, $p(z^{i\ell} \mid X^{i\ell})$ is equal to 1 since $X^{i\ell}$ fully determines $z^{i\ell}$ ($z^{i\ell} =\mid X^{i\ell} \mid$). Expression 10 is more conveniently written as:

$$\prod_{i\in\mathcal{T}} p(z^i, X^i, a^i, \sigma^i, \mathcal{M}^i \mid \mathcal{C}^i) = \prod_{i\in\mathcal{T}} p(a^i \mid z^i, X^i, \sigma^i)\, p(\sigma^i) \left(\prod_{i\in\mathcal{T}} \prod_{l=1}^{s^i} \frac{1}{m^{\mathcal{C}^{i\ell}}} \right). \tag{12}$$

Expression 12 exhibits the two terms $p(a^i \mid z^i, X^i, \sigma^i)$ and $p(\sigma^i)$. To propose priors for a^i and σ^i, we rely on the specific scheme proposed by Andrieu and Doucet. In their precursor works on RJMCMCs based on a multivariate linear model, the regression coefficients are assumed to follow a Gaussian distribution, conditional on predictor set X^i. Before we give the Gaussian distribution, we need the following definitions:

Definition 4.1
Given X^i, the whole set of predictors associated with target i, $D_{X^i}(x)$ is the matrix defined over $\{0,\ 1,\ 2\}$, of dimension $p \times \left((\sum_{\ell=1}^{s^i} z^{i\ell}) + 1\right)$, where p is the number of subjects observed. First column is a vector of 1s while each cell $D_{X_i}(x)_{o,j+1}$ describes j^{th}regressor of set X^i, for subject o. We define matrix Σ_{X^i} as: ${\Sigma_{X^i}}^{-1} = D^T_{X^i}(x)\ D_{X^i}(x)$.

Definition 4.2

$$\Sigma_{X^i} = D^T_{X^i}(x)\ D_{X_i}(x). \tag{13}$$

Under the zero-mean Gaussian assumption with covariance $(\sigma^i)^2\ \Sigma_{X^i}$, we write:

$$p(a^i \mid z^i, X^i, \sigma^i) =\mid 2\pi(\sigma^i)^2\Sigma_{X^i} \mid^{-1/2} \exp\left(-\frac{a^T_{X^i}\Sigma^{-1}_{X^i}a_{X^i}}{2(\sigma^i)^2} \right). \tag{14}$$

Regarding scale parameter σ^i, as $p(\sigma^i)$ and $p(\sigma^i)^2$ are equal, the prior distribution of variable $(\sigma^i)^2$ is given instead. $(\sigma^i)^2$ is assumed to follow a conjugate inverse-Gamma law:

$$(\sigma^i)^2 \sim \mathcal{IG}(\upsilon_0/2, \gamma_0/2). \tag{15}$$

Andrieu and Doucet recommend to choose $(\upsilon_0/2, \gamma_0/2) = (0, 0)$, to obtain Jeffrey's uninformative prior $p((\sigma^i)^2) \propto \frac{1}{(\sigma^i)^2}$.

Thanks to equations 8, 4, 9, 12, 15 and 14, expression $p(\eta)$ is now entirely explicit.

4.3 Step 3 - Analytical formulation of likelihood $p(y \mid \eta)$

The likelihood is expressed as:

$$p(y \mid \eta) = \prod_{i \in \mathcal{T}} p(y^i \mid \eta). \tag{16}$$

Since a linear model is assumed ($y^i - D_{X^i}(x)\, a_{X^i} = \varepsilon_i$), we now state that the noise ε_i is zero-mean Gaussian with variance $(\sigma^i)^2$:

$$p(y_i \mid \eta) = \frac{1}{(2\pi\ (\sigma^i)^2)^{p/2}} \exp\left[-\frac{(y^i - D_{X^i}(x)\ a_{X^i})^T\ (y^i - D_{X^i}(x)\ a_{X^i})}{2(\sigma^i)^2} \right]. \tag{17}$$

4.4 Step 4 - Temporary algebraic expression for posterior distribution $p(\eta \mid y)$

Combining the explicit derivations for $p(\eta)$ (see step 2, Equality8) and $p(y \mid \eta)$ (see step 3, Equalities 16 and 17) in Equality 7, we obtain a temporary algebraical expression of the joint posterior distribution $p(\eta \mid y)$:

$$\begin{aligned} p(\eta \mid y) \propto & \frac{\lambda^k}{k!} \frac{1}{\mathcal{C}_q^k} \left(\prod_{i \in \mathcal{T}} \prod_{l=1}^{s^i} \frac{1}{m^{\mathcal{C}^{il}}} \right) \\ & \prod_{i \in \mathcal{T}} \mid 2\pi(\sigma^i)^2 \Sigma_{X^i} \mid^{-1/2} \exp\left[-\frac{a_{X^i}^T \Sigma_{X^i}^{-1} a_{X^i}}{2(\sigma^i)^2} \right] \frac{\frac{\gamma_0}{2}^{\frac{\upsilon_0}{2}}}{\Gamma(\frac{\upsilon_0}{2})} ((\sigma^i)^2)^{-\frac{\upsilon_0}{2}-1} \\ & \exp\left[-\frac{\frac{\gamma_0}{2}}{(\sigma^i)^2} \right] \frac{1}{(2\pi\ (\sigma^i)^2)^{p/2}} \exp\left[-\frac{(y^i\ D_{X^i}(x)\ a_{X^i})^T\ (y^i\ D_{X^i}(x)\ a_{X^i})}{2(\sigma^i)^2} \right]. \end{aligned} \tag{18}$$

However, the obtained expression does not straightforwardly lend itself to integration with respect to a and σ parameters.

4.5 Step 5 - Definite algebraic expression for posterior distribution $p(\eta \mid y)$

Before we perform the integration process, we need to transform Equality 18 into a more appropriate algebraic expression. The transformation process applied by Adrieu and Doucet in their work on the detection of noisy sinusoidal signals was not much detailed. However, three clues allowed us to derive again the transformation process. The three clues consisted of three definitions, which were as many starting points to guess the derivation. In our specific case, these definitions write:

$$P^i = I - D_{X^i(x)}\ M^i\ D_{X^i(x)}^T \tag{19}$$

$$M^i = \frac{\delta^2}{\delta^2 + 1} \left(D_{X^i(x)}^T\ D_{X^i(x)} \right)^{-1} \tag{20}$$

$$d^i = M^i\ D_{X^i(x)}^T\ y^i. \tag{21}$$

Appendix 1 reports the steps of the process providing the final expression to be marginalized:

$$p(\eta \mid y) \propto \frac{\lambda^k}{k!}\,\frac{1}{\mathcal{C}_q^k}\,\left(\prod_{i\in\mathcal{T}}\prod_{l=1}^{s^i}\frac{1}{m^{\mathcal{C}^{i\ell}}}\right)$$
$$\prod_{i\in\mathcal{T}}\left\{\mid 2\pi(\sigma^i)^2\Sigma_{X^i}\mid^{-1/2}\,\exp\left[-\frac{(a_{X^i}-d^i)^T(M^i)^{-1}(a_{X^i}-d^i)}{2(\sigma^i)^2}\right]\exp\left[-\frac{\gamma_0+(y^i)^TPy^i}{2(\sigma^i)^2}\right]\,(\sigma^{i\,2})^{-\frac{\upsilon_0}{2}-1-\frac{p}{2}}\right\}. \tag{22}$$

4.6 Step 6 - Integration of the nuisance parameters

The expression obtained for $p(\eta \mid y)$ through step 5 is now convenient for elimination of a and σ parameters. From the expression of $p(\eta \mid y)$, we will thus obtain an expression of the posterior distribution $p(\zeta \mid y)$. In this case, the nuisance parameters can be eliminated on a theoretical basis, through integral calculus. Note that the MCMC is not used for this purpose.

Appendix 2 details the integration process. Finally, we obtain:

$$p(\zeta \mid y) \propto \frac{\lambda^k}{k!}\,\frac{1}{\mathcal{C}_q^k}\,\left(\prod_{i\in\mathcal{T}}\prod_{l=1}^{s^i}\frac{1}{m^{\mathcal{C}^{i\ell}}}\right)\,\left[\frac{(\frac{\gamma_0}{2})^{\frac{\upsilon_0}{2}}}{\Gamma(\frac{\upsilon_0}{2})}\,\Gamma(\frac{\upsilon_0+p}{2})\right]^{\tau}\,\frac{1}{(2\pi)^{\frac{\tau p}{2}}}\,\prod_{i\in\mathcal{T}}\mathcal{P}(target\ i),$$
$$\text{with } \mathcal{P}(target\ i) = \left(\frac{\gamma_0+(y^i)^TP^iy^i}{2}\right)^{-\frac{\upsilon_0+p}{2}}. \tag{23}$$

5 Calculation of acceptance probabilities

5.1 Addition of a co-mRNA (move A)

At line 4 in Algorithm 4, the drawn multi-loci pattern will determine the set of predictors to be added to the current predictor set of some target $i*$, if move A is further accepted (see line 7). In line 9, the new co-mRNA $\ell*$ is taken into account ($k^+ = k+1$; ${s^{i*}}^+ = s^{i*}+1$; ${\mathcal{C}^{i*}}^+ = \mathcal{C}^{i*} \oplus \ell*$; ${\mathcal{M}^{i*}}^+ = \mathcal{M}^{i*} \oplus m*$); ${z^{i*\ell*}}^+ =\mid \mathcal{M}^{i*\ell*} \mid$), where $\oplus$ designates the appending operation and $m*$ is the new multi-locus pattern. The predictor set ${X^{i*}}^+$ is computed as X^{i*} augmented with the variables corresponding to the SNPs of $m*$. The set of regression coefficients a^{i*} is augmented accordingly, to yield ${a^{i*}}^+$.

The acceptance probability for move A is:
$\alpha(\zeta,\zeta^+) = min(1, r(\zeta,\zeta^+))$, where $r(\zeta,\zeta^+)$ (indeed $r_{k,k+1}(\zeta,\zeta^+)$) writes as:

$$r(\zeta,\zeta^+) = \frac{p(\zeta^+ \mid y)}{p(\zeta \mid y)}\,\frac{d_{k+1}}{a_k}\,\frac{q(\zeta \mid \zeta^+)}{q(\zeta^+ \mid \zeta)}. \tag{24}$$

Following Equation 23, the terms appearing in $r(\zeta,\zeta^+)$ are respectively evaluated as:

$$p(\zeta^+ \mid y) = \frac{\lambda^{k+1}}{(k+1)!}\,\frac{1}{C_q^{k+1}}\,\left(\prod_{i\in\tau}\prod_{\ell=1}^{s^i}\frac{1}{m^{\mathcal{C}^{i\ell}}}\right)\,\mathcal{P}(\textit{target } i, \textit{co-mRNA } \ell*, \textit{multi-loci pattern } m*) \tag{25}$$

Algorithm 4 proposeMoveA

if $(q - k > 0)$
  propose a new co-mRNA $\ell*$ uniformly drawn in the $q - k$ current non active co-mRNAs.
  /* The drawn co-mRNA corresponds to a given target $i*$ */
  propose a multi-loci pattern $m*$ uniformly drawn in $M^{i*\ell*}$
  compute α_{ζ,ζ^+}
  sample $u \sim U_{[0,1]}$
  if $(u \leq \alpha_{\zeta,\zeta^+})$
    sample σ^{i*+} and a^{i*+}
    modify the Markov chain state into $(k+1, s^+, \mathcal{C}^+, \mathcal{M}^+, z^+, X^+, a^+, \sigma^+)$
  else
    keep the chain in state $(k, s, \mathcal{C}, \mathcal{M}, z, X, a, \sigma)$
  end if
end if

$$p(\zeta \mid y) = \frac{\lambda^k}{k!} \frac{1}{C_q^k} \left(\prod_{i \in \tau} \prod_{\ell=1}^{s^i} \frac{1}{m^{c^{i\ell}}} \right) \mathcal{P}(\textit{target } i). \tag{26}$$

From Equalities 25 and 26, we derive the posterior distribution ratio:

$\frac{p(\zeta^+|y)}{p(\zeta|y)} = \frac{\lambda}{q-k} \frac{1}{m^{i*\ell*}} \frac{\mathcal{P}(\textit{target } i*, \textit{transcript } \ell*, \textit{multi-locus pattern } m*)}{\mathcal{P}(\textit{target } i*)}$.

The first term in the proposal ratio, $\frac{d_{k+1}}{a_k} \frac{q(\zeta|\zeta^+)}{q(\zeta^+|\zeta)}$, simplifies in:

$$\frac{d_{k+1}}{a_k} = \frac{k+1}{\lambda}, \tag{27}$$

whereas the second term is calculated as shown below.

Given $k - q$, the total number of inactive co-mRNAs over all targets, we easily derive:

$$q(\zeta^+ \mid \zeta) = \frac{1}{q-k}. \tag{28}$$

However, the calculation of $q(\zeta \mid \zeta^+)$ is not so straighforward; it involves that of term $q_d(\zeta^+)$:

$$q(\zeta \mid \zeta^+) = \frac{1}{q_d(\zeta^+)}, \tag{29}$$

referring to definition 5.1.

Definition 5.1 *For state ζ, $q_d(\zeta)$ is the number of active co-mRNAs, over all targets showing a number of active co-mRNAs strictly greater than* 1.

$q_d(\zeta^+)$ is derived from $q_d(\zeta)$. Figure 3 shows how q_d decreases by one from state ζ^+ to state ζ, on a simple case. Nevertheless, another (unique) case has to be considered, as depicted in Figure 4: therein, the active co-mRNA of state ζ^+, whose dismissing yields state ζ, is one of the two active co-mRNAs of a target. This entails that the remaining active co-mRNA of this target can no more contribute to $q_d(\zeta)$ calculation. Thus, the following property holds:

Property 5.1

$$q_d(\zeta^+) = \begin{cases} q_d(\zeta) + 2 & \textit{if the co-mRNA is added to a target with only one active co-mRNA} \\ q_d(\zeta) + 1 & \textit{otherwise.} \end{cases}$$

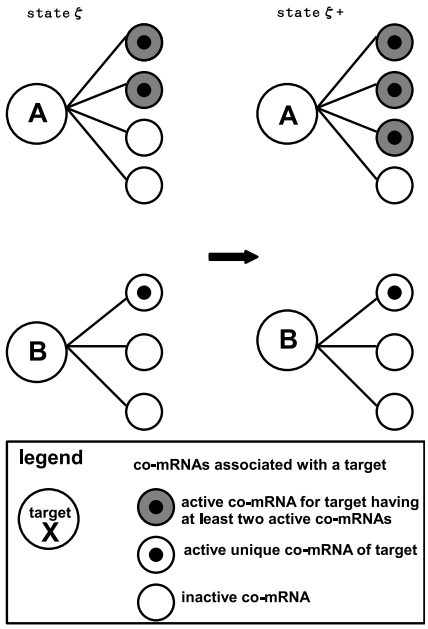

Figure 3: Increase of q_d between states ζ and ζ^+, in a simple case. See text, Definition 5.1. A and B denote targets. q: number of possible co-mRNAs over all targets; k: number of active co-mRNAs over all targets; $q = 7$; $k(\zeta) = 3$; $k(\zeta^+) = 4$; $q_d(\zeta) = 2$; $q_d(\zeta^+) = 3$.

From Equalities 28 and 29, the second term in the proposal ratio writes as:

$$\frac{q(\zeta \mid \zeta^+)}{q(\zeta^+ \mid \zeta)} = \frac{q - k}{q_d(\zeta^+)}. \tag{30}$$

Finally, in Equation 24 we substitute the expressions respectively obtained in 5.1, 27 and 30, respectively for posterior distribution ratio, first and second terms of proposal ratio. Therefore, the definite formula derived for $r(\zeta, \zeta^+)$ is the following:

$$r(\zeta, \zeta^+) = \frac{k+1}{q_d(\zeta^+)} \; \frac{1}{m^{i\ell *}} \; \frac{\mathcal{P}(\textit{target } i*, \textit{co-mRNA } \ell*, \textit{multi-locus pattern } m*)}{\mathcal{P}(\textit{target i*})}. \tag{31}$$

In move A, parameter a has to be sampled. This sampling depends on parameter σ. Finally, both parameters are sampled. From equation 22, it can be easily derived that:

$$(\sigma^i)^2 \mid y^i \zeta \sim \mathcal{IG}(\frac{\upsilon_0 + p}{2}, \frac{\gamma_0 + {y^i}^T \, P \, y^i)}{2} \tag{32}$$

and

$$a_{X^i} \mid y^i, \zeta, \sigma^i \sim \mathcal{N}(M^i \, D_{X^i}^T \, y^i, (\sigma^i)^2 \, M^i). \tag{33}$$

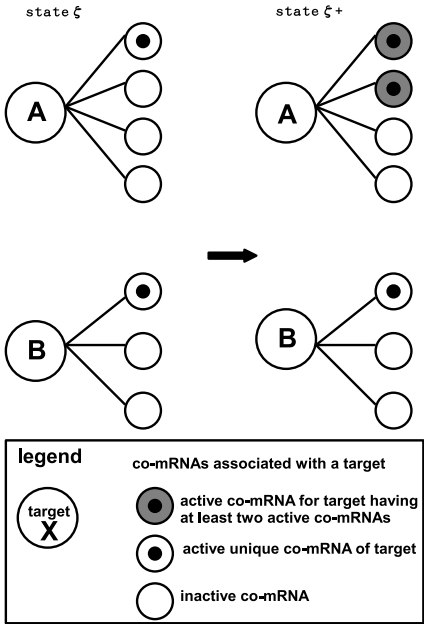

Figure 4: Increase of q_d between states ζ and ζ^+, in the unique alternative of the case presented Figure 3. $q = 7$; $k(\zeta) = 2$; $k(\zeta^+) = 3$; $q_d(\zeta) = 0$; $q_d(\zeta^+) = 2$.

5.2 Deletion of a co-mRNA (move D)

Quite symmetrically with respect to move A, denoting $i*$ the target for which a co-mRNA $\ell*$ is dismissed, and using Equation 31, we obtain:

$$r(\zeta, \zeta^-) = r_{k,k-1}(\zeta, \zeta^-) = \left(r_{k-1,k}(\zeta^-, \zeta)\right)^{-1} = \frac{q_d(\zeta)}{k}\, m^{i\ell*}\, \frac{\mathcal{P}(\textit{target } i)}{\mathcal{P}(\textit{target } i*, \textit{co-mRNA } \ell*, \textit{multi-locus pattern } m*)}, \quad (34)$$

where $m*$ denotes the active multi-locus pattern associated with co-mRNA $\ell*$. The pseudo-code of this move is depicted in Algorithm 5.

Algorithm 5 proposeMoveD

if $(q_d(\zeta) > 0)$
  propose a co-mRNA ℓ^* uniformly drawn amongst the $q_d(\zeta)$ current active co-mRNAs.
  /* The drawn co-mRNA corresponds to a given target $i*$ */
  compute α_{ζ,ζ^-}
  sample $u \sim U_{[0,1]}$
  if $(u \leq \alpha_{\zeta,\zeta^-})$
    sample a^{i*^-}
    modify the Markov chain state into $(k-1, s^-, \mathcal{C}^-, \mathcal{M}^-, z^-, X^-, a^-, \sigma)$
  else
    keep the chain in state $(k, s, \mathcal{C}, \mathcal{M}, z, X, a, \sigma)$
end if

5.3 Substitution of a co-mRNA (Move C)

Move C implements the replacement of an active co-mRNA ℓ with another active co-mRNA $\ell*$, for a given target $i*$. m is the active multi-locus pattern associated with replaced co-mRNA ℓ, while $m*$ denotes the active multi-locus pattern associated with replacing co-mRNA $\ell*$. Move C is typically a Metropolis update, meaning that the term $r(\zeta, \zeta^*)$, indeed $r_k(\zeta, \zeta^*)$, merely writes as follows:

$$r(\zeta, \zeta^*) = \frac{p(\zeta^* \mid y)}{p(\zeta \mid y)} \frac{q(\zeta \mid \zeta^*)}{q(\zeta^* \mid \zeta)}. \tag{35}$$

From Equation 23, the posterior distribution ratio is easily derived in:

$$\frac{p(\zeta^* \mid y)}{p(\zeta \mid y)} = \frac{m^{i\ell}}{m^{i*\ell*}} \frac{\mathcal{P}(\textit{target } i, \textit{co-mRNA } \ell*, \textit{multi-locus pattern } m*)}{\mathcal{P}(\textit{target } i*, \textit{co-mRNA } \ell, \textit{multi-locus pattern } m)}. \tag{36}$$

To evaluate the terms involved in the proposal ratio ($\frac{q(\zeta|\zeta^*)}{q(\zeta^*|\zeta)}$), we need define a new term, q_c:

Definition 5.2 *For state ζ, $q_c(\zeta)$ is the number of active co-mRNAs, over all targets showing a number of possible co-mRNAs strictly greater than their number of active co-mRNAs.*
In other terms, each such target t is characterized by $q^t - s^t > 0$.

This definition entails the following property:

Property 5.2 *Since each target must always present at least one active co-mRNA in any state, move C only concerns a target t possessing at least two possible co-mRNAs ($q^t > 1$).*

To calculate $q(\zeta^* \mid \zeta)$, we first have to uniformly draw the replaced active co-mRNA amongst the $q_c(\zeta)$ co-mRNAs allowed. Now knowing the target $i*$ which is subject to the replacement of one of its active co-mRNAs (ℓ), we uniformly draw the replacing active co-mRNA amongst $q^{i*}(\zeta) - s^{i*}(\zeta)$ valid candidates. Therefore, the denominator of the proposal ratio ($q(\zeta^* \mid \zeta)$) writes as:

$$q(\zeta^* \mid \zeta) = \frac{1}{q_c(\zeta)} \frac{1}{q^{i*}(\zeta) - s^{i*}(\zeta)}. \tag{37}$$

Similarly, the numerator of the proposal ratio expresses as:

$$q(\zeta \mid \zeta^*) = \frac{1}{q_c(\zeta^*)} \frac{1}{q^{i*}(\zeta^*) - s^{i*}(\zeta^*)}. \tag{38}$$

The strong constraint implied by Property 5.2 indicates that there are not many cases to be studied to establish the straightforward relationship between $q_c(\zeta)$ and $q_c(\zeta^*)$:

Property 5.3 $q_c(\zeta^*) = q_c(\zeta)$.

Figure 5 allows a quick understanding of Property 5.3.

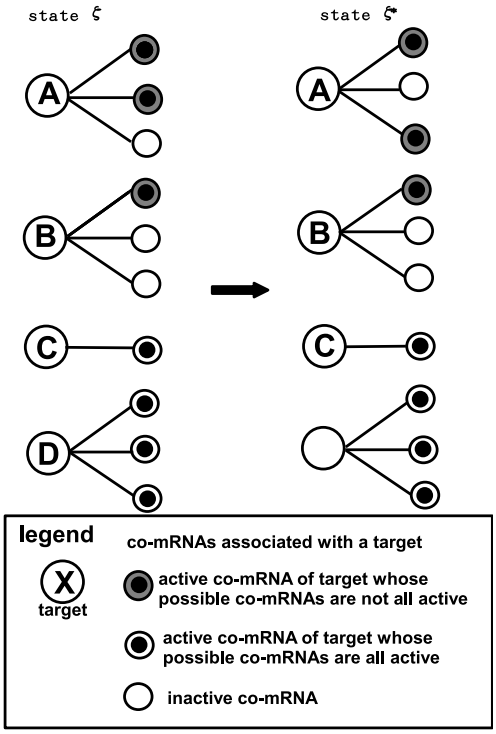

Figure 5: Relationship between $q_c(\zeta)$ and $q_c(\zeta^*)$, for states ζ and ζ^*, corresponding to move C, unique possible case. See text, Definition 5.2 and Property 5.2. A, B, C and D denote targets. q: number of possible co-mRNAs over all targets; k: number of active co-mRNAs over all targets; $q = 10$; $k(\zeta) = k(\zeta^*) = 7$; $q_c(\zeta) = q_c(\zeta^*) = 3$.

Besides, in the case of move C, $s^{i*}(\zeta^*)$ and $s^{i*}(\zeta)$ are equal. Together with Property 5.3, this equality entails that:

$$\frac{q(\zeta \mid \zeta^*)}{q(\zeta^* \mid \zeta)} = 1. \tag{39}$$

Therefore, using Equation 36, $r(\zeta, \zeta^*)$ is merely computed as:

$$r(\zeta, \zeta^*) = \frac{p(\zeta^* \mid y)}{p(\zeta \mid y)}. \tag{40}$$

The scheme of this move is depicted in Algorithm 6. At line 2, the replaced co-mRNA is determined (ℓ). ℓ determines a target i. Conditional on target i, a replacing co-mRNA is drawn at line 4. Conditional on the target and the replacing co-mRNA, a multi-locus pattern is drawn at random ($m*$). If the move is accepted (see line 8), the set of predictors corresponding to this multi-locus pattern will be added to the current predictor set of target i. The set of predictors corresponding to the replaced co-mRNA will be dismissed from target i's predictors. In line 10, the co-mRNA substitution from ℓ into $\ell*$ is

acknowledged: $\mathcal{C}^{i^*} = \mathcal{C}^i \ominus \ell \oplus \ell*$; $\mathcal{M}^{i^*} = \mathcal{M}^i \ominus m \oplus m*$); $z^{i\ell*^*} =\mid \mathcal{M}^{i\ell*} \mid$). Operation $\ominus$ performs the deletion operation on a vector. m is the multi-locus pattern associated with replaced co-mRNA ℓ, whereas $m*$ refers to the multi-locus pattern connected to the replacing co-mRNA. The predictor set X^{i^*} is computed as X^i dismissed of the variables corresponding to the SNPs in m, and further augmented with the variables corresponding to the SNPs in $m*$. The set of regression coefficients a^i is accordingly updated into a^{i^*}.

Algorithm 6 proposeMoveC

if $(q_c(\zeta) > 0)$
  propose a replaced co-mRNA ℓ uniformly drawn in the $q_c(\zeta)$ current active co-mRNAs
  /* The drawn co-mRNA corresponds to a given target i */
  propose a replacing co-mRNA $i\ell*$ uniformy drawn in the $q^i - s^i$ non active co-mRNAs of target i
  propose a multi-locus pattern uniformly uniformly drawn in $M^{i\ell*}$
  compute $\alpha_{\zeta,\zeta'}$
  sample $u \sim U_{[0,1]}$
  if $(u \leq \alpha_{\zeta,\zeta'})$
    sample $a^{i\ell*}$
    modify the Markov chain state into $(k, s, \mathcal{C}^*, \mathcal{M}^*, z^*, X^*, a^*, \sigma^*)$
  else
    keep the chain in state $(k, s, \mathcal{C}, \mathcal{M}, z, X, a, \sigma)$
  end if
end if

5.4 Substitution of a multi-locus pattern

Again a Metropolis update, move M replaces the active multi-locus pattern m associated with an active co-mRNA ℓ of, say, target i. The co-mRNA ℓ remains active while the MCMC evolves from state ζ to state ζ^*. The only difference between both states lies in that multi-locus pattern $m*$ now replaces m.

The term $r(\zeta, \zeta^*)$ is defined as in Equation 36. Therein, since $m^{i\ell}$ and $m^{i\ell*}$ are obviously equal ($\ell* = \ell$), the posterior distribution ratio now simplifies to:

$$\frac{p(\zeta^* \mid y)}{p(\zeta \mid y)} = \frac{\mathcal{P}(\textit{target } i, \textit{co-mRNA } \ell, \textit{multi-locus pattern } m*)}{\mathcal{P}(\textit{target } i, \textit{co-mRNA } \ell, \textit{multi-locus pattern } m)}. \tag{41}$$

The calculation of proposal ratio ($\frac{q(\zeta|\zeta^*)}{q(\zeta^*|\zeta)}$) requires a new definition:

Definition 5.3 *For state ζ, $q_m(\zeta)$ is the number of active co-mRNAs, over all targets showing a number of possible multi-locus patterns strictly greater than* 1.
In other terms, each such co-mRNA r, associated with, say, target t, is characterized by $m^{tr} > 1$.

Given this new definition, the denominator of the proposal ratio straighforwardly writes as:

$$q(\zeta^* \mid \zeta) = \frac{1}{q_m(\zeta)} \frac{1}{m^{i\ell}(\zeta) - 1}. \tag{42}$$

The first term in Equation 42 coerces the choice of the active multi-locus pattern to be replaced: a uniform draw amongst the $q_m(\zeta)$ co-mRNAs allowed ensures the satisfaction of the required constraint. Now knowing that active co-mRNA ℓ associated with target i is concerned by move M, the replacing

co-mRNA is uniformly drawn amongst the $m^{i\ell} - 1$ valid candidates. Indeed, Definition 5.3 guarantees that m is not the unique multi-locus pattern possible for co-mRNA ℓ.

It is obvious that $q_m(\zeta^*)$ and $q_m(\zeta)$ are equal. Therefore, the proposal ratio simplifies to 1 and as for move C, $r(\zeta, \zeta^*)$ is expressed as in Equation 40. This time, $r(\zeta, \zeta^*)$ will be computed using Equation 41. This move is described in Algorithm 7.

At line 2, as in move C, an active co-mRNA is drawn at random amongst the $q_m(\zeta)$ co-mRNAs allowed. Then, at line 4, conditional on target i and active co-mRNA $\ell*$, a multi-locus pattern is uniformly drawn. If this latter candidate, say $m*$ is accepted (see line 7), the set of predictors described by this multi-locus pattern will replace the current contribution of active co-mRNA $\ell*$, in the predictor set of target i. Line 9 akcnowledges this multi-locus pattern substitution: the predictor set X^{i^*} is computed as X^i dismissed of the variables corresponding to m, subsequently augmented with the variables associated with the replacing multi-locus pattern, $m*$. The set of regression coefficients a^i is updated accordingly.

Algorithm 7 Substitute-multi-locus-pattern

1: **if** $(q_m(\zeta) > 0)$
2: propose co-mRNA $i * \ell$ uniformly drawn in the $q_m(\zeta)$ current active co-mRNAs
3: /* The drawn co-mRNA corresponds to a given target $i*$ */
4: propose a replacing multi-loci pattern $i * \ell * m*$ uniformy drawn in the $(m^{\mathcal{C}^{i*\ell*}} - 1)$ non active multi-loci patterns of active co-mRNA $i * \ell*$
5: compute $\alpha_{\zeta,\zeta'}$
6: sample $u \sim U_{[0,1]}$
7: **if** $(u \leq \alpha_{\zeta,\zeta'})$
8: sample $a^{i*\ell*}$
9: modify the Markov chain state into $(k, s, \mathcal{C}^{'}, \mathcal{M}^{'}, z^{'}, X^{'}, a^{'}, \sigma^{'})$
10: **else**
11: keep the chain in state $(k, s, \mathcal{C}, \mathcal{M}, z, X, a, \sigma)$
12: **end if**
13: **end if**

5.5 Modification of the regression coefficients (Move R)

The last move proposed is the modification of the regression coefficients. It is simultaneously performed for all targets.

The reversible jump MCMC presented here can be viewed as implementing three levels. The first level (moves A and D) entails a variation of parameter k. As an MH update of the Markov chain (k unchanged), move C still lies in the scope of first level since it entails a substitution for co-mRNAs. Again an MH update, move M introduces the second level of our RJMCMC: this finer level only deals with multi-locus patterns. Third level restrains to the change of regression coefficients. When dealing with multivariate linear models, only two-step RJMCMCs had been proposed before [2, 14]. In this latter case, the updating of the linear model is fine-grained and the addition or dismissing of a (single) predictor is performed in the second (inner) level. In contrast, our approach modifies the structure of the linear model in the two upper levels. In the third level, for every target i in turn, move R updates σ^i and the whole set of regression coefficients a_{X^i}, relying on the two distributions given in 43 and 44. These

distributions are derived from Equation 22.

$$(\sigma^i)^2 \mid y^i, \zeta \sim \mathcal{IG}(\frac{\upsilon_0 + p}{2}, \frac{\gamma_0 + {y^i}^T \, P \, y^i}{2}) \tag{43}$$

$$a_{X^i} \mid y^i, \zeta, \sigma^i \sim \mathcal{N}(M^i \, {D_{X^i}}^T \, y^i, (\sigma^i)^2 \, M^i). \tag{44}$$

Algorithm 8 describes move R.

Algorithm 8 moveR

1: **for** each target i
2: update σ^i
3: update a^i
4: **end for**
5: **end for**

Acknowledgement

The author is deeply grateful to S. Lèbre for helpful discussions. A. Philippe and N. Depauw have to be thanked for their precious help too.

References

1. Albrechtsen, A., Castella, S., Andersen, G., Hansen, T., Pedersen, O., Nielsen, R. 2007. A Bayesian multilocus association method: allowing for higher-order interaction in association studies. *Genetics*, **176**, 1197–1208.

2. Andrieu, C., Doucet, A. 1999. Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Transactions on Signal Processing*, **47**(10).

3. Chen, X., Liu, C.T., Zhang, M., Zhang, H. 2007. A forest-based approach to identifying gene and gene-gene interactions. *Proceedings of the National Academy of Sciences of the USA*, **104**(49), 19199–19203.

4. Cordell, H.J. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, **11**(20), 2463–2468.

5. Cordell, H.J. 2009. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, **10**, 392-404 doi:10.1038/nrg2579.

6. Culverhouse, R. 2007. The use of the Restricted Partition Method with case-control data. *Human Heredity*, **63**(2), 93–100, doi:10.1159/000099181.

7. Clayton, D., Leung, H.-T. 2007. An R package for analysis of whole-genome association studies. *Human Heredity*, **64**, 45–51, doi: 10.1159/000101422.

8. Green, P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**(4), 711–732.

9. Hahn, L.W., Ritchie, M.D., and Moore, J.H. 2003. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, **19**, 376–382.

10. Halperin, E., Stephan, D.A. 2009. SNP imputation in association studies. *Nature Biotechnology*, **27**, 349–351, doi:10.1038/nbt0409-349.

11. Heidema, A.G., Boer, J.M.A., Nagelkerke, N., Mariman, E.C.M., van der A, D.L., Feskens, E.J.M. 2006. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *BMC Genetics*, **7**(23), doi: 10.1186/1471-2156-7-23.

12. Kelemen, A., Vasilakos, A.V., Liang, Y. 2009. Computational intelligence for genetic association study in complex diseases: review of theory and applications. *International Journal of Computational Intelligence in Bioinformatics and Systems Biology*, **1**(1), 15–31.

13. Kristensen, V.N., Edvardsen, H., Tsalenko, A., Nordgard, S.H., Sørlie, T., Sharan, R., Vailaya, A., Ben-Dor, A., Lønning, P.E., Lien, S., Omholt, S., Syvänen, A.-C., Yakhini, Z., Børresen-Dale, A.-L. 2006. Genetic variation in putative regulatory loci controlling gene expression in breast cancer. *Proceedings of the National Academy of Sciences of the USA*, **103**, 7735–7740.

14. Lèbre, S. 2007. Stochastic process analysis for genomics and dynamic Bayesian networks inference. *PhD thesis*. University of Evry, France.

15. Manolio, T.A. 2010. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, **363**, 166–176.

16. Marchini, J., Donnelly, P., Cardon, L.R. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, **37**, 413–417, doi:10.1038/ng1537.

17. Moore, J.H., Andrews, P.C, Barney, N., White, B.C. 2008. Development and evaluation of an open-ended computational evolution system for the genetic analysis of susceptibility to common human diseases. *Lecture Notes in Computer Science*, **4973**, 129–140, doi: 10.1007/978-3-540-78757-0.

18. Moore, J.H., Barney, N., Tsai, C.-T., Chiang, F.-T., Gui, J., White, B.C. 2007. Symbolic modeling of epistasis. *Human Heredity*, **63**(2), 120–133, doi: 10.1159/000099184.

19. Mourad, R., Sinoquet, C., Leray, P. 2010. Learning hierarchical Bayesian networks for genome-wide association studies. *In Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT)*, Paris, France, 549–556.

20. Motsinger, A.A., Ritchie, M.D., Reif, D.M. 2007. Novel methods for detecting epistasis in pharmacogenomics studies. *Pharmacogenomics*, **8**(9),1229–41.

21. Nelson, M.R., Kardia, S.L., Ferrell, R.E., Sing, C.F. 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research.*, **11**, 458–470.

22. Park, M.Y., Hastie, T. 2007. Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9**, 30–50.

23. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–575.

24. Ritchie, M.D., Hahn, L.W., Roody, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**,138–47.

25. Ruczinski, I., Kooperberg, C., LeBlanc, M. 2004. Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications. *Journal of Multivariate Analysis*, **90**, 178–195.

26. Schwender, H., Ickstadt, K. 2008. Identification of SNP interactions using logic regression. *Biostatistics*, **9**(1),187–198, doi:10.1093/biostatistics/kxm024.

27. Stephens, M. and Balding, D.J. (2009) Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10, 681-690. doi:10.1038/nrg2615.

28. Suchard, M.A., Weiss, R.E., Dorman, K.S., Sinsheimer, J.S. 2003. Inferring spatial phylogenetic variation along nucleotide sequence: a multiple change-point model. *Journal of the American Statistical Association* **98**, 427–437.

29. Tang, W., Wu, X., Jiang, R., Li, Y. 2009. Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy. *PLoS Genet.*, **5**(5)e1000464. doi:10.1371/journal.pgen.1000464

30. Török, H.P., Glas, J., Endres, I., Tonenchi, L., Teshome, M.Y., Wetzke, M., Klein, W., Lohse, P., Ochsenkühn, T., Folwaczny, M., Göke, B., Folwaczny, C., Müller-Myhsok, B., Brand, S. 2009. Epistasis between toll-tike receptor-9 polymorphisms and variants in NOD2 and IL23R modulates susceptibility to Crohn's Disease. *The American journal of gastroenterology*, **104**, 7, 1723–1733.

31. Verzilli, C.J., Stallard, N., Whittaker, J.C. 2006. Bayesian graphical models for genomewide association studies. *American Journal of Human Genetics*, **79**(1), 100–112.

32. Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N.L., Yu, W. 2009. MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. *BMC Bioinformatics*, **10**, 13.

33. Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N.L.S., Weichuan, Y. 2010. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics*, **26**(1), 30–37 doi: 10.1093/bioinformatics/btp622.

34. Xu, Q., Jia, Y.-B., Zhang, B.Y., Zou, K., Tao, Y.-B., Wang, Y.-P., Qiang, B.-Q., Wu, G.-Y., Shen, Y., Ji, H.-K., Huang,Y., Sun, X.-Q., Ji, L., Li, Y.-D., Yuan, Y.-B., Shu, L., Yu, X., Shen, Y.-C., Yu, Y.-Q., Ju, G.-Z. 2006. Association study of an SNP combination pattern in the dopaminergic pathway in paranoid schizophrenia: a novel strategy for complex disorders. *Molecular Psychiatry*, **9**, 510–521, doi:10.1038/sj.mp.4001472.

35. Xu, Q., Jia, Y.-B., Zhang, B.Y., Zou, K., Tao, Y.-B., Wang, Y.-P., Qiang, B.-Q., Wu, G.-Y., Shen, Y., Ji, H.-K., Huang, Y., Sun, X.-Q., Ji, L., Li, Y.-D., Yuan, Y.-B., Shu, L., Yu, X., Shen, Y.-C., Yu, Y.-Q., Ju, G.-Z. 2006. Association study of an SNP combination pattern in the dopaminergic pathway in paranoid schizophrenia: a novel strategy for complex disorders. *Molecular Psychiatry*, **9** 510–521, doi:10.1038/sj.mp.4001472.

36. Zhang, X., Huang, S., Zou, F., Wang, W. 2010. TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, **26**(12), i217-i227 doi:10.1093/bioinformatics/btq186.

37. Zhang, Y., Liu, J.S. 2007. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, **39**, 1167–1173.

38. Zubenko, G.S., Hughes, H.B., Stiffler, J.S. 2001. D10S1423 identifies a susceptibility locus for Alzheimer's disease in a prospective, longitudinal, double-blind study of asymptomatic individuals. *Mol. Psychiatry*, **6**, 413–419.

Appendix 1 - Transformation of the algebraic expression of the posterior distribution

Given the three matrices

$$P^i = I - D_{X^i}(x)\, M^i\, D_{X^i}(x)^T,\;\; M^i = (D_{X^i}(x)^T\, D_{X^i}(x))^{-1},\;\; d^i = M^i\, D_{X^i}(x)^T\, y^i,$$

the transformation of

$$p(\eta \mid y) \propto \frac{\lambda^k}{k!}\,\frac{1}{\mathcal{C}_q^k}\left(\prod_{i\in\mathcal{T}}\prod_{l=1}^{s^i}\frac{1}{m^{\mathcal{C}^{il}}}\right) \prod_{i\in\mathcal{T}} \mid 2\pi(\sigma^i)^2\Sigma_{X^i}\mid^{-1/2} \exp\left[-\frac{a_{X^i}^T\Sigma_{X^i}^{-1}a_{X^i}}{2(\sigma^i)^2}\right] \frac{\frac{\gamma_0}{2}^{\frac{\upsilon_0}{2}}}{\Gamma(\frac{\upsilon_0}{2})}\,((\sigma^i)^2)^{-\frac{\upsilon_0}{2}-1} \exp\left[-\frac{\frac{\gamma_0}{2}}{(\sigma^i)^2}\right]\,\frac{1}{(2\pi\,(\sigma^i)^2)^{p/2}}\exp\left[-\frac{(y^i\,D_{X^i}(x)\,a_{X^i})^T\,(y^i\,D_{X^i}(x)\,a_{X^i})}{2(\sigma^i)^2}\right].$$

into

$$p(\eta \mid y) \propto \frac{\lambda^k}{k!}\,\frac{1}{\mathcal{C}_q^k}\left(\prod_{i\in\mathcal{T}}\prod_{l=1}^{s^i}\frac{1}{m^{\mathcal{C}^{il}}}\right) \prod_{i\in\mathcal{T}}\left\{\mid 2\pi(\sigma^i)^2\Sigma_{X^i}\mid^{-1/2}\;\exp\left[-\frac{(a_{X^i}-d^i)^T(M^i)^{-1}(a_{X^i}-d^i)}{2(\sigma^i)^2}\right]\exp\left[-\frac{\gamma_0+(y^i)^TPy^i}{2(\sigma^i)^2}\right]\;(\sigma^{i2})^{-\frac{\upsilon_0}{2}-1-\frac{p}{2}}\right\}.$$

requires that the following equality be true:

$$a_{X^i}^T\Sigma^{-1}a_{X^i} + (y^i - D_{X^i}(x)a_{X^i})^T(y^i - D_{X^i}(x)\,a_{X^i}) = (a_{X^i}-d^i)^T(M^i)^{-1}(a_{X^i}-d^i) + y^{iT}\,P^i\,y^i. \qquad (45)$$

In the sequel, for lisibility, we will drop the indexes:

$$a^T\Sigma^{-1}a + (y - Da)^T(y - Da) = (a-d)^T M^{-1}(a-d) + y^T\,P\,y.$$

We now derive the equality:

$$\begin{aligned}
&(a-d)^T\,M^{-1}\,(a-d) + y^T\,P\,y\\
=&(a - M\,D^T\,y)^T\,D^T\,D\,(a - M\,D^T\,y) + y^T\,P\,y\\
=&(a^T - y^T\,D\,M^T)D^T\,D\,(a - M\,D^T\,y) + y^T\,(I - D\,(D^T\,D)^{-1}\,D^T)\,y\\
=&a^T\,D^T\,D\,a - a^T\,D^T\,D\,M\,D^T\,y - y^T\,D\,M^T\,D^T\,D\,a + y^T\,D\,M^T\,D^T\,D\,M\,D^T\,y + y^T\,y - y^T\,D\,(D^TD)^{-1}\,D^T\,y\\
=&a^T\,D^T\,D\,a - a^T\,D^T\,D\,(D^TD)^{-1}\,D^T\,y - y^T\,D\,((D^T\,D)^{-1})^T\,D^T\,D\,a + y^T\,D\,((D^T\,D)^{-1})^T\,D^T\,D\,(D^T\,D)^{-1}\,D^T\,y + y^T\,y - y^T\,D\,(D^TD\\
=&a^T\,D^T\,D\,a - a^T\,D^T\,y - y^T\,D\,((D^T\,D)^T)^{-1}\,D^T\,D\,a + y^T\,D\,((D^TD)^T)^{-1}\,D^T\,D\,(D^TD)^{-1}\,D^T\,y + y^T\,y - y^T\,D\,(D^TD)^{-1}\,D^T\,y\\
=&a^T\,D^T\,D\,a - a^T\,D^T\,y - y^T\,D\,(D^T\,D)^{-1}\,D^T\,D\,a + y^T\,D\,(D^T\,D)^{-1}\,D^T\,D\,(D^T\,D)^{-1}\,D^T\,y + y^T\,y - y^T\,D\,(D^TD)^{-1}\,D^T\,y\\
=&a^T\,D^T\,D\,a - a^T\,D^T\,y - y^T\,D\,a + y^T\,D\,(D^T\,D)^{-1}\,D^T\,y + y^T\,y - y^T\,D\,(D^T\,D)^{-1}\,D^T\,y\\
=&a^T\,D^T\,D\,a - a^T\,D^T\,y - y^T\,D\,a + y^T\,y.
\end{aligned}$$

We recall that in Section 4.2, we defined Σ_{X^i}:

$$\Sigma_{X^i} = D_{X^i}^T(x)\ D_{X_i}(x).$$

More, concisely, this writes: $\Sigma = D^T\ D$.
Thus, we derive

$$\begin{aligned}
&(a-d)^T\ M^{-1}\ (a-d) + y^T\ P\ y \\
=&a^T\ \Sigma^{-1}\ a + a^T\ D^T\ D\ a - a^T\ D^T\ y - y^T\ D\ a + y^T\ y \\
=&a^T\ \Sigma^{-1}\ a + (a^T\ D^T - y^T)\ D\ a + (y^T - a^T\ D^T)\ y \\
=&a^T\ \Sigma^{-1}\ a + (y - Da)^T(y - Da)\square.
\end{aligned}$$

Appendix 2 - Marginalization over nuisance parameters a and σ

In formula 22 od Section 4.5, we identify two expressions related to two well-known probability density functions: that of a multidimensional normal distribution (for parameter a^i) and that of an inverse gamma distribution (for parameter $(\sigma^i)^2$). Therefore, marginalization over the nuisance parameters is straightforward, through integral calculus. The probability density function of the multidimensional distribution of parameters (d^i, $(\sigma^i)^2 M^i$) appears. Integrating on parameter a^i provides 1.

The inverse gamma distribution of parameters (α,β) is:

$$\mathcal{IG}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\ z^{-\alpha-1}\ \exp(\frac{\beta}{z}).$$

In formula 22, we recognise $z^{-\alpha-1}\ \exp(\frac{\beta}{z})$,

where $z = (\sigma^i)^2$, $\alpha = \frac{v_0+p}{2}$, $\beta = \frac{\gamma_0+{y^i}^T\ {P^i}\ y^i}{2}$. Integrating $z^{-\alpha-1}\ \exp(\frac{\beta}{z})$ on the domain of variation of z yields $\frac{\Gamma(\alpha)}{\beta^\alpha}$,

in our case: $\frac{\Gamma(\frac{v_0+p}{2})}{\left(\frac{\gamma_0+(y^i)^T\ P^i\ y^i}{2}\right)^{\frac{v_0+p}{2}}}$.

Bayesian multi-locus pattern selection and computation through reversible jump MCMC

Christine Sinoquet

Abstract

In the human genome, susceptibility to common diseases is likely to be determined by interactions between multiple genetic variants. We propose an innovative Bayesian method to tackle the challenging problem of multi-locus pattern selection in the case of quantitative phenotypes. For the first time, in this domain, a whole Bayesian theoretical framework has been defined to incorporate additional transcriptomic knowledge. Thus we fully integrate the relationships between phenotypes, transcripts (messenger RNAs) and genotypes. Within this framework, the relationship between the genetic variants and the quantitative phenotype is modeled through a multivariate linear model. The posterior distribution on the parameter space can not be estimated through direct calculus. Therefore we design an algorithm based on Markov Chain Monte Carlo (MCMC) methods. In our case, the number of putative transcripts involved in the disease is unknown. Moreover, this dimension parameter is not fixed. To cope with trans-dimensional moves, our sampler is designed as a reversible jump MCMC (RJMCMC). In this document, we establish the whole theoretical background necessary to design this specific RJMCMC.

LINA, Université de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3