Detection of Gene by Gene and Gene by Environment Interactions in Genome-Wide Association Studies (GWAS) through Bayesian Graphical Models

Laurent Briollais[1,2]

[1] *Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada*
[2] *Dalla Lana School of Public Health, University of Toronto, Toronto, Canada*

Résumé.

L'émergence des études d'association sur l'ensemble du génome (GWASs) dans le domaine de la génétique humaine constitut un effort sans précédent pour découvrir de nouveaux variants génétiques associés à des maladies et des traits complexes. Jusqu'à présent, il y eu 1461 GWASs repertoriés dans la base de données HuGE et plus de 9900 postions génomiques identifiées. Pourtant, la majeure partie de la contribution génétique sous-jacente à la plupart des maladies humaines communes reste pour l'essentiel inexpliquée et l'intérêt biologique de beaucoup de postions génomiques identifiées à partir des GWASs n'ont pas été caractérisées. La recherche d'inéractions Gène-Gène (GxG) et Gène-Environnement (GxE) dans le contexte des GWASs est une manière de mieux exploiter les découvertes des GWASs en intégrant des facteurs de risque épidémiologiques et d'autres gènes dans l'analyse. Malgré leur intérêt, ces analyses soulèvent de nombreux défis méthodologiques qui ont débouché sur très peu de développements spécifiques. Dans ce travail, notre but principal est de développer un cadre statistique général basé sur les modèles graphiques Bayésiens pour la détection d'inéractions Gène-Gène (GxG) et Gène-Environnement (GxE) dans les GWASs.

Mots-clés. Gène; Environnement; GWAS; Bayésien; Modèle Graphique; Cancer du Sein.

Abstract.

The emergence of Genome-Wide Association Studies (GWASs) in the field of human genetics is an unprecedented effort to discover new genetic variants associated with complex human diseases and complex traits. To date, there have been 1,461 GWASs reported in the HuGE database and more than 9,900 hits identified. Yet, the bulk of genetic contribution underlying most common human diseases remains largely unexplained and the biological relevance of many loci identified through GWASs have not been characterized. The search for Gene by Gene (GxG) and Gene by Environment (GxE) interactions in the context of GWAS is a way to leverage GWAS discoveries by integrating epidemiological risk factors and other genes into the analysis. Despite their interest, these analyses raise

many methodological challenges and has seen very limited specific developments. In this work, our main purpose is to develop a general statistical framework based on Bayesian graphical modeling for the detection of GxG and GxE interactions in GWASs.

Keywords. Gene; Environment; GWAS; Bayesian; Graphical Model; Breast Cancer.

1 Bayesian Graphical Model (BGM)

A graphical model is a family of probability distributions which are Markov with respect to a given graph G. A discrete graphical model is a graphical model where the random variables are discrete.

Let X be the random vector of interest. Suppose that a random sample $\mathbf{x} = (x_1, \ldots, x_n)$ of values of X has been observed. Let us assume that we choose to do a model search in the family of models $\mathcal{M}_1, \ldots, \mathcal{M}_k$. We write the models as

$$\mathcal{M}_j = \{p(x|\vartheta), \vartheta \in \Theta_j\}, j = 1, \ldots, k \tag{1}$$

where ϑ is a parameter in the parameter set Θ_j and $p(x|\vartheta)$ is a probability density function. In the particular case where the model is a graphical model, the parameter space is defined by the underlying graph G and we identify models $\mathcal{M}_j$ with their underlying graph G_j.

In a Bayesian framework we assume a prior probability $P(\mathcal{M}_j), j = 1, \ldots, k$ on the set of models $(\mathcal{M}_1, \ldots, \mathcal{M}_k)$ and a prior probability on the parameters ϑ, and want to derive the posterior model probabilities $P(\mathcal{M}_j|\mathbf{x})$ for each one of the models $\mathcal{M}_1, \ldots, \mathcal{M}_k$, that is, the conditional distribution of $\mathcal{M}_j$ given the data.

The Bayesian solution is to choose the model with the highest posterior probability. According to the Bayes' theorem, the posterior probability for $\mathcal{M}_j$ is

$$P(\mathcal{M}_j|\mathbf{x}) = \frac{P(\mathbf{x}|\mathcal{M}_j)P(\mathcal{M}_j)}{\sum_{i=1}^{k} P(\mathbf{x}|\mathcal{M}_i)P(\mathcal{M}_i)} \tag{2}$$

The term $\sum_{i=1}^{k} P(\mathbf{x}|\mathcal{M}_i)P(\mathcal{M}_i)$ in (2) is a constant. Therefore we can write

$$P(\mathcal{M}_j|\mathbf{x}) \propto \underbrace{P(\mathbf{x}|\mathcal{M}_j)}_{(the \quad Marginal \quad Likelihood)} \quad \underbrace{P(\mathcal{M}_j)}_{(the \quad Model \quad Prior)} \tag{3}$$

Suppose that the random vector of interest X consists of r binary factors $\{X_1, \ldots, X_r\}$, each having two levels: 0 and 1. Take N individuals, and classify them accordingly to the r binary factors in $V = \{1, 2, \ldots, r\}$, let $\mathcal{E}$ be the collection of all non empty subsets of $V$and $\mathcal{E}_0$ the collection of possible subsets of V including $\emptyset$, then the elements F in $\mathcal{E}_0$ are in 1-1 correspondence with the cells in the contingency table and we can use p_F to denote the cell probability

$$p_F = P(X_v = 1, v \in F, X_v = 0, v \notin F). \tag{4}$$

Similarly, we can use $n_F, F \in \mathcal{E}_0$ to denote the cell counts. We know that since N is fixed, the cell counts follow a multinomial distribution with the following distribution function

$$f((n), p) = \binom{N}{(n)} p_\emptyset^{N - \sum_{F \in \mathcal{E}} n_F} \prod_{F \in \mathcal{E}} p_F^{n_F} \tag{5}$$

For a given contingency table, the transformation from the cell counts $\{n_F\}$ to the marginal cell counts $\{y_F\}$ is a simple linear transformation with Jacobian equal to 1. If we assume a multinomial distribution for the cell counts, we readily obtain, up to a multiplicative constant, the following natural exponential family distribution for the marginal cell counts $y = (y_E, E \in \mathcal{E}_0)$ as a natural exponential family:

$$\begin{aligned} f(y; \theta, G) \quad &= \quad \exp\Big(\sum_{D \in \mathcal{D}} \theta_D y_D - N \log(1 + \sum_{E \in \mathcal{E}} \exp(\sum_{D \subseteq_G E} \theta_D))\Big) \\ &\quad \text{with } \theta = (\theta_D, D \in \mathcal{D}) \end{aligned} \tag{6}$$

The conjugate priors for distributions in the exponential family have density

$$\pi_G(\theta | s, \alpha) = I_G(s, \alpha)^{-1} \exp\{\sum_{D \in \mathcal{D}} \theta_D s_D - \alpha \log\Big(1 + \sum_{E \in \mathcal{E}} \exp(\sum_{D \subseteq_G E} \theta_D)\Big)\} \ , \tag{7}$$

for some $s = (s_D, D \in \mathcal{D}) \in \Re^{|\mathcal{D}|}$ and $\alpha \in \Re$, where $I_G(s, \alpha)$ is the normalizing constant.

2 Specification of the Prior

A method to construct hyperparameters of a proper prior $\pi_{\mathcal{D}}(\theta_{\mathcal{D}} | (s, \alpha))$ is to start with a fictive prior contingency table with all cell counts n_F positive, not necessarily integers. With α denoting the total count in the given fictive contingency table, y_D denoting the marginal cell counts, we can take as hyperparameters $\alpha = N$ and $s_D = y_D,\ D \in \mathcal{D}$. Lack of prior information can be expressed through what is sometimes called a flat prior by taking all the fictive cell entries to be equal and equal to $\frac{\alpha}{|\mathcal{I}|}$. We used this latter prior specification in our application to the breast cancer GWAS.

3 Posterior of a Model

The posterior of G is proportional to the ratio of the two normalizing constants:

$$P(G \mid Y) \propto I_G(y+s, n+\alpha)/I_G(s,\alpha). \tag{8}$$

For G decomposable, the prior $\pi(\theta|\alpha, s)$ is identical to the hyper Dirichlet (see Massam et al., 2009). It therefore follows that the normalizing constants I_G can be computed analytically when the graph G is decomposable. When G is non decomposable, I_G needs to be computed numerically.

4 Regression Induced by a Graphical Model

In GWAS, we are interested in modelling the response variable (i.e. case control status) as a function of the SNP variables. Let $Y = X_r$, $r \in V$ be a response variable and X_A, $A \subset V \backslash \{r\}$ be a set of explanatory variables (i.e. the SNPs). Denote by $(n)_{A\cup\{r\}}$ and $(n)_A$ the corresponding marginal tables. Here (n), $(n)_{A\cup\{r\}}$ and $(n)_A$ are cross-classifications involving X_V, $X_{A\cup\{r\}}$ and X_A, respectively. The connection between log-linear models and the regressions derived from them has been explored in Dobra et al. (2010). They showed that the marginal likelihood of the regression $[Y|X_A]$ can be expressed as the ratio between the marginal likelihoods of the saturated log-linear models for $(n)_{A\cup\{r\}}$ and $(n)_A$.

5 SNP selection in GWAS Setting

Let R denote a set of possible regression models. We associate with each candidate model $r \in R$ a neighbourhood $\text{nbd}(r) \subset R$. Any two models $r, r' \in R$ are connected through a path $r = r_1, r_2, ..., r_l = r'$ such that $r_j \in \text{nbd}(r_{j-1})$ for $j = 2, ..., l$. The neighbourhood of $r = Y|X_A$ is obtained by addition moves, deletion moves, and replacement moves. In an addition move, we individually include in A any variable in $V \backslash A$. In a deletion move, we individually delete any variable that belongs to A. For a replacement move, we individually replace any one variable in A with any one variable in $V \backslash A$. The first stage of the MOSS procedure is as follows.

We make use of a current list of regressions S that is updated during the search. Define

$$S(c) = \left\{ r \in S : P(r) \geq c \max_{r' \in R} P(r') \right\}$$

where $c \in (0,1)$. A regression $r \in S$ is called explored if all of its neighbours $r' \in \text{nbd}(r)$ have been visited.

1. Initialize a starting list of regressions S. For each $r \in S$, calculate and record its marginal likelihood $P(r)$. Mark r as unexplored.

2. Let L be the set of unexplored regressions in S. Sample an $r \in L$ according to probabilities proportional with $P(r)$ normalized within L. Mark r as unexplored.

3. For each $r' \in \text{nbd}(r)$, check if r' is currently in S. If it is not, evaluate and record its marginal likelihood $P(r')$. Eliminate the regressions $S \backslash S(c')$ for some pre-chosen value $0 < c' < c$.

4. With probability q eliminate from S the regressions in $S \backslash S(c)$.

5. If all the regressions in S are explored STOP. Otherwise return to step 2.

The role of the parameters c, c', and q is to limit the number of regressions that need to be visited to a manageable number. At the end of the first stage, we will have a set of top regressions each involving a small number of variables. The first stage uses saturated models and thus includes interactions between SNPs (Genes) and between SNPs and environmental factors.
The second stage is to search the space of BGMs to identify the most relevant interactions among the variables in each of the top regressions. By using the generalized hyper Dirichlet prior of Massam et al. (2009), the computations in both steps can be done efficiently. Once a set of promising log-linear models has been found (at the end of stage two), model averaging can be used to build a classifier for predicting the response. The efficacy of the classifier can be assessed using k-fold cross validation.

6 APPLICATION

We used the CGEM breast cancer data. The genome-wide association studies (GWAS) for breast cancer has been completed in the Nurses' Health Study (NHS) with nearly 550,000 SNPs genotyped. The analysis includes 1,145 individuals who developed breast cancer during the observational period and 1,142 age-matched individuals who did not develop breast cancer during the same time period. Both the genotype data and the pre-computed analyses based on the genotype data were retrieved from the following website (http://cgems.cancer.gov/). The project team has developed easy access to pre-computed results of the data including SNP frequencies and single SNP association analyses.

The BGM selected 12 SNPs with MAF varying between 0% and 42%. Three SNPs have a MAF lower than 5%. In general, frequentist approaches applied to GWAS would not be able to perform a test statistic for these SNPs. Interestingly, the BGM identifies an interactions between two SNPs, one in the gene FRMD4A and one in the gene FGFR2. The effects of these two SNPs are in opposite direction ($\beta = -1.25$).

7 Discussion

Our real application of BGM to a breast cancer GWAS data set confirms the interest of this approach and its relevance for genetic research. Many of the genetic associations we found, especially the GxG interactions, would not have been found by conventional approaches, which generally cannot evaluate multi-SNP models and rare variants.

Bibliographie

[1] Dobra, A. and Massam, H. (2010), The Mode Oriented Stochastic Search (MOSS) for Log-linear Models With Conjugate Priors. *Statistical Methodology.* 7, 240–253.

[2] Massam, H. and Liu, J. and Dobra, A. (2009), A Conjugate Prior for Discrete Hierarchical Log-linear Models. *Annals of Statistics.* 37, 3431–3467.