Gene selection heuristic algorithm for nutrigenomics studies

Damien D. Valour, Isabelle Hue, Bénédicte Grimard, B. B. Valour

Gene selection heuristic algorithm for nutrigenomics studies

D. Valour, I. Hue, B. Grimard and B. Valour

Physiol. Genomics 45:615-628, 2013. First published 30 April 2013;
doi: 10.1152/physiolgenomics.00139.2012

Physiol Genomics 45: 615–628, 2013.
First published April 30, 2013; doi:10.1152/physiolgenomics.00139.2012.

Gene selection heuristic algorithm for nutrigenomics studies

D. Valour,[1,2] I. Hue,[1,2] B. Grimard,[1,2] and B. Valour[3]

[1]*INRA, UMR 1198 Biologie du Développement et Reproduction, Jouy-en-Josas, France;* [2]*Université Paris-Est, ENVA, UMR1198, Maisons-Alfort, France; and* [3]*Pôle Supérieur de Bansac, Département de Mathématiques, Clermont-Ferrand, France*

Submitted 23 October 2012; accepted in final form 28 April 2013

Valour D, Hue I, Grimard B, Valour B. Gene selection heuristic algorithm for nutrigenomics studies. *Physiol Genomics* 45: 615–628, 2013. First published April 30, 2013; doi:10.1152/physiolgenomics.00139.2012.— Large datasets from -omics studies need to be deeply investigated. The aim of this paper is to provide a new method (LEM method) for the search of transcriptome and metabolome connections. The heuristic algorithm here described extends the classical canonical correlation analysis (CCA) to a high number of variables (without regularization) and combines well-conditioning and fast-computing in "R." Reduced CCA models are summarized in PageRank matrices, the product of which gives a stochastic matrix that resumes the self-avoiding walk covered by the algorithm. Then, a homogeneous Markov process applied to this stochastic matrix converges the probabilities of interconnection between genes, providing a selection of disjointed subsets of genes. This is an alternative to regularized generalized CCA for the determination of blocks within the structure matrix. Each gene subset is thus linked to the whole metabolic or clinical dataset that represents the biological phenotype of interest. Moreover, this selection process reaches the aim of biologists who often need small sets of genes for further validation or extended phenotyping. The algorithm is shown to work efficiently on three published datasets, resulting in meaningfully broadened gene networks.

canonical correlation analysis; local models determination; PageRank method; Markov and non-Markov chains; structure matrix

1 INTRODUCTION

With the increase of functional genomics studies that aim at identifying key factors affected by particular physiological contexts or treatments, the combination of different -omics techniques such as transcriptomics and metabolomics is becoming a popular manner to better understand the biology of complex systems. The underlying purpose of those integrative approaches is to find links between heterogeneous datasets to assess whether connecting them is a more powerful way to analyze such a large quantity of data and give them a fine-tuned biological sense at a higher level of understanding than using classical analyses.

Canonical correlation analysis (CCA) methods (13) have emerged as an efficient exploratory tool to correlate two datasets acquired on a same experimental unit. However, a key assumption in CCA is that the number of biological replicates has to be higher than the number of variables to correlate in each set. This assumption is never verified in costly experiments. Furthermore, CCA has been extended to generalized canonical correlation analysis (GCCA; Refs. 4, 5), to regularized canonical correlation analysis (RCCA, Refs. 11, 18, 32) and to regularized generalized CCA (RGCCA, Ref. 29) to overcome the limitations of the method (replicates numbers, vectors colinearity, more than two sets to correlate). Correlations analyses are particularly useful in nutrigenomics studies and RCCA, for example, was previously performed to link metabolism and transcriptome in mouse (21) and bovine work (31).

However, none of those methods fully answers mathematical and biological problems. For example, RGCCA reveals convergence and matrix conditioning issues because of a too global approach, which makes the results difficult to interpret. Moreover, CCA and its extended methods isolate highly correlated variables. This approach would be suitable for our purposes if the neglect of some variables was possible in the two datasets. However, in nutrigenomics studies the analysis is based on a huge transcriptomic dataset and a more restricted metabolic dataset in which parameters are biologically chosen on their relevance to represent the final phenotype of interest. This implies that metabolic variables should not be individualized but taken together since they describe a global physiological context that will be used to select for a few genes that explain the observed phenotype.

The purpose of this paper is to provide a new heuristic method that gives the linear links between two sets of column vectors in which each row is issued from measurements of an individual data. Applied to our examples the first dataset represents a set of genes; we hope that some of them will be able to give a linear explanation of the whole metabolic/clinical second set behavior. The algorithm developed here is designed to take into account the limited number of individuals by a multiple exploitation of reduced local CCA models, while the classical method would fail when applied to the entire set of genes.

The method consecutively selects restrictive subsets of vectors within the first dataset on the basis of their explicative levels, measuring the most efficient gene vectors to optimize models. By increasing a required significance level we force the process to explore the whole first dataset of gene vectors. The result is a number of nonredundant linear models. We describe here that such reduced models can be summarized in PageRank matrices (23). Thus, we obtain a stochastic matrix, product of the PageRank matrices, which is the result of the course of the algorithm.

We established as one of the main points of this paper that a Markov process applied to the previous stochastic matrix determines our gene selection method. Singularly, this process is built from a self-avoiding walk i.e., a non-Markov process. The sequence of powers of the Markov matrix gives local extended models (LEM) that take into account the persistence of some genes through several local models (connexity). The

Address for reprint requests and other correspondence: B. Valour, Pôle Supérieur de Bansac, Département de Mathématiques, F-63000, Clermont-Ferrand, France (e-mail: bernval1@gmail.com).

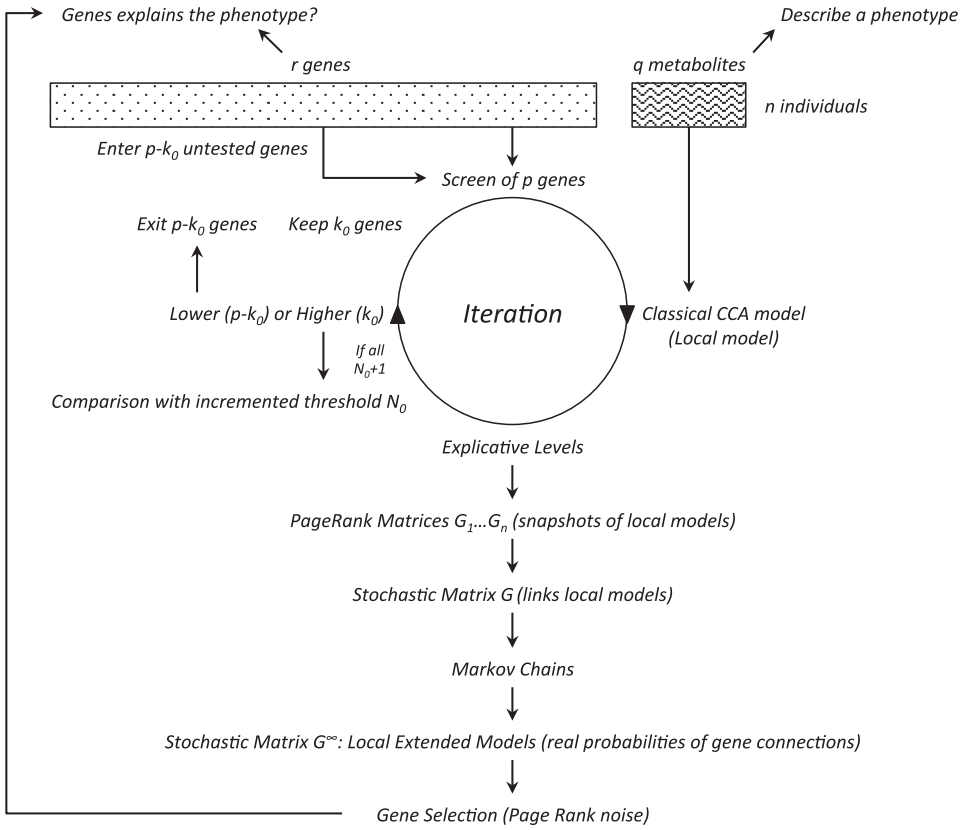

Fig. 1. Principle of the gene selection method.

principle of the gene selection method (or LEM method) is presented in Fig. 1.

So doing, we also propose a consistent method for structure matrix determination, while GCCA and RGCCA are unable to choose a structural partition into sets of genes and the proper weight for each of these partitions. In fact, the application of RGCCA block-wise or not (29) to our problems involves coefficients estimation of the linear combination of the whole first vector set, in the hope that some will be neglected because of their values close to zero. RGCCA transforms our mathematical problem into a continuous optimization problem solved by Gauss-Seidel-like methods and does not take into account many interesting local models.

As proof-of-principle, the current method has been applied to two nutrigenomics studies of similar complexities, one concerning the fertility of dairy cows (31) and the other the role of PPARα in the regulation of hepatic metabolism in the mouse (19). In addition, we extended the use of the method to a larger dataset dealing with the clinical effect of varying doses of acetaminophen on rats (3), evidencing in all cases reduced computing times and increased biological outputs. In this article we will *1*) develop the biological datasets used, *2*) describe the mathematics supporting the algorithm, and *3*) conclude on biological advantages brought by this new method. We also give an end-user "R" program (25).

2 BIOLOGICAL DATASETS

Datasets

The datasets analyzed with the gene selection algorithm are presented in Table 1. The two nutrigenomics datasets, referred to as NutriBov and NutriMous, were of similar complexities.

Table 1. *Datasets*

	NutriBov	NutriMous	LiverTox
Diet/Dose	n = 2, UF vs. CTRL	n = 5, COC, REF, SUN, LIN, FISH	n = 4, low (2), high (2)
Tissues	n = 3, OVI, ENDO, CL	n = 1, LIV	n = 1, LIV
Units	n = 12, cows 4/diet/tissue	n = 40, mice 20/genotype (2) 8/diet	n = 64, male rats
Variables	n = 2, genes (OVI: 293) blood metabolites (6)	n = 2, genes (120) hepatic fatty acids (21)	n = 2, genes (3,116) clinical measurements (10)
Studied correlation	genes/metabolites	genes/metabolites	genes/measurements
Data matrices	12 × 293 (genes) & 16 × 6 (metabolites)	40 × 120 (genes) & 40 × 21 (FA)	64 × 3,116 (genes) & 64 × 14 (measurements)
Methodology	RCCA	RCCA, sPLS	RCCA, sPLS
Reference list no.	31	12, 14, 19	12

RCCA, regularized canonical correlation analysis; sPLS, sparse partial least squares; UF, underfed; CTRL, control; COC, coconut oil diet; REF, reference diet; SUN, sunflower oil diet; LIN, linseed oil diet; FISH, fish oil diet; OVI, oviduct; ENDO, endometrium; CL, corpus luteum; LIV, liver; FA, fatty acid.

They respectively described the effect of diets (two to five) on tissues (one to three) of dairy cows or genotyped mice (PPARα−/− vs. wild type), while measuring blood metabolites or hepatic fatty acids and concomitantly evaluating gene expression changes in the dedicated tissues: oviduct, endometrium, corpus luteum (OVI, ENDO, CL) or liver (LIV). To add upon these first datasets where a limited number of physiological measurements (6–21 metabolites) were correlated to hundreds of genes (120–293), we extended our tests to a study (LiverTox) where 10 clinical measurements were concomitantly analyzed to many more genes (3,116). NutriMous and LiverTox datasets are publicly available in the "mixOmics" package (15). For the NutriBov dataset see (31).

RCCA and nonregularized GCCA (sparse partial least squares) analysis

So far, these datasets were analyzed by RCCA or sparse partial least squares (sPLS) (11, 12, 16, 31). In our case studies (see section 5), we re-analyzed the data with RCCA or sPLS too, evidencing gene/metabolite correlations for one half or one third of the physiological or clinical measurements (NutriBov, 3 out of 6 metabolites; NutriMous, 7/21; LiverTox, 7/14) and at best, half of the interesting gene changes (NutriBov, 151/293; NutriMous, 27/120; LiverTox, 1,032/3,116). These results depended on the selected correlation thresholds. In the published reports there were as follows: NutriBov: $r > 0.6$, NutriMous: $r > 0.5$; we choose in our study $r > 0.5$ for LiverTox.

RCCA and sPLS limitations

However, using those methods we observed several limitations. In practice, the results obtained by RCCA were highly sensitive to the quality of the determination of the regularization parameters. Those difficulties increased positively with the size of the datasets until the convergence of the method was not ensured in a reasonable time on the LiverTox dataset (section 5.4). The sPLS method is derived from the concept of RGCCA (29) and includes the convergence and matrix conditioning issues mentioned earlier. To overcome those limitations we present a heuristic algorithm (LEM method) that doesn't need regularization and combines fast computing in R software (25). This method also aims to maximize the number of metabolites/clinical measurements to correlate to gene expression, thus proposing a data-mining tool complementary to RCCA and sPLS, and biologically: new correlations and/or hypotheses.

3 GENE SELECTION ALGORITHM BASED ON CCA

3.1 Notations, Assumptions, and Objectives

Let us consider two matrices X_1 (size $n \times r$) and X_2 (size $n \times q$) respectively evaluating "gene expressions" and "metabolism patterns," n being the number of individuals and $r >> n > q$. Thus, we own a huge number of r genes to test according to q metabolic variables.

When the number of individuals is smaller than the number of variables (7) or in case of strong multi-colinearity in X_1 or X_2, nonregularized CCA sometimes fails. Then, the determination of shrinkage estimations and shrinkage constants is needed (17).

We describe in section 3.3, a heuristic method to extend CCA to r genes instead of regularization methods. We propose to test successive classical models of CCA including p genes ($p \leq q < n \ll r$). Now, the size of X_1 is $n \times p$.

Let us consider the classical model of canonical analysis using the concatenated

$$X = \underset{n\times p}{[X_1]}\underset{n\times q}{[X_2]}, \quad (1)$$

block-wise defined matrix [size $n \times (p + q)$], where we aim to find a linear relationship between column vectors of X_1 and column vectors of X_2.

The heuristic method allows the choice of a sufficiently small number p of variables for each performed CCA. Thus, we reasonably hope X_1 and X_2 to be full rank matrices [$\text{Rank}(X_1) = p$ and $\text{Rank}(X_2) = q$]. We also assume that the columns of X_1 and X_2 are centered and normalized by using D-metrics (here $D = \underset{nxn}{Id}$, Euclidian-metrics).

We want to extract, among $\binom{r}{p}$ potential CCA models, those giving a maximal explicative level (a notion to be defined hereafter).

3.2 Explicative Level of a CCA Model, Definition

We know (27) that the number of nonzero eigenvalues in a canonical analysis is less than, or equal to, $\text{Min}(p, q) = p$ (here $p \leq q$). The multiplicity order of the eigenvalue 1 is $\text{Dim}(W_1 \cap W_2)$, with:

$$W_1 = \left\{x \in \mathbb{R}^n / x = X_1 a, a \in \mathbb{R}^p\right\},$$
$$W_2 = \left\{y \in \mathbb{R}^n / y = X_2 b, b \in \mathbb{R}^q\right\}. \quad (2)$$

Then,

$$\forall z \in W_1 \cap W_2, \exists (a, b) \in \mathbb{R}^p \times \mathbb{R}^q / z = X_1 a = X_2 b, \quad (3)$$

and for such a z, a linear combination of the column vectors y_j, $j \in [\![1, q]\!]$ of X_2 is given by a linear combination of the column vectors x_k, $k \in [\![1, p]\!]$ of X_1.

We thus obtain a relation between the eigenvalues taken by the initial variables (i.e., to test) from the first set "gene expressions" (X_1) and those of the second set "metabolism patterns" (X_2).

The eigenvectors associated with an eigenvalue 0 generate the D-orthogonal parts between W_1 and W_2 so that the figuring vectors in one or the other part are representing totally independent variables.

Canonical analysis gives the proximity to zero (Bartlett test) for eigenvalues (27), which works with multinormalized samples (we will further suppose multinormalization of our samples).

Let us consider the whole set of eigenvalues $\lambda_1,\ldots, \lambda_k, \lambda_{k+1},\ldots, \lambda_p$ written with their multiplicity order, classified by decreasing values. We keep in mind:

$$\forall i \in [\![1, p]\!], 0 \leq \lambda_i \leq 1. \quad (4)$$

Then, the H_0 null hypothesis is:

If $\lambda_1,\ldots, \lambda_k$ have a significance level ≤5% (or any acceptable value), those $\lambda_1,\ldots, \lambda_k$ are judged significantly different from zero. Then, we can test $\lambda_{k+1},\ldots, \lambda_p$ for nullity.

We remind that the error of first kind is $P(\text{reject } H_0 | H_0 \text{ TRUE})$ and that the significance level is the smallest value of

the error of first kind that we could have chosen while still rejecting H_0.

More explicitly, we own the quantity:

$$-\left(n-1-k-\frac{1}{2}(p+q+1)+\sum_{i=1}^{k}\frac{1}{\lambda_i}\right)\ln\left(\prod_{i=k+1}^{p}(1-\lambda_i)\right), \quad (5)$$

where $\prod_{i=k+1}^{p}(1-\lambda_i)$ is the Wilks lambda.

If the theoretical value of $\lambda_{k+1},\ldots,\lambda_p$ is zero, then for a big enough $n-2k$, the real random variable Z associated with this quantity is an estimation of the proximity to zero of $\lambda_{k+1},\ldots,\lambda_p$ and it approximately follows a $\chi^2_{(p-k)(q-k)}$ law.

If Z takes the value z_k, assuming H_0, the probability to observe a so large value of Z is

$$P(Z \geq z_k) = P_k \quad (6)$$

with P_k significance level of the nullity of every term λ_i, $i \in [\![k+1, p]\!]$. This P_k is generally obtained by a reverse reading in the χ^2 table of fractiles, where

$$P_k = 1 - P(Z < z_k). \quad (7)$$

Besides, if the columns x_k of X_1 and y_j of X_2 are D-normalized, then the canonical analysis gives pairs (ξ_i,η_i) of D-normalized eigenvectors, ξ_i and η_i being associated with the same eigenvalue $\lambda_i = \cos^2(\xi_i, \eta_i)$. The canonical coefficient of correlation is $\sqrt{\lambda_i} = \cos(\xi_i, \eta_i)$ [$\cos(\xi_i, \eta_i) \geq 0$ by construction], where these ξ_i are pairwise D-orthogonal in W_1 and these η_i are pairwise D-orthogonal in W_2.

Now, we only consider pairs in which the significance degree associated with the Bartlett test is ≤5% (or any acceptable value). The number of Bartlett-significant pairs is s, with $s \leq \mathrm{Min}(p,q)$.

The canonical analysis then gives the coordinates of $Pr(x_k)$ and $Pr(y_j)$, where Pr represents the D-orthogonal projection from $\mathbb{R}^n$ to the vector space generated by these (ξ_i), $i \in [\![1, s]\!]$ or, alternatively, by these (η_i), $i \in [\![1, s]\!]$.

Then, the canonical analysis coefficients are calculated:

$$Pr(x_k) = \sum_{i=1}^{s}\alpha_k^i\xi_i;\ Pr(y_j) = \sum_{i=1}^{s}\beta_j^i\xi_i \quad (8)$$

or alternatively,

$$Pr(x_k) = \sum_{i=1}^{s}\gamma_k^i\eta_i;\ Pr(y_j) = \sum_{i=1}^{s}\delta_j^i\eta_i. \quad (9)$$

Since, by hypothesis, the vectors x_k and y_j are D-normalized, $|\alpha_k^i| = |\cos(x_k, \xi_i)|$ represents the coefficient of correlation between the initial variable x_k and the canonical variable ξ_i, $(k, i) \in [\![1, p]\!] \times [\![1, s]\!]$, similarly $|\beta_j^i| = |\cos(y_j, \xi_i)|$ represents the coefficient of correlation between the initial variable y_j and the canonical variable ξ_i, $(j, i) \in [\![1,q]\!] \times [\![1, s]\!]$. The same applies to the alternative choice.

For a pair of chosen values (ξ_i, η_i), if $\cos(\xi_i, \eta_i)$ is significantly different from zero, we will consider that the initial variable x_k in the first group X_1 is related to the initial variable y_j in the second group X_2 when $|\alpha_k^i| \geq 0.25$ (or any chosen level in [0,1], a level of $1/\sqrt{p}$ performing at least a correlation with one of the canonical variables) and simultaneously $|\beta_j^i| \geq 0.25$.

The initial variable x_k is the vertex of degree

$$\mathrm{card}\{\beta_j^i / |\beta_j^i| \geq 0.25, j \in [\![1, q]\!]\} \quad (10)$$

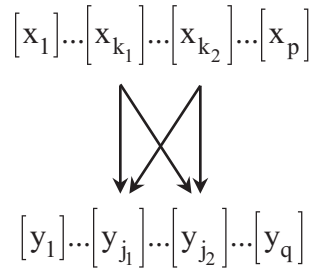

Fig. 2. Directed bipartite graph with respect to the pairs of values (ξ_i, η_i), where x_k (k from 1 to p) is the gene sequence; y_j (j from 1 to q) is the metabolite sequence. An arrow represents the association between a gene and a metabolite.

of a multifunction bipartite graph. We thus obtain a directed bipartite graph with respect to the pairs of values (ξ_i, η_i) presented in Fig. 2.

Then, we assign to x_k its degree $\deg_k^i$ and we will call the explicative level of x_k in a CCA model the integer

$$N_k = \sum_{i=1}^{s}\deg_k^i. \quad (11)$$

The integer

$$N = \sum_{k=1}^{p}N_k \quad (12)$$

will be called the explicative level of the canonical analysis, with respect to the tried set $X_1 = [x_1]\ldots[x_p]$, while $X_2 = [y_1]\ldots[y_q]$ is a priori fixed.

3.3 Heuristic Strategy to optimize Local CCA Models

Integers N_k, $k \in [\![1, p]\!]$ as well as the integer N allow, for a fixed p, a strategy of choice of the most relevant CCA models among the $\binom{r}{p}$ possible models. Our strategy consists in choosing the models that maximize the N_k and in fine N. Nevertheless, calculating the integers N_k and N for each of the $\binom{r}{p}$ possible models would be extremely time-consuming and will lead to too many results to analyze in the end.

One of the major originalities of our work is that we consequently adopt a heuristic strategy to optimize the research of genes highly correlated with metabolism. Genes enter or exit a screen of size p used to generate different local CCA models until the whole set of r genes is tested. The explicative levels of the genes are used to characterize those that are suitable to be retained or exited of the models.

This strategy is formulated as follows:

1) We rank the list of genes by increasing the index numbers: $(x_1,\ldots, x_r)$.

2) We choose the first p genes $(x_1,\ldots,x_p)$ and we compute the list of pairs $[(x_1, N_1),\ldots,(x_p, N_p)]$.

3) We then rank these pairs by decreasing explicative levels (maybe with some duplicated levels).

Through a reorganization of the indexes, $N_1 \geq \ldots \geq N_k \geq \ldots \geq N_p$ will be assumed. Then, it is also assumed that we are given a certain integer threshold level N_0.

4) We only keep $(x_1,\ldots, x_{k0})$ in the list of genes, where k_0 is the greatest integer such that $N_{k0} \geq N_0$. The following nonexplored genes are now introduced to fulfill the list. By this, we obtain a p-terms list (with at most p incoming terms) and we reiterate the process by recomputing the list of pairs.

However, if we choose $N_0 = 0$, there is no progression of the process and we just retain the very first model given by the first p genes. To increase the explicative level of models, we propose to increment N_0 from 0 to the first value of N_0 until all x_k, $k \in [\![1, r]\!]$ included in the initial list have been transferred to canonical analysis.

4 MATHEMATICAL STUDY OF THE ALGORITHM

4.1 Construction of PageRank Matrices

Let us consider a directed bipartite graph made of the set Z of r + q vertices:

$$Z = \{x_1, \ldots, x_r, y_1, \ldots, y_q\}, \tag{13}$$

where $\{x_1,\ldots, x_r\}$ represents the r genes to test and $\{y_1,\ldots, y_q\}$ the q metabolites to explain.

The set E consisting of all arcs of the graph is built as follows:

After a CCA on p genes selected among r according to the rules of the algorithm outlined above, each gene x_k, $k \in [\![1, r]\!]$ is associated with a set of metabolites

$$\{y_j / j \in J_k, J_k \subset [\![1, q]\!]\}, \tag{14}$$

which may be empty. We create arcs with origin x_k and with extremity each value y_j, $j \in J_k$.

Then, we build a PageRank type stochastic square matrix $G(g_{kj})$ with r rows and r columns (2, 23):

We choose $\rho \in [0, 1[$.A reminder:

$$\forall k \in [\![1, r]\!], N_k = \sum_{i=1}^{s} \deg_k^i, \tag{15}$$

with s number of significant eigenvectors used in the CCA and

$$N = \sum_{k=1}^{r} N_k \tag{16}$$

(assumed nonzero) for p tested genes.

Note that with this writing, if x_k is not among the p vectors tested in the CCA necessarily for $k \in [\![1, r]\!]$, $N_k = 0$.

Then, will be denoted by X_m the real random variable corresponding to the index of a gene in the model at step m. $X_m = z_m$, where $z_m \in [\![1, r]\!]$.

We set:

For $k \in [\![1, r]\!]$, index of a tested gene,

$$g_{kk} = \rho + (1-\rho)\frac{N_k}{N} = P(X_{m+1} = k \,|\, X_m = k), \tag{17}$$

$$\forall j \in [\![1, r]\!], j \neq k, g_{kj} = (1-\rho)\frac{N_j}{N} = P(X_{m+1} = j \,|\, X_m = k). \tag{18}$$

If k is not the index of a tested gene,

$$g_{kk} = 1 \text{ and } \forall j \in [\![1, r]\!], g_{kj} = 0. \tag{19}$$

Consequently, for a tested gene:

$$\text{If } N_k = 0, \text{ then } g_{kk} = \rho \text{ and } \forall j \in [\![1, r]\!], g_{jk} = 0. \tag{20}$$

The probability given to the gene j at step m + 1 knowing the probability given to the gene k at step m depends only on the ratio coming from CCA at step m. This ratio is formed by the explicative level of the gene j divided by the explicative level of the general model for which the gene k, but the other tested genes also potentially have a contribution. This probability is not generally invariant under a transposition of indexes j and k.

G is stochastic because:

1) G is clearly positive.

2) For $k \in [\![1, r]\!]$, index of a tested gene,

$$\sum_{j=1}^{r} g_{kj} = g_{kk} + \sum_{\substack{j=1 \\ j\neq k}}^{r} g_{kj} = \rho + (1-\rho)\frac{N_k}{N} + \sum_{\substack{j=1 \\ j\neq k}}^{r} (1-\rho)\frac{N_j}{N} \tag{21}$$

$$= \rho + \frac{(1-\rho)}{N}\sum_{j=1}^{r} N_j = \rho + \frac{(1-\rho)}{N}N = 1.$$

If k is not a tested gene, it is clear that

$$\sum_{j=1}^{r} g_{kj} = 1. \tag{22}$$

Now, we describe the matrices and tests used by the algorithm:

If, as suggested for the initialization step, we choose a priori the p first genes $(x_1,\ldots, x_p)$ associated with the row probability vector V^0 of coordinates $v_k^0 = P(X_0 = k)$, where $\forall k \in [\![p + 1, r]\!]$, $v_k^0 = 0$, we obtain for the first iteration:

1) $G_1(g_{kj}^1)$ stochastic matrix r × r.

2) $V^1(v_k^1)$ row vector with $V^1 = V^0G_1$.

Thus, V^1 is a probability vector, where:

$$\forall j \in [\![1, r]\!], v_j^1 = P(X_1 = j) = \sum_{k=1}^{r} v_k^0 g_{kj}^1 = v_j^0 g_{jj}^1 + \sum_{\substack{k=1 \\ k\neq j}}^{r} v_k^0 g_{kj}^1. \tag{23}$$

In obvious notations, for a $\rho_1 \in [0,1[$ chosen:

$$\forall j \in [\![1, p]\!], v_j^1 = v_j^0\left(\rho_1 + (1-\rho_1)\frac{N_j^1}{N_1}\right) + \sum_{\substack{k=1 \\ k\neq j}}^{p} v_k^0\left((1-\rho_1)\frac{N_j^1}{N_1}\right) = v_j^0\rho_1 + (1-\rho_1)\frac{N_j^1}{N_1}, \tag{24}$$

and

$$\forall j \in [\![p + 1, r]\!], v_j^1 = 0. \tag{25}$$

The block-wise matrix G_1, the vector V^0 and V^1, are represented in the Fig. 3. The square blocks in the main diagonal of G_1 are from the left to the right of respective dimensions p and r − p.

The coefficient $(1 - \rho_1)/N_1$ being strictly positive, if N_0^1 is a fixed strictly positive threshold level for this step, the equivalence

$$\forall j \in [\![1, p]\!], N_j^1 < N_0^1 \Leftrightarrow v_j^0\rho_1 + (1-\rho_1)\frac{N_j^1}{N_1} < v_j^0\rho_1 + (1-\rho_1)\frac{N_0^1}{N_1} \tag{26}$$

allows an exclusion test for genes.

Let us consider for this step the matrix G_1^0 obtained by replacing in G_1 every N_j^1 by N_0^1, where j describes the set of indexes of the p tested genes in the CCA. Then, the exclusion

$$G_1 = \left(\begin{array}{cccc|ccc} \rho_1+(1-\rho_1)\frac{N_1^1}{N_1} & (1-\rho_1)\frac{N_2^1}{N_1} & \cdots & (1-\rho_1)\frac{N_p^1}{N_1} & 0 & \cdots & 0 \\ (1-\rho_1)\frac{N_1^1}{N_1} & \ddots & & \vdots & \vdots & & \vdots \\ \vdots & & \ddots & (1-\rho_1)\frac{N_p^1}{N_1} & \vdots & & \vdots \\ (1-\rho_1)\frac{N_1^1}{N_1} & \cdots & (1-\rho_1)\frac{N_{p-1}^1}{N_1} & \rho_1+(1-\rho_1)\frac{N_p^1}{N_1} & 0 & \cdots & 0 \\ \hline 0 & \cdots & \cdots & 0 & 1 & 0 & 0 \\ \vdots & & & \vdots & 0 & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & 0 & 1 \end{array}\right)$$

$$V^0\left(\, v_1^0 \;,\ldots,\; v_p^0 \;,\; 0 \;,\ldots,\; 0 \,\right)$$

$$V^1\left(\, v_1^0\rho_1+(1-\rho_1)\frac{N_1^1}{N_1} \;,\ldots,\; v_p^0\rho_1+(1-\rho_1)\frac{N_p^1}{N_1} \;,\; 0 \;,\ldots,\; 0 \,\right)$$

Fig. 3. The block-wise matrix G_1, the vector V^0 and V^1.

test for genes collects those l_1 indexes that break the constraint $V^0G_1 \geq V^0G_1^0$ (term to term inequality for row vectors) because for j index of the p tested genes, if $(V)_j$ is the jth coordinate of the row vector V,

$$\left(V^0G_1\right)_j \geq \left(V^0G_1^0\right)_j \Leftrightarrow N_j^1 \geq N_0^1. \tag{27}$$

Now we suppose $0 < l_1 \leq p$ and $p + l_1 \leq r$. The next step is a CCA in which are removed these l_1 genes, and then we get the 1_1 following ones in such a way that p genes are still tested. We chose $\rho_2 \in [0, 1[$.The CCA is used to build a block-wise stochastic matrix $G_2(g_{kj}^2)$ where, to simplify the notation without loss of generality, it is assumed here

$$\forall j \in [\![p - 1_1 + 1, p]\!], \left(V^0G_1\right)_j < \left(V^0G_1^0\right)_j. \tag{28}$$

We obtain the row vector $V^2(v_k^2)$ using $V^2 = V^1G_2$. Thus, V^2 is a probability vector where

$$\forall j \in [\![1, r]\!], v_j^2 = P\left(X_2 = j\right). \tag{29}$$

The block-wise matrix G_2 and the vector V^2 are represented in Fig. 4. The square blocks in the main diagonal of G_2 are from the left to the right of respective dimensions $p - l_1$, l_1, l_1, $r - (p + l_1)$.

Let us consider for this step the matrix G_2^0 obtained by replacing in G_2 every N_j^2 by N_0^2 (fixed threshold level with $N_0^2 \geq N_0^1$), where j describes the set of indexes of the p tested genes in the CCA. Then, the gene exclusion test consists in collecting those l_2 indexes, which break the constraint $V^1G_2 \geq V^1G_2^0$ because for j index of the p tested genes,

$$\left(V^1G_2\right)_j \geq \left(V^1G_2^0\right)_j \Leftrightarrow N_j^2 \geq N_0^2. \tag{30}$$

$$G_2 = \left(\begin{array}{cccc|ccc|ccc|ccc} \rho_2+(1-\rho_2)\frac{N_1^2}{N_2} & (1-\rho_2)\frac{N_2^2}{N_2} & \cdots & (1-\rho_2)\frac{N_{p-l_1}^2}{N_2} & 0 & \cdots & 0 & (1-\rho_2)\frac{N_{p+1}^2}{N_2} & \cdots & (1-\rho_2)\frac{N_{p+l_1}^2}{N_2} & 0 & \cdots & 0 \\ (1-\rho_2)\frac{N_1^2}{N_2} & & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & & (1-\rho_2)\frac{N_{p-l_1}^2}{N_2} & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ (1-\rho_2)\frac{N_1^2}{N_2} & \cdots & (1-\rho_2)\frac{N_{p-l_1-1}^2}{N_2} & \rho_2+(1-\rho_2)\frac{N_{p-l_1}^2}{N_2} & 0 & \cdots & 0 & (1-\rho_2)\frac{N_{p+1}^2}{N_2} & \cdots & (1-\rho_2)\frac{N_{p+l_1}^2}{N_2} & \vdots & & \vdots \\ \hline 0 & \cdots & \cdots & 0 & 1 & 0 & 0 & 0 & \cdots & 0 & \vdots & & \vdots \\ \vdots & & & \vdots & 0 & \ddots & 0 & \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & 0 & 0 & 1 & 0 & \cdots & 0 & \vdots & & \vdots \\ \hline (1-\rho_2)\frac{N_1^2}{N_2} & \cdots & \cdots & (1-\rho_2)\frac{N_{p-l_1}^2}{N_2} & 0 & \cdots & 0 & \rho_2+(1-\rho_2)\frac{N_{p+1}^2}{N_2} & (1-\rho_2)\frac{N_{p+2}^2}{N_2} & \cdots & (1-\rho_2)\frac{N_{p+l_1}^2}{N_2} & \vdots & \vdots \\ \vdots & & & \vdots & \vdots & & \vdots & (1-\rho_2)\frac{N_{p+1}^2}{N_2} & & \vdots & \vdots & & \vdots \\ \vdots & & & \vdots & \vdots & & \vdots & \vdots & & (1-\rho_2)\frac{N_{p+l_1}^2}{N_2} & \vdots & & \vdots \\ (1-\rho_2)\frac{N_1^2}{N_2} & \cdots & \cdots & (1-\rho_2)\frac{N_{p-l_1}^2}{N_2} & \vdots & & \vdots & (1-\rho_2)\frac{N_{p+1}^2}{N_2} & \cdots & (1-\rho_2)\frac{N_{p+l_1-1}^2}{N_2} & \rho_2+(1-\rho_2)\frac{N_{p+l_1}^2}{N_2} & 0 & \cdots & 0 \\ \hline 0 & \cdots & \cdots & 0 & \vdots & & \vdots & 0 & \cdots & 0 & 1 & 0 & 0 \\ \vdots & & & \vdots & \vdots & & \vdots & \vdots & & \vdots & 0 & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \end{array}\right)$$

$$V^2\left(\, V_1^1\rho_2+(1-\rho_2)\frac{N_1^2}{N_2}\alpha \,,\ldots,\, V_{p-l_1}^1 \;\rho_2+(1-\rho_2)\frac{N_{p-l_1}^2}{N_2}\alpha,\, V_{p-l_1+1}^1 \;,\ldots,\, V_p^1,\, (1-\rho_2)\frac{N_{p+1}^2}{N_2}\alpha \,,\ldots\, ,(1-\rho_2)\frac{N_{p+l_1}^1}{N_2}\alpha \,,0,\ldots,0 \,\right)$$

Fig. 4. The block-wise matrix G_2 and the vector V^2, where α is $\sum_{k=1}^{p-l_1} v_k^1$.

We assume further $0 < l_2 \leq p$ (the model progresses) and $p + \sum_{i=1}^{2} l_i \leq r$ (it remains genes to be tested).

4.2 Algorithm With the Previous Notations, in a Pseudolanguage

The algorithm as described in the previous section can be associated with a stochastic matrices sequence (G_n) of PageRank type according to the following iterative method:

Initialization. Choice of V^0 probability row vector/$\forall k \in [\![p+1, r]\!]$, $v_k^0 = 0$ and $J_0 = [\![1, p]\!]$.

Step 1: 1). Choice of $\rho_1 \in [0, 1[$.

2) CCA for $\{x_i\}_{i \in J_0}$ and computation of G_1.

3) Computation of $V^1 = V^0 G_1$. $\forall j \in [\![1, r]\!]$, $(V^1)_j = P(X_1 = j)$.

4) Choice of the threshold level N_0^1.

5) Exclusion test $(V^0 G_1)_j < (V^0 G_1^0)_j$ for j describing J_0.

6) Computation of l_1 (number of excluded genes).

- ○ If $l_1 = 0$ increment N_0^1 with 1 and return to *step 5*).
- ○ Else ≪ stop testing ≫
  - ● If $p + l_1 \leq r$, computation of J_1.
  - ● Else End.

J_1 is the set of the p incoming genes into the CCA for the next step. J_1 is the union of the set of indexes corresponding to the $p - l_1$ preserved genes belonging to J_0 and the set $[\![p+1, p+l_1]\!]$ of indexes corresponding to the l_1 incoming genes.

Step n (n ≥ 2). 1) Choice of $\rho_n \in [0, 1[$.

2) CCA for $\{x_i\}_{i \in J_{n-1}}$ and computation of G_n.

3) Computation of $V^n = V^{n-1} G_n = V^0 G_1 \ldots G_n$. $\forall j \in [\![1, r]\!]$, $(V^n)_j = P(X_n = j)$.

4) Choice of the threshold level $N_0^n \geq N_0^{n-1}$.

5) Exclusion test $(V^{n-1} G_n)_j < (V^{n-1} G_n^0)_j$ for j describing J_{n-1}.

6) Computation of l_n (number of excluded genes).

- ○ If $l_n = 0$ increment N_0^n with 1 and return to *step 5*).
- ○ Else ≪ stop testing ≫
  - ● If $p + \sum_{i=1}^{n} l_i \leq r$, computation of J_n.
  - ● Else End.

J_n is the set of indexes of the p incoming genes into the CCA for the next step. J_n is the union of the set of indexes corresponding to the $p - l_n$ preserved genes belonging to J_{n-1} and the set $[\![p + \sum_{i=1}^{n} -1\ l_i + 1,\ p + \sum_{i=1}^{n} l_i]\!]$ of indexes corresponding to the l_n incoming genes.

This algorithm is programmed in R language and is available in the Supplementary Data 1.[1] Just like in our R program, after the last step one can force the introduction of eventually remaining genes (strictly less than p). So, it gives one more step.

4.3 Theoretical Results From the Method

For this section 4.3 only, we will retain the notation N to denote the number of steps processed by the algorithm.

4.3.1 Finiteness of the algorithm. This algorithm ends after N steps with clearly $N \leq r - p + 1$.

4.3.2 Stability result. The stability of high explicative levels of persistent genes through several connected CCA models is ensured by a demonstrated theorem (available on demand).

4.3.3 Lemma. As in section 4.1, we consider

$$Z = \{x_1, \ldots, x_r, y_1, \ldots, y_q,\}, \tag{31}$$

where $\{x_1, \ldots, x_r\}$ represents the r genes to test and $\{y_1, \ldots, y_q\}$ the q metabolites.

Under the constraints $N_0^1 = 1$ and for every $n \geq 1$,

$$N_0^{n+1} = \text{Min}\{M \in \mathbb{N}^* / \left(M \geq N_0^n\right) \text{ and } \left(M \text{ threshold level at step n } \Rightarrow 1_n > 0\right)\}, \tag{32}$$

then,

$$\begin{aligned} [0, 1[^N &\xrightarrow{\psi} (\text{Sto}_r)^N \\ (\rho_1, \ldots, \rho_N) &\mapsto (G_1, \ldots, G_N) \end{aligned} \tag{33}$$

is an injective application, where Sto_r denotes the set of stochastic matrices with r rows and r columns.

PROOF. First of all, this is not so clear that ψ is an application, because the determination of the stochastic matrices $G_1, \ldots, G_N$ also depends on the genes exclusion method. As a result, this determination depends on the calculation of the explicative levels, which must be the same regardless of the algorithm re-execution with the same initial vectors and constraints on the N_0^n sequence. For this purpose, it is sufficient that the CCA gives the same eigenvectors for the eigensubspaces associated with Bartlett-significant multiple eigenvalues. Among a various number of iterative methods, one of them is chosen for the determination of eigenvectors by CCA. This method generally begins with initial randomly generated vectors (10, 24). For any re-execution of a CCA, the choice of the initial vectors must be the same. Then the number of steps and the exclusion of genes at each step are the same for any $\rho_1, \rho_2, \ldots$ sequence. So, clearly, the definition set of the application ψ is $[0,1[^N$.

Moreover, if

$$(\rho_1, \ldots, \rho_N) \neq (\rho'_1, \ldots, \rho'_N) \tag{34}$$

and if i_0 is the smallest index such that $\rho_{i_0} \neq \rho'_{i0,}$ then

$$\left(\psi\left(\rho_{i_0}\right)\right)_{i_0} \neq \left(\psi\left(\rho'_{i_0}\right)\right)_{i_0} \tag{35}$$

because $\forall i \in [\![1, N]\!]$, ρ_i is obtained by subtracting any term (which is not 1) of the main diagonal of G_i from an extra diagonal term of the same column.

4.3.4 Self-avoiding walk. The stochastic matrices (G_n) are the result of a non-Markov process. In fact, the equality

$$P\left(X_{m+1} = z_{m+1} \middle| X_{0:m} = z_{0:m}\right) = P\left(X_{m+1} = z_{m+1} \middle| X_m = z_m\right), \tag{36}$$

with $m \in [\![0, N-1]\!]$ and $P(X_{0:m} = z_{0:m}) > 0$, is not generally ensured. We have, due to the exclusions of genes, a random self-avoiding walk. Indeed, the particularity of the process is that each step results in the definitive exclusion of some genes.

4.3.5 PageRank matrices properties. Each matrix G_n, $n \in [\![1, N]\!]$ is stochastic by construction and possibly reducible because, with indexes permutation, we can usually put it in the form $\begin{pmatrix} A & 0 \\ B & C \end{pmatrix}$ with A and C square matrices of nonzero dimension. Thus, if G_n is reducible, G_n is nonergodic. Then, $G = G_1 \ldots G_N$ is still stochastic, possibly reducible, and so nonergodic in this case.

4.3.6 Markov chains, local extended models, and structure matrix. Now we use the stochastic matrix $G = G_1 \ldots G_N$. If G is considered as a matrix associated with a constant transition kernel $\nu(.,.)$ on every step, we thus define a homogeneous Markov chain. It has been shown in section 4.3.5 that G is

[1] The online version of this article has supplemental material.

generally reducible therefore nonergodic. If we restrict the kernel to the set Ess of essential points, we obtain a partition of Ess into equivalence classes $C_1,\ldots, C_t$ for the relation R said "communication relation."

For $(p,q) \in [\![1, r]\!] \times [\![1, r]\!]$, R is defined by

$$(pRq) \Leftrightarrow \left(\exists (n_1, n_2) \in \mathbb{N} \times \mathbb{N} / G^{n_1}_{p,q} > 0 \text{ and } G^{n_2}_{q,p} > 0\right), \quad (37)$$

where p and q are the indexes of genes x_p and x_q, $G^n_{i,j}$ formally representing the term of the row i and the column j inside G to the power n (see Ref. 1).

We call an LEM the set of genes whose indexes belong to the same communication class. According to the Kolmogorov-Doeblin theorem (28), for a given initial distribution Π_0 on $[\![1, r]\!]$ set of genes indexes, each trajectory of the Markov chain reaches (because our kernel ν has only a finite number of states) an essential point. Then the trajectory stays in a communication class of this point. The transition kernels obtained by restriction of the kernel to classes $C_1,\ldots, C_t$ are always irreducible. ν^m formally represents the iterated kernel with the associated matrix G^m. If for $i \in [\![1, t]\!]$, $\nu|c_i$ is periodic of period d_i, the relation S defined for $(p,q) \in c_i \times c_i$ by

$$(pSq) \Leftrightarrow \left(\exists n \in \mathbb{N} / v^{nd_i}(p, q) > 0\right), \quad (38)$$

is an equivalence relation. This relation owns d_i equivalence classes $D_1,\ldots, D_{d_i}$ stable with respect to ν^{di}, where for $j \in [\![1,d_i]\!]$, $\nu^{di}|D_j$ is an irreducible aperiodic kernel that admits a limit distribution Π_∞. This Π_∞ is independent from initial distributions Π_0

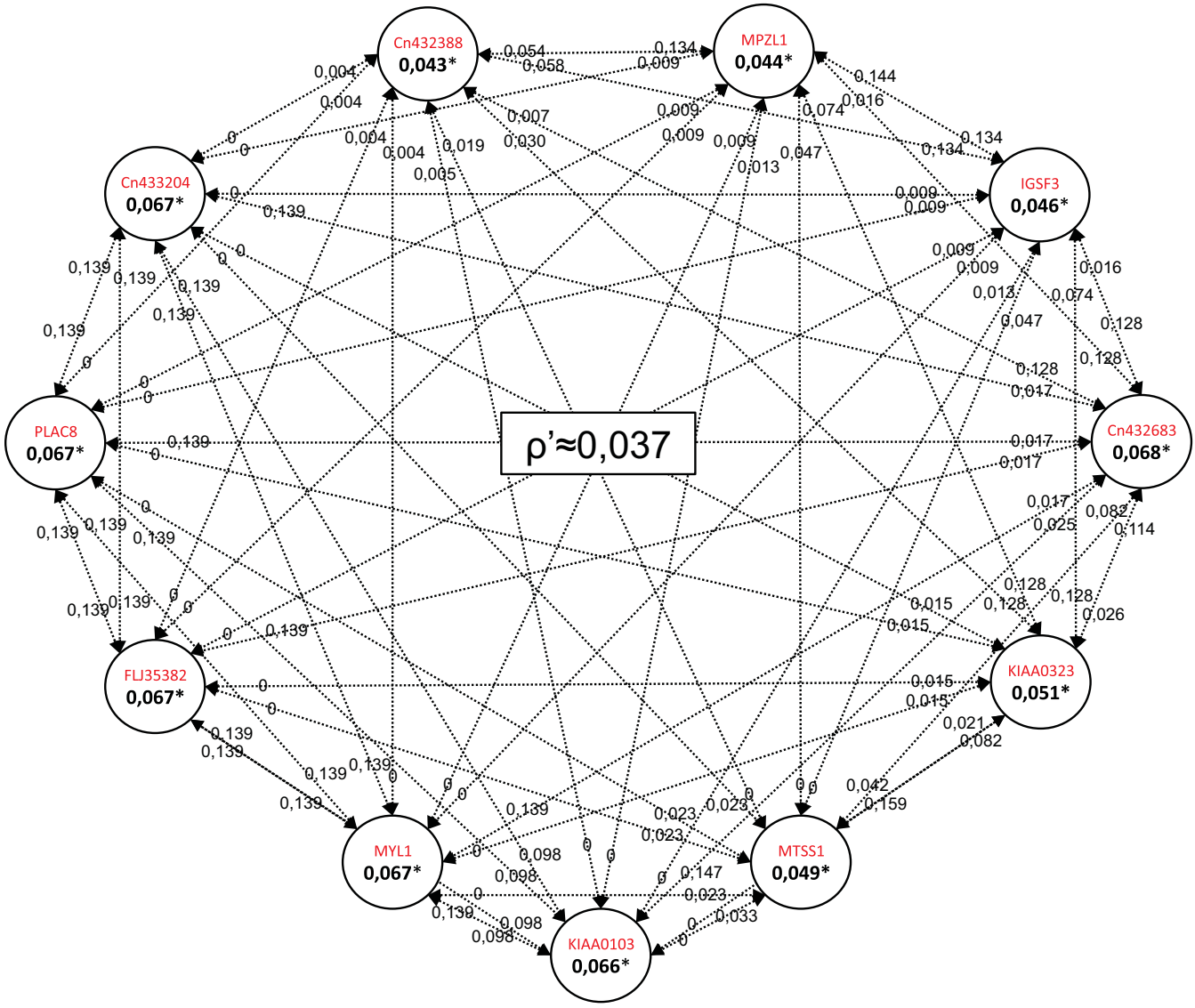

Fig. 5. Weighted Markovian graph of the major local extended model (LEM) issued from the canonical correlation analysis (CCA)-based heuristic algorithm. The stochastic matrix G including the genes association probabilities was computed. G elevated to a certain power is a diagonal block-wise matrix of which the diagonal submatrices are corresponding to genes of the same LEM. The sequence of the successive powers of G was iterated in order converge each column of the submatrices to a nearly constant vector as expected in the Markovian theory (see R program). It was sufficient to compute G^{10}. Then, we can exclude nonessential and nonsignificant genes from the LEM. Here are shown the 11 selected genes (among 27) with a column mean P verifying $P > \rho'$ (ρ' is the threshold of PageRank noise). They are presented inside circles and are labeled with their own probability extracted from G^{10} (*). Those probabilities represent the importance of the gene in the LEM. Probabilities on the connective arrows (dotted lines) are issued from the matrix G. For simplicity extra connections with nonessential or nonsignificant genes are not represented.

whose supports are included in D_j (see Ref. 1). Π_∞ then defines a probability vector invariant under this kernel matrix.

Finally, for each $i \in [\![1, t]\!]$ we obtain an LEM composed of the set of genes whose indexes appear in c_i.

$\nu|_{c_i}$ is an irreducible kernel whose matrix thus admits a probability vector Π_i invariant giving the influence of each gene in this LEM. If c_i is aperiodic, Π_i is the limit distribution. The set of LEMs determines blocks of related genes that define a structure matrix, while other methods (GCCA, RGCCA) have difficulties to provide the rationale for its determination.

5 NUMERICAL RESULTS FROM CASE STUDIES

The algorithm was applied to the three datasets (NutriBov, NutriMous, and LiverTox). One is particularly detailed here: the NutriBov study.

5.1 Computing the G matrix and Gene Selection in the NutriBov Study: the Mathematical Interpretation

The algorithm was processed on the NutriBov study with CCA models including six genes (p) and six metabolites (q) at each iteration step. We obtained a diagonal block-wise matrix G that presented 27 square blocks (submatrices), each defining an LEM as described in 4.3.6.

In our example, it was sufficient to compute G^{10} in order to converge each column of each submatrix to a nearly constant column vector (deviation from the mean value less than or equal to 10^{-3}). Each row of a submatrix exponentially converges to an invariant probability vector Π_∞. p' is the number of genes included in an LEM, $\rho' = 1/p'$ will be called the threshold of PageRank noise.

Then, a gene is selected in an LEM if it presents a corresponding column mean P verifying $P > \rho'$. Here, the submatrices of G are ergodic because the corresponding submatrices of G^{10} have all their terms strictly greater than zero. If these submatrices are irreducible but eventually nonergodic, the previous selection criteria uses a Cesaro average convergence (term to term for a sequence of matrices) here applied to the sequence of powers of the matrix G (9).

We avoided local models presenting only six genes, which meant that all the genes where excluded at the next step of the iteration process. A total of 93 genes out of 293 were selected in G^{10} within the 27 LEMs.

5.2 Algorithm Highlights Genes of Interest, Example With One LEM: from Mathematics to Biology

Among the 27 LEMs computed by the LEM method on the NutriBov study, we will comment in this section on the model including the highest number of genes.

Twenty-seven genes were associated in this single interesting LEM in the matrix G. We selected 11 genes when computing G^{10} because they were involved with probabilities $P >$ 1/27 ($\approx$0.037). The probabilities of genes association for this model are then summarized in a weighted Markovian graph (see Fig. 5) dedicated to the genes selected. Note that this kind of graph could be computed for any other of the 27 LEMs to highlight the process of gene selection.

The functions highlighted in the genes selected by the LEM method were highly coherent as the biological processes gathered them: mainly cytoskeleton coherence, dynamic and organization, and immunity. Moreover, the LEM method highlighted *PLAC8* (Placenta-specific 8) in the oviduct, a gene that regulates embryo-maternal interactions. Higher expression of *PLAC8* in bovine blastocysts (*day 7* embryo postfertilization previously in interface with the oviduct) was reported to be a good marker of pregnancy success (8). This result is of biological interest and is specific to the use of the

Table 2. *Comparison of biological results*

	Cellular Location					Networks
LEM vs. Other Methods	Total	Extracell	Plasma Membrane	Cytoplasm	Nucleus	Directlinks Only (score >20)
			NutriBov			
IPA ready						
RCCA	151	6	6	33	19	3
LEM	93	4	9	26	13	3
shared	**50**	2	4	14	8	2
RCCA spe	101	4	6	28	16	4
LEM spe	**43**	2	6	11	6	2
			NutriMous			
IPA ready						
RCCA	28	4	3	18	2	1
LEM	29	5	4	14	4	2
Shared	**11**	0	2	8	1	1
RCCA spe	17	4	1	11	1	1
LEM spe	**18**	5	2	7	3	1
			LiverTox			
IPA analysis ready						
sPLS	500	29	57	184	124	9
LEM	810	58	84	312	204	14
shared	**229**	10	21	86	53	8
sPLS spe	271	18	33	95	70	8
LEM spe	**581**	42	60	216	144	8

IPA, Ingenuity Pathway Analysis; LEM, local extended model; spe, specific to the method. Boldface indicates what is LEM specific and what is shared with RCCA and/or sPLS methods.

LEM method, as it did not appear in previous analyses by RCCA (31).

5.3 Biological Results on the Other Studies: NutriMous and LiverTox

Applied to the NutriMous study, the algorithm highlighted correlations between all hepatic fatty acids and 29 genes (out of 120). Applied to the LiverTox dataset, the algorithm highlighted correlations between all clinical measurements and 1,015 genes (out of 3,116). As for the NutriBov study, the algorithm identified genes (correlated to physiological or clinical measurements) that were also identified by RCCA or sPLS analyses (Table 2). Moreover, the LEM method also identified new genes that were "algorithm-specific." Fortunately, the cellular distribution of all these genes (shared or specific) was fairly similar whatever the method (RCCA, sPLS, or LEM method), suggesting that these computations highlighted different gene correlations inside common pathways. To validate this, we looked at the LiverTox dataset using the Ingenuity Pathway Analysis software (Ingenuity Systems, http://www.ingenuity.com) and confirmed that genes specifically or commonly identified with sPLS or with the algorithm (LEM method): *1*) belonged to similar gene networks such as "cell morphology" (see Fig. 6) and *2*) helped drawing extended gene networks (Fig. 7 and Fig. 8).

5.4 Global Overview of LEM Method Results Compared With RCCA or sPLS: Advantages and Limits

Benchmarking and time computations. We performed a benchmarking of the three methods on the different datasets using an Intel Core i7-2760QM central processing unit (CPU, 2.4 GHz, 4 cores, and 8 processors), according to R version 2.15.3 (2013-03-04, R Foundation for Statistical Computing), to assess their respective performances during realistic simulations on real biological datasets (Table 3). We principally focused on "user time" and "system time," as the user time is the CPU time charged for the execution of user instructions of the calling process, and the system time is the CPU time

algo-specific

Score	Focus Molecules	Top Functions
51	32	Cell Morphology, Cellular Function and Maintenance, Hereditary Disorder
45	30	Hematological System Development and Function, Humoral Immune Response, DNA Replication, Recombination, and Repair
44	29	Gene Expression, Hereditary Disorder, Cell-To-Cell Signaling and Interaction
37	27	Gastrointestinal Disease, Inflammatory Disease, Tissue Morphology
36	26	Cell Death and Survival, Cardiovascular System Development and Function, Organ Morphology
36	28	Neurological Disease, Cellular Function and Maintenance, Molecular Transport
33	24	Post-Translational Modification, Cell Death and Survival, Cell-To-Cell Signaling and Interaction
27	21	Cancer, Cellular Movement, Gastrointestinal Disease

sPLS-specific

Score	Focus Molecules	Top Functions
46	26	Infectious Disease, Cell Death and Survival, Cellular Development
40	24	RNA Damage and Repair, Cellular Assembly and Organization, Cellular Function and Maintenance
38	27	Cell Death and Survival, Cell Morphology, Cellular Assembly and Organization
32	20	Nucleic Acid Metabolism, Small Molecule Biochemistry, DNA Replication, Recombination, and Repair
23	16	Developmental Disorder, Endocrine System Disorders, Hereditary Disorder
22	15	Cell-To-Cell Signaling and Interaction, Nervous System Development and Function, Infectious Disease
21	15	Carbohydrate Metabolism, Drug Metabolism, Lipid Metabolism

algo-sPLS-shared

Score	Focus Molecules	Top Functions
47	26	Cell Death and Survival, Cancer, Cardiac Necrosis/Cell Death
44	25	Cell Cycle, Hereditary Disorder, Neurological Disease
25	16	Hereditary Disorder, Lipid Metabolism, Neurological Disease
23	15	Carbohydrate Metabolism, Lipid Metabolism, Small Molecule Biochemistry
23	15	Gastrointestinal Disease, Hepatic System Disease, Hereditary Disorder
23	15	Cellular Assembly and Organization, Inflammatory Disease, Inflammatory Response
23	15	Carbohydrate Metabolism, Drug Metabolism, Small Molecule Biochemistry
20	14	Cell Morphology, Nervous System Development and Function, Tissue Morphology

Fig. 6. We compared biological networks generated with the genes identified as algorithm-specific, sparse partial least squares (sPLS)-specific, or shared, while studying the LiverTox datasets. Similar functions appear in each gene list, though not always with the same rank [or Ingenuity Pathway Analysis (IPA) score]. Gene IDs are those recognized by the IPA software. Displayed networks correspond to those listed in Table 2.

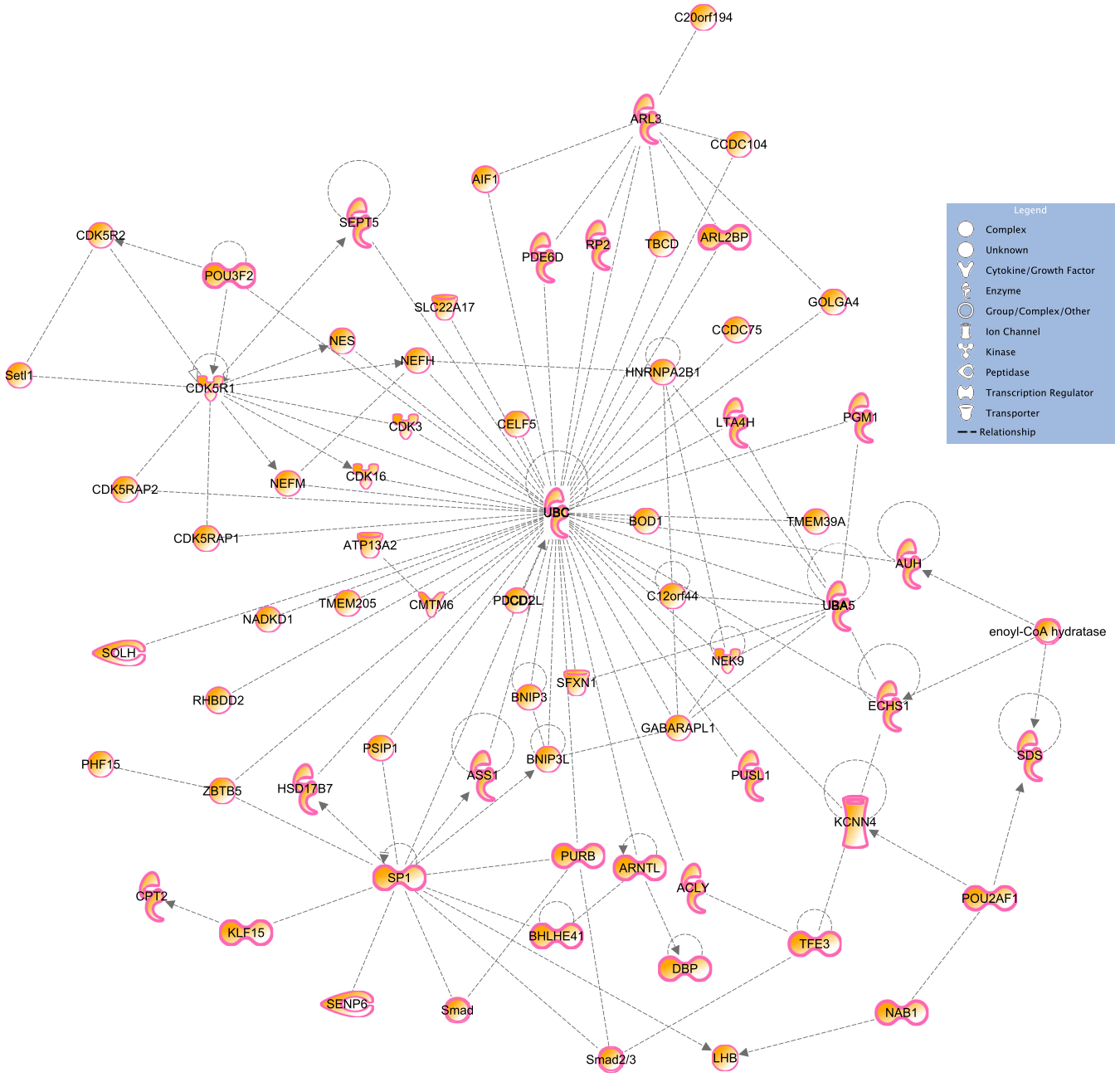

Fig. 7. We compared gene networks related to "cell morphology," involving both algorithm-specific genes and genes in common between sPLS and algorithm analyses. These networks come from the LiverTox datasets and appear complementary. Gene IDs are those recognized by the IPA software. Displayed networks correspond to those from Fig. 6, namely: algo-specific/network 1 vs. algo-sPLS-shared/network 8.

charged for execution by the system on behalf of the calling process. We observed a huge difference between regularized methods such as RCCA and nonegularized ones (e.g., LEM method and sPLS). Indeed, both user time and system time were particularly low with LEM or sPLS on moderated size datasets such as NutriBov or NutriMous (maximum 4.626 s), while RCCA needed 26.06 or 23.12 min to complete the process.

Interestingly, the largest dataset, LiverTox, diverged the computation times. The LEM method only took 2.16 h to complete the analysis on the 3,116 genes $\times$ 10 clinical measurements $\times$ 64 individuals. Comparatively, the search of regularization parameters for RCCA never converges in our conditions on those data, and we estimated the user time $>>$ 11.57 days. Finally, the sPLS only took 1.549 s to estimate the correlations, supporting the fact that sPLS is a direct calculation method (converges to a classical CCA) but is still unable to estimate the coefficients to affect to the structure matrix.

Degree of gene selection. The LEM method was a quite selective method with recommended selection parameters (section 3.2) in the Nutribov study as it highlighted 93 nonredundant genes (out of 293) vs. 151 genes in the classical model of RCCA ($r > 0.6$). As the algorithm was built to provide a

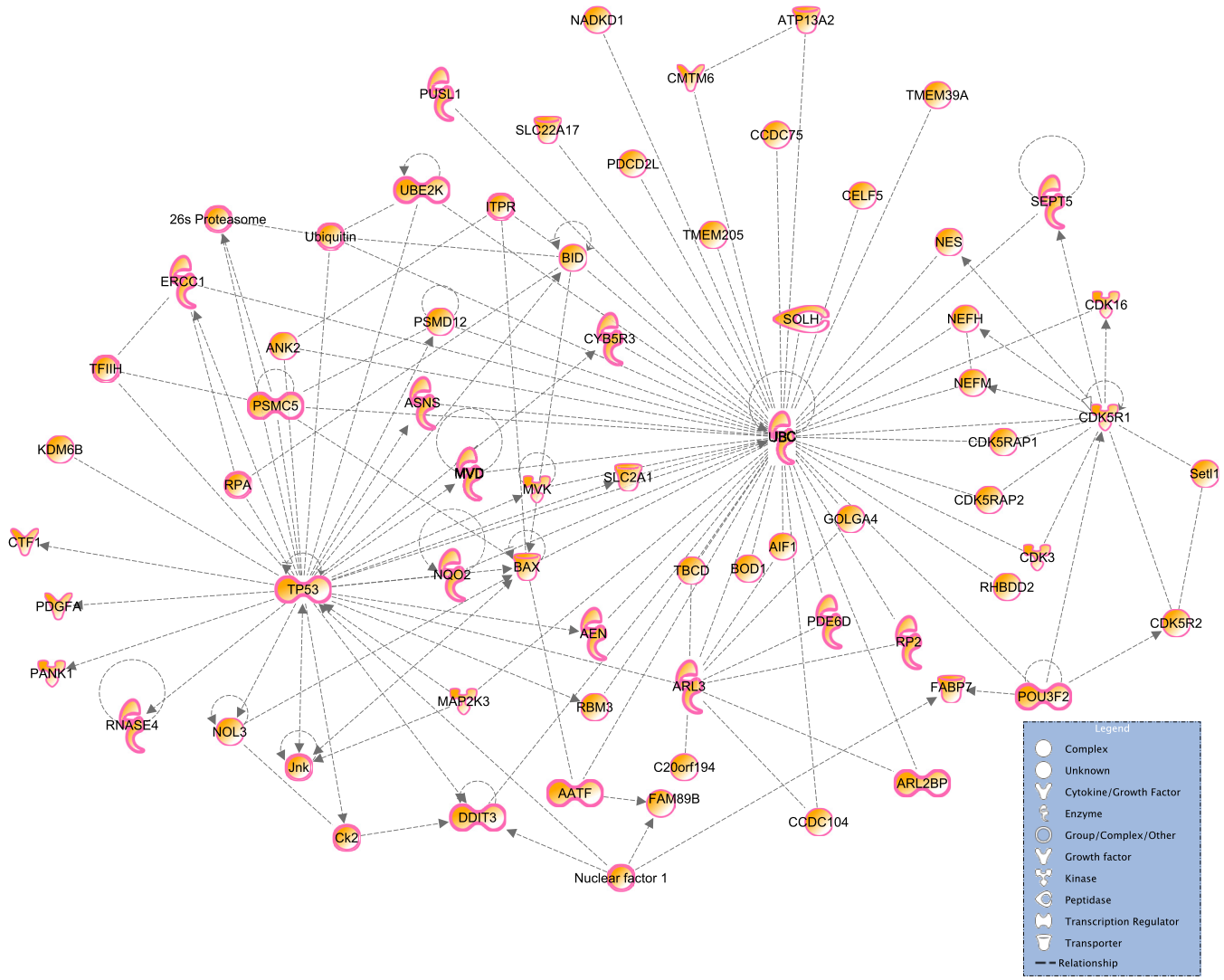

Fig. 8. We compared gene networks related to "cell morphology," involving both sPLS-specific genes and genes in common between sPLS and algorithm analyses. These networks come from the LiverTox datasets and appear complementary. Gene IDs are those recognized by IPA software. Displayed networks correspond to those from Fig. 6, namely: sPLS-specific/network 3 vs. algo-sPLS-shared/network 8.

selection within the transcriptomic dataset, we hope to bring by this exploitation of multiple local models a restrictive set of genes highly related to the whole metabolic context. Those results reach the purpose of biologists who need to find small sets of genes that could be further validated (RT-qPCR) or used as markers in candidate-gene approaches. The levels of gene selection were more similar on the LiverTox and the NutriMous dataset (respectively, 1,015: LEM method vs. 1,032: sPLS and 29: LEM method vs. 27: RCCA).

Mathematical advantages and limits. The LEM method doesn't need a regularization procedure for high dimensional datasets, aims to maximize the correlations between the genes and the metabolic/clinical measurements, proposes a method for the determination of structure matrices, and is fast while computing in R (25). Moreover, the stability of the method was mathematically demonstrated. However, this method remains a heuristic, i.e., an exploratory method, but proposes new concepts for the search of correlations between two biological datasets (see section 4.3). On the other hand, this method does not detail much about which metabolites/clinical measurements contribute the most to the correlations with the genes for each CCA model.

Table 3. *Comparison of computing times*

	NutriBov		NutriMous		LiverTox	
	User Time, s	System Time, s	User Time, s	System Time, s	User Time, s	System Time, s
sPLS	0.092	0.003	0.057	0.003	1.549	0.026
RCCA	1,563.656 (26.06 min)	39.528	1,387.179	17.316	$>>10^6$ (>> 11.57 days)	$>> 5 \times 10^3$
LEM method	4.626	0.040	1.250	0.007	7,785.15 (2.16 h)	164.48

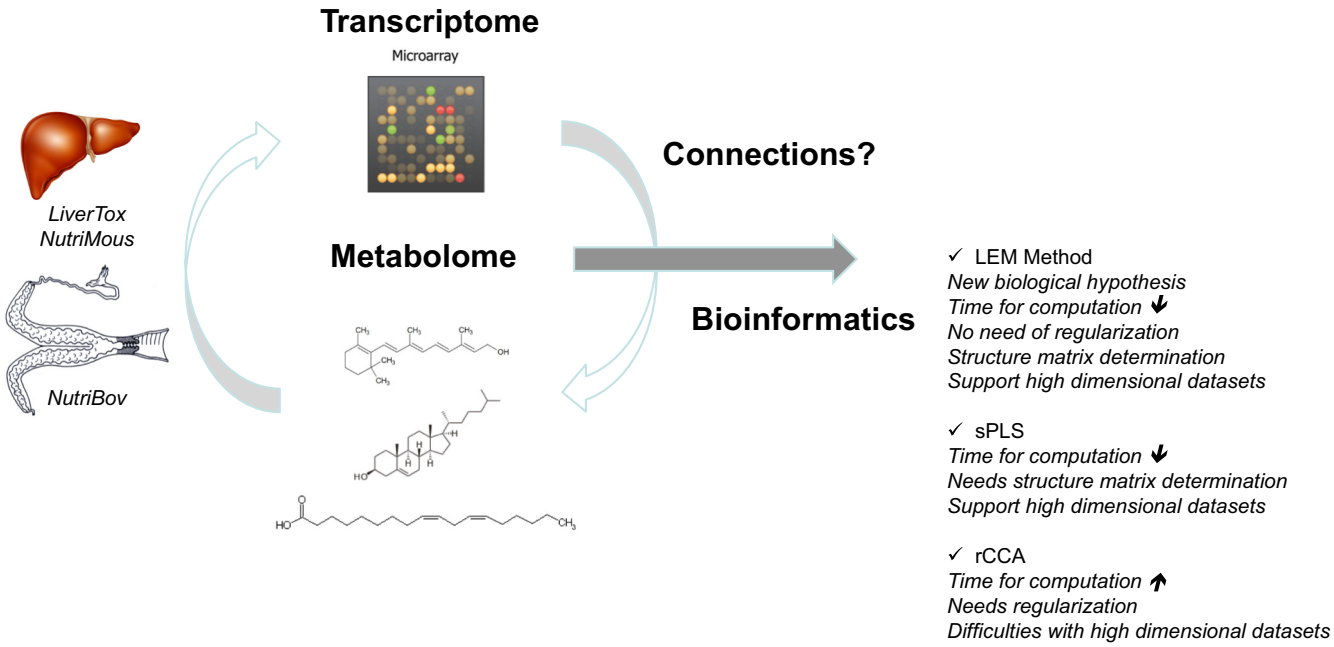

Fig. 9. Summary of the biological questions and the advantages and limits of the LEM method, the sPLS, and the regularized CCA (RCCA).

Biological advantages and limits. As a result of their mathematical differences, these methods do not reveal exactly the same genes or networks as being correlated to metabolites. These methods could be used independently or complementarily. The complementarities of statistical tools for biological analysis should be considered as a suitable methodology to shed light on more aspects of gene regulations than only one would as proposed in section 5.3.

Figure 9 proposes a summary of advantages and limits proposed by the LEM method.

CONCLUSIONS

Here, we built a heuristic CCA-based algorithm able to make a gene selection within transcriptomic datasets. Genes are selected through metabolic parameters used to define a phenotype of interest. For each CCA model across the iteration process, we computed the explicative level of each gene: a parameter assessing the degree of correlation between a gene and the whole metabolic parameters. We optimized the strategy of research of local models by retaining the genes on the basis of their good explicative levels. This method is equivalent to a self-avoiding walk process that gives a stochastic diagonal block-wise matrix G representing the association probability between genes. We considered that this association between genes is fixed, and we had a homogeneous Markov process, which establishes a partition of related genes, gathered in LEM. G was then elevated to a certain power to converge the probabilities to select genes by eliminating nonsignificant or nonessential genes.

The algorithm was simulated on three datasets of different complexities. The comparison with the established models (RCCA, sPLS) revealed many common genes, which makes us confident in our approach. Moreover, the LEM method helped in the search of new and/or complementary genes networks and was very competitive on high dimensional datasets.

Finally, the metabolites that appear in the selected datasets correspond to blood metabolites, including hormones, or to hepatic fatty acids. However, they could virtually be of any origin since the LEM method does not preclude of the biological variables that are analyzed concomitantly to gene expression changes. The LEM could thus apply to any such variable such as animals, plants, or bacteria, in vivo or in vitro biological systems, at the cellular or subcellular level (mitochondria for example), as long as these variables are measurable, even at a single cell level (22). On the basis of recent studies, when these "nongenic" variables happen to be metabolites, they could be those of interest for animal or vegetal physiology (6, 30), disease (14), or development but also animal or vegetal stem cell fate (20, 26, 33).

ACKNOWLEDGMENTS

The authors are glad to thank Kim-Anh Lê Cao, Ignacio González, Sébastien Déjean, the Editor, and two referees for constructive comments that contributed to improving the quality of the paper. The authors also thank Sébastien Déjean for contributions in the programming.

GRANTS

This work received the financial support of the National Research Agency of France and APIS-GENE (GENANIMAL Program). Damien Valour was a PhD student with a DGER-INRA grant followed by a one-year grant thanks to the Reproseq project (INRA/UNCEIA).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

Author contributions: D.V., B.G., and B.V. conception and design of research; D.V. performed experiments; D.V., I.H., B.G., and B.V. analyzed data; D.V., I.H., B.G., and B.V. interpreted results of experiments; D.V., I.H., and B.V. prepared figures; D.V., I.H., B.G., and B.V. drafted manuscript; D.V., I.H., B.G., and B.V. edited and revised manuscript; D.V., I.H., B.G., and B.V. approved final version of manuscript.

REFERENCES

1. **Bérard J.** *Chaînes de Markov*. Cours publié de l'université Lyon I, 2009.
2. **Brezinski C, Redivo-Zaglia M.** The PageRank Vector: properties, computation, approximation, and acceleration. *SIAM J Matrix Anal Appl* 28: 551–575, 2006.
3. **Bushel PR, Wolfinger RD, Gibson G.** Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Syst Biol* 1: 15, 2007.
4. **Carroll JD, Green PE, Chaturvedi A.** *Mathematical Tools for Applied Multivariate Analysis*. San Diego, CA: Academic, 1997.
5. **Carroll JD.** A generalization of canonical correlation analysis to three or more sets of variables. *Proc 76th Ann Conv Am Psych Assoc*: 227–228, 1968.
6. **Dupont J, Reverchon M, Cloix L, Froment P, Rame C.** Involvement of adipokines, AMPK, PI3K and the PPAR signaling pathways in ovarian follicle development and cancer. *Int J Dev Biol* 56: 959–967, 2012.
7. **Eaton ML, Perlman MD.** The non-singularity of generalized sample covariance matrices. *Ann Stat* 1: 710–717, 1973.
8. **El-Sayed A, Hoelker M, Rings F, Salilew D, Jennen D, Tholen E, Sirard MA, Schellander K, Tesfaye D.** Large-scale transcriptional analysis of bovine embryo biopsies in relation to pregnancy success after transfer to recipients. *Physiol Genomics* 28: 84–96, 2006.
9. **Foata P, Fuchs A.** *Processus stochastiques: Processus de Poisson, chaînes de Markov et martingales*. Paris: Dunod, 2002.
10. **Gmür T.** *Dynamique des structures: Analyse modale numérique*. Lausanne: Presse polytechniques et universitaires romandes, 1997.
11. **González I, Déjean S, Martin P, Baccini A.** CCA: an R package to extend canonical correlation analysis. *J Stat Software* 23: 2008.
12. **Gonzalez I, Le Cao KA, Davis MJ, Dejean S.** Visualising associations between paired 'omics' data sets. *BioData Min* 5: 19, 2012.
13. **Hotteling H.** Relations between two sets of variables. *Biometrika* 28: 321–377, 1936.
14. **Kaelin WG Jr, McKnight SL.** Influence of metabolism on epigenetics and disease. *Cell* 153: 56–69, 2013.
15. **Lê Cao KA, González I, Déjean S.** integrOmics: an R package to unravel relationships between two omics data sets Bioinformatics. *Bioinformatics* 25: 2855–2856, 2009.
16. **Le Cao KA, Rossouw D, Robert-Granie C, Besse P.** A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol* 7: Article 35, 2008.
17. **Ledoit O, Wolf M.** A well conditioned estimator for large-dimensional covariance matrices. *J Multivar Anal* 88: 365–411, 2004.
18. **Leurgans SE, Moyeed RA, Silverman BW.** Canonical correlation analysis when the data are curves. *J Roy Stat Soc B* 55: 725–740, 1993.
19. **Martin PG, Guillou H, Lasserre F, Dejean S, Lan A, Pascussi JM, Sancristobal M, Legrand P, Besse P, Pineau T.** Novel aspects of PPARalpha-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. *Hepatology* 45: 767–777, 2007.
20. **Meissen JK, Yuen BT, Kind T, Riggs JW, Barupal DK, Knoepfler PS, Fiehn O.** Induced pluripotent stem cells show metabolomic differences to embryonic stem cells in polyunsaturated phosphatidylcholines and primary metabolism. *PLoS One* 7: e46770, 2012.
21. **Morine MJ, McMonagle J, Toomey S, Reynolds CM, Moloney AP, Gormley IC, Gaora PO, Roche HM.** Bi-directional gene set enrichment and canonical correlation analysis identify key diet-sensitive pathways and biomarkers of metabolic syndrome. *BMC Bioinformat* 11: 499, 2010.
22. **Nemes P, Rubakhin SS, Aerts JT, Sweedler JV.** Qualitative and quantitative metabolomic investigation of single neurons by capillary electrophoresis electrospray ionization mass spectrometry. *Nat Protoc* 8: 783–799, 2013.
23. **Page L, Brin S, Motwani R, Winograd T.** *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford InfoLab, 1999.
24. **Quarteroni A, Sacco R, Saleri F.** *Méthodes numériques pour le calcul scientifique*. Springer, 2000.
25. **R Development Core Team.** *R: a Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2011.
26. **Rafalski VA, Mancini E, Brunet A.** Energy metabolism and energy-sensing pathways in mammalian embryonic and adult stem cell fate. *J Cell Sci* 125: 5597–5608, 2012.
27. **Saporta G.** *Probabilités, analyse de données et statistique*. Paris: TECHNIP, 2006.
28. **Sonin I.** The asymptotic behaviour of a general finite nonhomogeneous Markov chain (the decomposition-separation theorem). *Stat Prob Game Theory* 30: 337–346, 1996.
29. **Tenenhaus A, Tenenhaus M.** Regularized generalized canonical correlation analysis. *Psychometrika* 76: 257–284, 2011.
30. **Turktas M, Inal B, Okay S, Erkilic EG, Dundar E, Hernandez P, Dorado G, Unver T.** Nutrition metabolism plays an important role in the alternate bearing of the olive tree (Olea europaea L.). *PLoS One* 8: e59876, 2013.
31. **Valour D, Hue I, Degrelle S, Dejean S, Marot G, Dubois O, Germain G, Humblot P, Ponter A, Charpigny G, Grimard B.** Pre- and post-partum mild underfeeding influences gene expression in the reproductive tract of cyclic dairy cows. *Reprod Domest Anim* 48: 484–499, 2013.
32. **Vinod HD.** Canonical ridge and econometrics of joint production. *J Econometrics* 6: 129–137, 1976.
33. **Xiong Y, McCormack M, Li L, Hall Q, Xiang C, Sheen J.** Glucose-TOR signalling reprograms the transcriptome and activates meristems. *Nature* 496: 181–186, 2013.