Gene selection heuristic algorithm for nutrigenomics studies

D. Valour,¹,² I. Hue,¹,² B. Grimard,¹,² and B. Valour³
¹INRA, UMR 1198 Biologie du Développement et Reproduction, Jouy-en-Josas, France; ²Université Paris-Est, ENVA, UMR1198, Maisons-Alfort, France; and ³Pôle Supérieur de Bansac, Département de Mathématiques, Clermont-Ferrand, France

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

The purpose of this paper is to provide a new heuristic method that gives the linear links between two sets of columns in 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 establish as one of the main points of this paper that a Markov process applied to the previous stochastic matrix converges and matrix conditioning issues because of a too classical method would fail when applied to the previous stochastic matrix converges and matrix conditioning issues because of a too 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 resumee 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.

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 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 establish as one of the main points of this paper that a Markov process applied to the previous stochastic matrix converges and matrix conditioning issues because of a too 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 resumee 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.
The 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

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

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 in 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, 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 X₁ (size n × r) and X₂ (size n × 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 X₁ or X₂, 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 ≤ q < n ≪ r). Now, the size of X₁ is n × p.

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

\[ X = [X_1]\_n\times_p [X_2]\_n\times_q \]

block-wise defined matrix [size n × (p + q)], where we aim to find a linear relationship between column vectors of X₁ and column vectors of X₂.

The heuristic method allows the choice of a sufficiently small number p of variables for each performed CCA. Thus, we reasonably hope X₁ and X₂ to be full rank matrices [Rank(X₁) = p and Rank(X₂) = q]. We also assume that the columns of X₁ and X₂ are centered and normalized by using D-metrics (here D = 1.0, Euclidian-metrics).

We want to extract, among \( \binom{p}{q} \) 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, \( \min(p, q) \) (here \( p ≤ q \)). The multiplicity order of the eigenvalue 1 is \( \text{Dim}(W_1 \cap W_2) \), with:

\[ W_1 = \{ x \in \mathbb{R}^n / x = X_1 a, a \in \mathbb{R}^p \}, \]

\[ W_2 = \{ y \in \mathbb{R}^n / y = X_2 b, b \in \mathbb{R}^q \}. \]

Then,

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

and for such a z, a linear combination of the column vectors \( y_j \), \( j \in [1, q] \) of X₂ is given by a linear combination of the column vectors \( x_k \), \( k \in [1, p] \) of X₁.

We thus obtain a relation between the eigenvalues taken by the initial variables (i.e., to test) from the first set “gene expressions” (X₁) and those of the second set “metabolism patterns” (X₂).

The eigenvectors associated with an eigenvalue 0 generate the D-orthogonal parts between W₁ and W₂ 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 \( \lambda_1,..., \lambda_k, \lambda_{k+1},..., \lambda_p \) written with their multiplicity order, classified by decreasing values. We keep in mind:

\[ \forall i \in [1, p], 0 ≤ \lambda_i ≤ 1. \]

Then, the \( H_0 \) null hypothesis is:

If \( \lambda_1,..., \lambda_k \) have a significance level ≤5% (or any acceptable value), those \( \lambda_1,..., \lambda_k \) are judged significantly different from zero. Then, we can test \( \lambda_{k+1},..., \lambda_p \) 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:
\[
- \left( n - 1 - k - \frac{1}{2}(p + q + 1) + \sum_{i=1}^{k} \frac{1}{\lambda_i} \right) \ln \left( \prod_{i=k+1}^{p} (1 - \lambda_i) \right),
\]
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 $X^2_{(p-k)(q-k)}$ law.

If $Z$ takes the value $z_k$, assuming $H_0$, the probability to observe a so large value of $Z$ is
\[
P(Z \geq z_k) = P_k
\]
with $P_k$ significance level of the nullity of every term $\lambda_i, i \in \{ k+1, \ldots, P \}$. This $P_k$ is generally obtained by a reverse reading in the $\chi^2$ table of fractiles, where
\[
P_k = 1 - P(Z < z_k).
\]

Besides, if the columns $x_k$ of $X_1$ and $y_j$ of $X_2$ are D-normalized, then the canonical analysis gives pairs $(\xi_i, \eta_i)$ of D-normalized eigenvectors, $\xi_i$ and $\eta_i$ 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 $\xi_i$ are pairwise D-orthogonal in $W_1$ and these $\eta_i$ are pairwise D-orthogonal in $W_2$.

Now, we only consider pairs in which the significance degree associated with the Bartlett test is $\leq 5\%$ (or any acceptable value).

The number of Bartlett-significant pairs is $s$, with $s \leq \text{Min}(p,q)$.

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

Then, the canonical coefficients are calculated:
\[
\Pr(x_k) = \sum_{i=1}^{s} a_i^k \xi_i; \quad \Pr(y_j) = \sum_{i=1}^{s} b_i^j \xi_i
\]
or alternatively,
\[
\Pr(x_k) = \sum_{i=1}^{s} a_i^k \eta_i; \quad \Pr(y_j) = \sum_{i=1}^{s} b_i^j \eta_i.
\]

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

For a pair of chosen values $(\xi_i, \eta_i)$, 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 $|a_i^k| \geq 0.25$ (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 $|b_j^i| \geq 0.25$.

The initial variable $x_k$ is the vertex of degree
\[
\text{card}\{|b_j^i| \geq 0.25, j \in \{ 1, q \}\}
\]
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:

\[
Z = \{x_1, \ldots, x_p, y_1, \ldots, y_q\},
\]

where \( \{x_1, \ldots, x_p\} \) represents the \( p \) 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

\[
\{y_j / j \in J_k, J_k \subset [1, q]\},
\]

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:

\[
\forall k \in [1, r], N_k = \sum_{i=1}^{s} \text{deg}_i^k,
\]

with \( s \) number of significant eigenvectors used in the CCA and

\[
N = \sum_{k=1}^{r} N_k
\]

(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,

\[
g_{kk} = \rho + (1 - \rho) \frac{N_k}{N} = \text{P}(X_{m+1} = k|X_m = k),
\]

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

If \( k \) is not the index of a tested gene,

\[
g_{kk} = 1 \quad \text{and} \quad \forall j \in [1, r], g_{kj} = 0.
\]

Consequently, for a tested gene:

If \( N_k = 0 \), then \( g_{kk} = \rho \) and \( \forall j \in [1, r], g_{jk} = 0 \).

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,

\[
\sum_{j=1}^{r} g_{kj} = g_{kk} + \sum_{j=1}^{r} g_{kj} = \rho + (1 - \rho) \frac{N_k}{N} + \sum_{j=1}^{r} (1 - \rho) \frac{N_j}{N} = 1.
\]

If \( k \) is not a tested gene, it is clear that

\[
\sum_{j=1}^{r} g_{kj} = 1.
\]

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 \( v^0_i = \text{P}(X_0 = k) \), where \( \forall k \in [p + 1, r], v^0_k = 0 \), we obtain for the first iteration:

1) \( G_1(g_{kj}) \) stochastic matrix \( r \times r \).
2) \( V^1(V^0) \) row vector with \( V^1 = V^0 G_1 \).

Thus, \( V^1 \) is a probability vector, where:

\[
\forall j \in [1, r], v_j^1 = \text{P}(X_j = j) = \sum_{k=1}^{r} v^0_{kj} g_{kj} = v^0_j + \sum_{k=1}^{r} v^0_{kj} g_{kj}.
\]

In obvious notations, for a \( \rho_1 \in [0,1[, \text{ chosen:}

\[
\forall j \in [1, p], v_j^1 = v^0_j \left( \rho_1 + (1 - \rho_1) \frac{N^1_j}{N^1} \right) + \sum_{k=1}^{p} v^0_k \left( (1 - \rho_1) \frac{N^1_k}{N^1} \right) = v^0_j \rho_1 + (1 - \rho_1) \frac{N^1_j}{N^1},
\]

and

\[
\forall j \in [p + 1, r], v_j^1 = 0.
\]

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 \((1 - \rho_1)N^1_j\) being strictly positive, if \( N^0_j \) is a fixed strictly positive threshold level for this step, the equivalence

\[
\forall j \in [1, p], N^1_j < N^0_j \Leftrightarrow v^0_j \rho_1 + (1 - \rho_1) \frac{N^1_j}{N^1} < v^0_j \rho_1
\]

allows an exclusion test for genes.

Let us consider for this step the matrix \( G_1^0 \) obtained by replacing in \( G_1 \) every \( N^1_j \) by \( N^0_j \), where \( j \) describes the set of indexes of the \( p \) tested genes in the CCA. Then, the exclusion
test for genes collects those $l_1$ indexes that break the constraint $V_0 G_1 \leq H_1$ (term to term inequality for row vectors) because for $j$ index of the $p$ tested genes, if $(V)_j$ is the $j$th coordinate of the row vector $V$,

$$
(V)_j \leq \left(\sum_{k=1}^{l_1} \alpha_{k,j} \right) \leq N_0.
$$

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 $l_1$ 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 $G_2(g_{ij}^2)$ where, to simplify the notation without loss of generality, it is assumed here

$$
\forall j \in [p - l_1 + 1, p], (V_0 G_1)_j < (V_0 G_0)_j.
$$

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

$$
\forall j \in [1, r], v^2_j = P(X_2 = j).
$$

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 $N_0^2 \geq N_0$), 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 $V^2 G_2 \leq H_1$ because for $j$ index of the $p$ tested genes,

$$
(V G_2)_j \leq (V G_2^0)_j \Leftrightarrow N_j^2 \leq N_0^2.
$$

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

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

Physiol Genomics • doi:10.1152/physiolgenomics.00139.2012 • www.physiolgenomics.org
We assume further $0 < l_2 \leq p$ (the model progresses) and $p + \sum_{i=1}^{n} l_i \leq r$ (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 Page-Rank type according to the following iterative method:

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

Step 1: For $j = 1$. Choice of $\rho_1 \in [0, 1]$.  
2) CCA for $\{x_i\}_{i=1}^p$ 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 $N^0_0$.
5) Exclusion test $(V^0 G_1)_j < (V^0 G_0)_j$ for $j$ describing $J_0$.
6) Computation of $I_1$ (number of excluded genes).
   - If $I_1 = 0$ increment $N^0_0$ with 1 and return to step 5).
   - Else $< stop testing >$
     - If $p + I_1 \leq r$, 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 $p - I_1$ preserved genes belonging to $J_0$ and the set $[p + 1, p + I_1]$ of indexes corresponding to the $I_1$ incoming genes.

Step $n$ ($n \geq 2$). 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 $V^n = V^{n-1} G_n = V^{n-1} G_{1} ... G_{n-1}, \forall j \in [1, r]$, $(V^n)_j = P(X_n = j)$.
4) Choice of the threshold level $N^n_0 \geq N^0_0 - 1$.
5) Exclusion test $(V^{n-1} G_n)_j < (V^{n-1} G_{n-1})_j$ for $j$ describing $J_{n-1}$.
6) Computation of $I_n$ (number of excluded genes).
   - If $I_n = 0$ increment $N^n_0$ 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 $p - I_n$ preserved genes belonging to $J_{n-1}$ and the set $[p + \sum_{i=1}^{n} l_i + 1, p + \sum_{i=1}^{n} l_i]$ of indexes corresponding to the $I_n$ incoming genes.

This algorithm is programmed in R language and is available in the Supplementary Data 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

\[
Z = \{x_1, \ldots, x_r, y_1, \ldots, y_q\},
\]

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

Under the constraints $N^0_0 = 1$ and for every $n \geq 1$,

\[
N^{n+1}_0 = \min\{M \in \mathbb{N}^+ / (M \geq N^0_0) \text{ and } (M \text{ threshold level at step } n \Rightarrow I_n > 0)\},
\]

then,

\[
[0, 1]^{\psi} \rightarrow \frac{\psi(0, 1) \cdots \psi(0, N)}{\psi(0, 1) \cdots \psi(0, N)}
\]

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

Proof. First of all, this is not so clear that $\psi$ 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 $N^0_0$ 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 $\psi$ is $[0, 1]^N$.

Moreover, if

\[
(\rho_1, \ldots, \rho_N) \neq (\rho_1, \ldots, \rho_N),
\]

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

\[
\psi(\rho_{i_0})_{i_0} \neq \psi(\rho_{i_0}')_{i_0}
\]

because $\forall i \in [1, N]$, $\rho_i$ 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

\[
P(X_{m+1} = z_{m+1} | X_{0:m} = z_{0:m}) = P(X_{m+1} = z_{m+1} | X_{m} = z_m),
\]

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 $(\bar{G}_n^0)$ with $A$ and $C$ square matrices of nonzero dimension. Thus, if $G_n$ is reducible, $G_n$ is nonergodic. Then, $G = G_1 ... G_N$ 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 $G = G_1 ... G_N$. If $G$ is considered as a matrix associated with a constant transition kernel $\psi(...) \ldots$ on every step, we thus define a homogeneous Markov chain. It has been shown in section 4.3.5 that $G$ is

\[\text{Physiol Genomics} \cdot \text{doi:10.1152/physiolgenomics.00139.2012} \cdot \text{www.physiolgenomics.org}\]
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, ..., C_t for the relation R said
“communication relation.”

For (p,q) ∈ [1, r] × [1, r], R is defined by

\[(pRq) ⇔ (∃ (n_1, n_2) ∈ \mathbb{N} × \mathbb{N}/G^{n_1}_{p,q} > 0 \text{ and } G^{n_2}_{q,p} > 0), \quad (37)\]

where p and q are the indexes of genes x_p and x_q, G_{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-
Dooblin theorem (28), for a given initial distribution \(\Pi_0\) on
[1, r] set of genes indexes, each trajectory of the Markov chain
reaches (because our kernel \(\nu\) has only a finite number of
states) an essential point. Then the trajectory stays in a com-
munication class of this point. The transition kernels obtained
by restriction of the kernel to classes C_1, ..., C_t are always
irreducible. \(\nu^m\) formally represents the iterated kernel with the
associated matrix \(G^m\). If for i ∈ [1, t], \(\nu^{d_i}\) is periodic of period
d_i, the relation S defined for (p,q) ∈ C_{i_1} × C_{i_2} by

\[(pS q) ⇔ (∃ n ∈ \mathbb{N}/\nu^{nd_i}(p, q) > 0), \quad (38)\]

is an equivalence relation. This relation owns d_i equivalence
classes D_{i_1}, ..., D_{i_t} stable with respect to \(\nu^{d_i}\), where for j ∈ [1, d_i],
\(\nu^{d_i}D_j\) is an irreducible aperiodic kernel that admits a limit distri-
bution \(\Pi_0\). This \(\Pi_0\) is independent from initial distributions \(\Pi_{0,1}\)

---

**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 = \(\rho'\approx 0.037\) (\(\rho'\) 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 $D_i$ (see Ref. 1). $II_\circ$ then defines a probability vector invariant under this kernel matrix.

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

$v_{\circ c_i}$ is an irreducible kernel whose matrix thus admits a probability vector $II_\circ$ invariant giving the influence of each gene in this LEM. If $c_i$ is aperiodic, $II_\circ$ 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 $\rho^\circ$. $p^\circ$ is the number of genes included in an LEM, $\rho^\circ = 1/p^\circ$ 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 $P > \rho^\circ$. 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 $P > 1/27 (=0.037)$. 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 post-fertilization 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 LEM method.

Table 2. Comparison of biological results

| LEM vs. Other Methods | Cellular Location | Networks | | | |
|-----------------------|------------------|----------|----------|----------|
|                       | Total | Extracell | Plasma Membrane | Cytoplasm | Nucleus | Directlinks Only (score >20) |
| 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        |
| 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        |
| 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

---

**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.
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 × 10 clinical measurements × 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. 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.
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

<table>
<thead>
<tr>
<th></th>
<th>NutriBov</th>
<th>NutriMous</th>
<th>LiverTox</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>User Time, s</td>
<td>System Time, s</td>
<td>User Time, s</td>
</tr>
<tr>
<td>sPLS</td>
<td>0.092</td>
<td>0.003</td>
<td>0.057</td>
</tr>
<tr>
<td>RCCA</td>
<td>1,563.656 (26.06 min)</td>
<td>39.528</td>
<td>1,387.179</td>
</tr>
<tr>
<td>LEM method</td>
<td>4.626</td>
<td>0.040</td>
<td>1.250</td>
</tr>
</tbody>
</table>
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 DGÉR-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

43. Gonzalez I, Dejean S, Martin P, Baccini A. Involvement of adipo...