Comparing gene expression networks in a multi-dimensional space to extract similarities and differences between organisms

Gaëlle Lelandais\textsuperscript{1,2,*}, Pierre Vincens\textsuperscript{2}, Anne Badel-Chagnon\textsuperscript{2}, Stéphane Vialette\textsuperscript{3}, Claude Jacq\textsuperscript{1} and Serge Hazout\textsuperscript{2}

\textsuperscript{1}Laboratoire de Génétique Moléculaire, CNRS UMR 8541, Ecole Normale Supérieure, 46 rue d’Ulm, 75230 Paris cedex 05, France, \textsuperscript{2}Equipe de Bioinformatique Génomique et Moléculaire, INSERM U726, Université Paris 7, case 7113, 2 Place Jussieu, 75251 Paris cedex 05, France and \textsuperscript{3}Laboratoire de Recherche en Informatique, UMR CNRS 8623, Faculté des Sciences d’Orsay, Université Paris Sud, 91405 Orsay, France

Received on October 25, 2005; revised on February 9, 2006; accepted on March 4, 2006

Advance Access publication March 9, 2006

Associate Editor: David Rocke

ABSTRACT

\textbf{Motivation:} Molecular evolution, which is classically assessed by comparison of individual proteins or genes between species, can now be studied by comparing co-expressed functional groups of genes. This approach, which better reflects the functional constraints on the evolution of organisms, can exploit the large amount of data generated by genome-wide expression analyses. However, it requires new methodologies to represent the data in a more accessible way for cross-species comparisons.

\textbf{Results:} In this work, we present an approach based on Multi-dimensional Scaling techniques, to compare the conformation of two gene expression networks, represented in a multi-dimensional space. The expression networks are optimally superimposed, taking into account two criteria: (1) inter-organism orthologous gene pairs have to be nearby points in the final multi-dimensional space and (2) the distortion of the gene expression networks, the organization of which reflects the similarities between the gene expression measurements, has to be circumscribed. Using this approach, we compared the transcriptional programs that drive sporulation in budding and fission yeasts, extracting some common properties and differences between the two species.

\textbf{Availability:} The source code is freely distributed to academic users upon request to the authors. More information can be found online at http://www.biologie.ens.fr/lgmgml/publication/comp3d/.

\textbf{Contact:} lelandais@biologie.ens.fr

\textbf{Supplementary information:} Supplementary data are available at http://www.biologie.ens.fr/lgmgml/publication/comp3d/SupData.php

1 INTRODUCTION

Comparing genomic properties of different species at varying evolutionary distances is a powerful approach for studying biological and evolutionary principles. Because whole-genome sequences are available for a large number of organisms (Bernal \textit{et al.}, 2001), gene and protein sequences have received the most attention as the basis for investigating evolutionary changes. It has been valuable to develop methodologies based on comparative analysis for identifying coding and functional non-coding sequences, as well as sequences that are unique for a given organism [see Frazer \textit{et al.} (2003) for review]. But evolution involves biological variations at many levels and one of the next major steps is to understand how the genes interact to perform particular biological processes. In this context, the accumulation of microarray data from multiple species (Barrett \textit{et al.}, 2005) provides new opportunities to (1) discover how the genes interact to perform specific biological process and (2) to study the evolution of expression network properties. Previous studies have initiated comparative analyses of large datasets of expression profiles from different organisms (Bergmann \textit{et al.}, 2004; Lefebvre \textit{et al.}, 2005; Stuart \textit{et al.}, 2003). Authors generally attempt to consolidate the classical compendium approach (Hughes \textit{et al.}, 2000; Kim \textit{et al.}, 2001) by identifying gene pairs exhibiting co-expression in multiple organisms and across a large number of arrays in each species (Frazer \textit{et al.}, 2003). In spite of very interesting results that demonstrate the potential of cross-species comparisons using expression data, the global approach that consists of the integration of large sets of unrelated microarray data precludes a precise comparison of the sets of genes involved in a particular cellular process. Analyses based on selected experiments that are as close as possible between species, appear to be a more promising approach. They constitute an alternative way to gain in informativity and accuracy in order to understand how conserved are regulatory sub-networks. This principle was first pursued by Alter \textit{et al.} (2003), who compared time points during the cell cycle between yeast and human. But since this pioneer study, the standardization of microarray protocol (Knudsen and Daston, 2005), now accessible to numerous research groups, and the DNA sequences data from closely related species (Dujon \textit{et al.}, 2004; Kellis \textit{et al.}, 2003), lead to a rapid accumulation of new expression data directly comparable between organisms. The challenge of comparing the transcriptional response of several organisms to equivalent environmental conditions is only starting to be addressed, but many aspects of this question are fascinating. For instance, will orthologous genes occupy similar positions in expression networks? To answer this question, it is necessary to (1) be able to analyze the intrinsic structure of a gene expression network and (2) compare the organization of two networks in different organisms. In the present paper, we present an approach based on Multi-dimensional Scaling (MDS) techniques that is...
particularly well-suited for solving these two problems. It permits the representation of the expression data in a space with a reduced dimension [for instance three-dimensional (3D), for visualization] while preserving well the relationships between the genes in highest dimension. The comparison of two gene expression networks is then performed through an original optimization of the superimposition in a multi-dimensional space. In addition, we provide different indexes for assessing the quality of the optimal superimposition. We demonstrate the utility of our approach by a comparative analysis of the transcriptional program that drives the developmental process of sporulation in the yeasts Saccharomyces cerevisiae (Chu et al., 1998) and Schizosaccharomyces pombe (Mata et al., 2002). We conclude that although the transcriptional programs in the two yeasts exhibit common overall properties that certainly mirror the robustness of the gene expression networks during evolution, they are not identical. In particular, orthologous genes were separated according to the conservation of their expression between the two yeasts and we could characterize different degrees of cross-species conservation over time.

2 METHODS

2.1 Notation

For the sake of clarity, we will name the two genomes to be compared \(G_1\) and \(G_2\). Using microarray data, gene expression matrices can be drawn up independently for each genome. The rows correspond to individual genes, the columns are individual microarray experiments and cells contain a measure of the gene activity. Note that the microarray experiments are either time-course gene expression data or gene expression values measured under different experimental conditions. In these matrices, let \(Y_{ij}\) [or \(Y_{i(j)}\)] be the gene expression level of the \(i\)-th gene in the \(n\)-th experiment (or the gene expression value at the time \(t_n\)), for \(i=1, \ldots, G_1\) and \(n=1, \ldots, N_1\) where \(G_1\) denotes the total number of genes in \(G_1\) and \(N_1\) the number of experiments (or time points). Likewise, let \(Y_{jm}\) [or \(Y_{j(m)}\)] be the gene expression level of the \(j\)-th gene in the \(m\)-th experiment (or the gene expression value at the time \(t_m\)), for \(j=1, \ldots, G_2\) and \(m=1, \ldots, N_2\) where \(G_2\) denotes the total number of genes in \(G_2\) and \(N_2\) the number of time points (or experiments). Therefore, each gene can be given a coordinate called its ‘expression profile’ which is, for the \(i\)-th gene, the vector \(Y_i = \{Y_{i1}, Y_{i2}, \ldots, Y_{in}\}\) and for the \(j\)-th gene, \(Y_j = \{Y_{j1}, Y_{j2}, \ldots, Y_{jn}\}\). This coordinate represents a point in an \(N\)-dimensional space, where \(N\) is the number of experiments (\(N_1\) or \(N_2\)).

2.2 Similarity measure between intra-organism gene expression profiles

For each genome (\(G_1\) and \(G_2\)), the matrices of distances between the gene expression profiles (denoted \(D_1\) and \(D_2\)) can be calculated independently. In our study, the ‘inter-gene distances’ are simply Euclidean distances after normalizing each gene expression vector to have a mean 0 and a variance 1. Thus, considering two genes, \(i\) and \(j\), belonging to the genome \(G_1\) the distance between their corresponding expression profiles is:

\[
D_{i,j} = \sqrt{\frac{1}{N_1} \sum_{k=1}^{N_1} (Y_{i,k} - Y_{j,k})^2},
\]

In this case, it can easily be demonstrated that this distance is directly related to the classical Pearson correlation [i.e. \(D^2 = (2N - 1)/N(1 - r)\), where \(D\) and \(r\) denote the Euclidean distance between expression profiles and correlation respectively, and \(N\) the number of measurements of expression]. Note that \(N_1\) and \(N_2\) are not necessarily equal, and therefore the distance is normalized by the number of experiments, \(N\), to avoid systematic bias between genomes \(G_1\) and \(G_2\).

2.3 Representation of a gene expression network in a space of reduced dimension

‘MDS’ is a well-established technique for dimensionality reduction and visualization. It can be used to display the high-dimensional microarray data in a space of lower dimension. Using a distance matrix like \(D_1\) or \(D_2\), this approach enables the transformation of a \(N\)-dimensional space into a \(p\)-dimensional space \((p < N)\), by finding the optimum low-dimensional configuration. The criterion to be optimized, i.e. numerically minimized to find the optimum configuration, maintains the small distances while approximating the large distances. Its expression for the genome \(G_1\) is as follows (\(i \neq j\)):

\[
C = \frac{1}{K} \sum_{i=1}^{G_1} \sum_{j=1}^{G_1} \frac{(D_{i,j} - D_{i,i})^2}{D_{i,j}},
\]

where \(D^4\) denotes the inter-gene distance between expression vectors in the \(N\)-dimensional space and \(D_{i,i}\) the Euclidean distance (also called ‘geometric distance’) in the low-dimensional space between the points corresponding to the \(i\)-th and \(i\)-th genes respectively. The normalization term \(K = \sum_{i=1}^{G_1} D_{i,i}\) transforms the criterion into a dimension-free number.

Consequently, each gene of \(G_1\) (or \(G_2\)) can be located in a low-dimensional space in such a way that the geometric distances in the low-dimensional space are as close as possible to the real distances measured between the gene expression profiles in the \(N_1\) (or \(N_2\))-dimensional space. For each genome, the resulting ‘gene expression network’ reflects the intrinsic structure of the system studied by microarray experiments: two genes with similar expression profiles will be represented by nearby points, whereas two genes with very different patterns of expression will be distant from one another.

2.4 Definition of inter-organism pairs of genes: orthology assignments

To compare the transcriptional responses in different organisms, pairs of related genes, one in each organism, have to be defined. Several types of information can be used. In this study we use orthology relationships to connect expression data and compare two transcriptional programs. We apply the INPARANOID program for generating orthologous gene pairs between genomes \(G_1\) and \(G_2\), as described previously (O’Brien et al., 2005; Remm et al., 2001). To summarize, the algorithm starts the detection of orthologs with calculation of all pairwise similarity scores between the complete sets of protein sequences from the two genomes. This is done with the BLAST program (Altschul et al., 1990). Then, sequence pairs with mutual best hits are detected and serve as central points around which mutual best hits are detected and serve as central points around which overlapping groups are resolved. Note that for some sequences, there is more than one ortholog in one or both species. One gene from genome \(G_1\) can have a significant similarity score with several genes in genome \(G_2\), and vice versa. Our methodology is able to manage these multiple inter-organism links.

2.5 Cross-species comparison of two gene expression networks in a multi-dimensional space

To compare the gene expression networks of two genomes in the same multi-dimensional space, we developed a procedure for optimal superimposition. This process needs first to define inter-organism pairs of genes (\(i, j\)) having correspondence (see previous section). Then, the selected pairs of genes constitute a set of ‘attractor pairs’ in the multi-dimensional space. In other words, the strategy aims at moving one genome (for instance \(G_1\)) towards the other (genome \(G_2\)) such that inter-organism gene pairs are brought as close as possible in the multi-dimensional space, while limiting the distortion of the displaced genome by keeping at the best inter-gene distances of the matrix \(D_{i,j}\) in the translation of points. More precisely, to find
the optimum superimposition configuration, we define a criterion \( E^* \) which is numerically minimized and has two components: the first, \( E \), quantifies the proximity of inter-organism related genes (i.e. the attractor pairs), and the second, \( C \), is a constraint measuring the conservation of the inter-gene distances of the moving genome. The quantity \( C \) corresponds to the initial value of \( C \), i.e. before displacing the point subset. Thus, the expression of \( E^* \) is

\[
E^* = E + \omega(C - C_0),
\]

where \( \omega \) is a weighting. A large \( \omega \)-value leads to a displacement of the set of points of genome \( G_j \), without distorting the expression network. The only 'authorized' displacements are thus orthogonal transformations because of the important constraint of maintaining all the inter-gene distances. In this case, the proximity between attractor genes can remain imperfect. Then, to refine the superimposition between attractors, plastic transformations are allowed by decreasing the \( \omega \)-value. A \( \omega \)-value of 0 in the criterion leads to the perfect superimposition of the attractors in the multi-dimensional space, but also a large distortion of the gene expression network of the moving genome \( G_i \). During the superimposition procedure, the quantity \( E \) is initially maximal and decreases during the displacement whereas the other quantity \( (C - C_0) \) initially equals zero and increases. The \( \omega \)-value is chosen to maintain a balance between the levels of the two components \( E \) and \( (C - C_0) \).

The proximity, \( E \), between inter-organism associated genes depends on the geometric distances in the multi-dimensional space between the points that compose attractor pairs. Its expression is similar to that used in the MDS technique:

\[
E = \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} d_{i,j}^2 / m_{ij},
\]

where \( d_{i,j} \) denotes the geometric distance between the points corresponding to a pair of attractors \((i, j)\) in the multi-dimensional space. The coefficient \( d_{i,j} \) indicates whether a given gene pair is defined as an 'attractor'. It equals 1 when the genes \( i \) and \( j \) are linked (orthologs in our study), otherwise it equals 0. The term \( K = \sum_{i=1}^{K} \sum_{j=1}^{K} D_{i,j} \) allows normalization of the quantity \( E \). We have introduced a particular weighting in the criterion \( E \). The term \( m_{ij} \) is the mean distance between the \( j \)-th gene and its \( i \)-nearest neighbor genes \((j \text{ is fixed at 10 in our study})\). As the geometric distances \( d_{i,j} \) between the attractor points must theoretically equal 0 (if we assume that the gene expression networks are entirely identical), the distances must be weighted by a quantity not biased by the large distances, choosing the choice of this weighting. Note that the component \( E \) can be divided into individual values \( E_i \) for each gene \( i \) in the displaced genome \( G_i \):

\[
E_i = \frac{1}{K} \sum_{j=1}^{K} d_{i,j}^2 / m_{ij}.
\]

The constraint \( C \) is identical to that used in the above MDS technique \((i \neq j)\):

\[
C = \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} (d_{i,j} - D_{i,j})^2 / D_{i,j}^2
\]

with

\[
K = \sum_{i=1}^{K} \sum_{j=1}^{K} D_{i,j}.
\]

### 2.6 The optimization result and its significance

The reduction of an \( N \)-dimensional space into a \( p \)-dimensional space \((p < N)\) involves a loss of information. The optimization procedure approximates, for each pair of genes \((i, j)\), the inter-gene distance \( D_{i,j} \) between the gene expression profiles to a geometric distance \( d_{i,j} \) between points in the \( p \)-dimensional space. Thus, considering a pair of genes \((i,j')\) belonging to the same genome, say \( G_i \), we can calculate the following indexes:

\[
e_{i,j'} = \frac{|d_{i,j'} - D_{i,j'}|}{\sqrt{D_{i,j'}}},
\]

and

\[
\Delta_i = \sum_{j} e_{i,j}^2.
\]

A low \( e_{i,j'} \) value means that the points corresponding to the genes \( i \) and \( j' \) are correctly mapped in the multi-dimensional space according to the measures of their gene expression. Moreover, the \( \Delta \)-value enables to capture the relevance of the position of a particular gene \( i \) in the light of all the other genes of genome \( G_i \). Note that the constraint \( C \) is \( \sum_{i} \Delta_i / K \) (see Equation \( 6.1 \)).

These two indices can be calculated after the initial mapping of genes into a multi-dimensional space, thus assessing the loss of information during dimensionality reduction. But they can also be used to quantify, for each gene, the local distortion of the gene expression network during the superimposition procedure described above. For instance, if a gene \( i \) is attracted in a direction that does not agree with its position in the gene expression network, it will result in a large local deformation of the network and a substantial increase in its \( \Delta \)-value (\( \Delta_{\text{initial}} = \Delta_{\text{optim}} - \Delta_{\text{initial}} \)). In contrast, if a gene is attracted in a direction that agrees with its position in the gene expression network, there will only be a small local deformation of the network and a small increase in its \( \Delta \)-value.

### 2.7 Technical information

All programs were implemented in the R programming language (http://cran.r-project.org/). Functions were numerically minimized using the quasi-Newton method (R function available in the BASE package). All calculations were carried out on a Dell precision 360 equipped with a 2.6 GHz processor running a Linux DEBIAN distribution. 3D visualizations and snapshots were performed using the RGL package (http://rgl.neosciencers.org/News.html).

### 2.8 A case study of sporulation: comparison of the transcriptional programs in yeasts

To illustrate our approach, we compared two species: the budding yeast *Sacharomyces cerevisiae* and the fission yeast *S.pombe*. Both have been entirely sequenced (Goffeau et al., 1996; Wood et al., 2002) and comprehensively annotated. These two model organisms were chosen because two very similar sets of microarray time-course experiments (one for each organism) were available. Indeed, two different laboratories have used DNA microarrays to study the transcriptional program that drives the developmental process of sporulation, in which diploid cells undergo meiosis to produce haploid germ cells. The first study, published in 1998 (Chu et al., 1998) led to a precise description of different temporal gene expression patterns of about 500 genes induced in *S.cerevisiae*. More recently, it has been demonstrated that around 1000 genes are also regulated in successive waves of transcription in *S.pombe* (Mata et al., 2002) during meiosis and sporulation. Both Chu et al. (1998) and Mata et al. (2002) measured the changes in the concentrations of mRNA transcripts for each gene at successive times after transfer of wild-type diploid yeast cells to a nitrogen-deficient medium that induces sporulation. We have collected the measured log(Ratio) expression values for each time point: 7 for *S.cerevisiae*, t\(_c\): [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 h] and 13 for *S.pombe*, t\(_c\): [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 h].
for the two species are similar (see also Supplementary Figure S2). For instance, we can clearly distinguish in both cases a highly connected region from all other genes, and there is a similar temporal organization (see below).

Temporal organization of sporulation in yeast. Genes induced during the time-course experiments were classified according to their expression profiles (Quackenbush, 2001) into four temporal classes by k-means clustering (Fig. 1B). In both species, these four temporal classes coincide with the major biological processes of sporulation as described in Mata et al. (2002): response to nutritional changes (‘Nitrogen’ class of genes), meiotic S phase and recombination (‘Early’ class), meiotic division (‘Middle’ class) and sporulation (‘Late’ class). To visualize the successive waves of transcription during the sporulation process on the 3D representation, each gene was colored according to its temporal class, showing that genes can be placed in a sequence that reflects the time of initial induction during the time-course (see also Supplementary Figure S3).

Common properties of the transcriptional programs of sporulation between the two yeasts. Taken at face value, the parallel observation of comparable temporal classes in S.pombe and S.cerevisiae (Fig. 1B) suggests that the overall process of sporulation is identical in fission and budding yeasts. However, to address evolutionary aspects, it is valuable to determine whether ‘orthologous’ genes in the two yeasts are implicated in the biological processes of sporulation. Orthology defines the relationship between genes in different species that originate from a single gene in the last common ancestor of these species (Fitch, 2000; Sonnhammer and Koonin, 2002). Such gene pairs are most likely to share the same function and should, in principle, occupy a similar position in the gene expression networks derived from the sporulation process. To test this hypothesis, the structure of the orthologous gene expression networks were compared in a 3D space, in order to systematically characterize both similarities and differences across species.

3.2 Cross-species comparison of orthologous genes

Gene orthology and sporulation in S.cerevisiae and S.pombe. Orthology relationships among the genes were inferred using the INPARANOID algorithm (O’Brien et al., 2005; Remm et al., 2001) (see Methods). We detected 5053 orthologous gene pairs between the complete genomes of S.cerevisiae and S.pombe. Then, we determined whether orthologous genes induced during sporulation in S.cerevisiae (518 genes) were also up-regulated in S.pombe (724 genes). We identified 122 orthologous gene pairs, which correspond to $G_1 = 85$ S.cerevisiae genes and $G_2 = 98$ S.pombe genes. We thus constructed a set of orthologs between the two yeasts, all implicated in the sporulation process, and serving as the kernel for the following cross-species comparison. All the other genes, i.e. those with no identifiable orthologous counterpart represent a set of ‘organism-specific’ genes (see Discussion).

Optimal superimposition in a 3D space. First, each set of orthologous genes was independently represented in a 3D space according to the expression measurements (Fig. 2A). For both species, highly connected regions can still be distinguished from all other genes. We also checked that the gene distribution into the previously characterized temporal classes was unchanged following selection of inter-organism gene pairs (Supplementary Materials S4). Next, the 3D-gene expression networks were mapped in the

Fig. 1. 3D-gene expression networks of yeast sporulation data (A) The 724 and 518 genes significantly induced during the sporulation process in the fission (green) and budding (blue) yeasts were mapped in a 3D space, according to the expression measurements during the time-course experiments described in Mata et al. (2002) and Chu et al. (1998). Points represent genes, and white segments link pairs of genes for which the inter-gene distance is less than a threshold $D_0 = 0.3$ (a conventional correlation >0.95). (B) Using the k-means algorithm, genes were classified according to their expression profiles into four classes named ‘Nitrogen’ (blue), ‘Early’ (green), ‘Middle’ (red) and ‘Late’ (yellow). Genes contained in each group are shown as different colors so that they can be identified in the gene expression network. Unlike in (A), only the first and the second coordinates are presented here.

>1.13). According to the notations defined in Methods, the resulting expression matrices comprised $G_1 = 518$ genes and $N_1 = 7$ time points for the S.cerevisiae genome ($G_1$), and $G_2 = 724$ genes and $N_2 = 13$ time points for the S.pombe genome ($G_2$).

Representations in a 3D space show similar organization in the two yeasts. In this case study of sporulation, we decided to use a 3D space for two reasons. First, principal component analyses applied to the S.cerevisiae and S.pombe sporulation data (Chapman et al., 2002) showed that, in both cases, >93% of the variability is found in three dimensions (Supplementary Figure S1). 3D representations are thus good pictures of the high-dimensional representations of the gene expression networks. Second, it has the advantage of allowing the visualization of each step of the cross-species comparison without any subsequent transformations. Therefore, we mapped the yeast sporulation data in a 3D space using the MDS procedure for dimensionality reduction described in Methods. Results are presented in Figure 1A. Note that because there are more measurements in the S.pombe (13 points) than in the S.cerevisiae (7 points) time-course, the pairwise relationships between gene expression profiles are better discriminated in the S.pombe 3D-gene expression network. Nevertheless, Figure 1A suggests that the overall organizations of the 3D-gene expression networks of the two species are similar (see also Supplementary Figure S2).
same space and associated genes between the two yeasts were connected with red segments (Fig. 2B). The length of a segment is a good visual indicator of the spatial proximity of orthologs. Then, to analyze the concordance between orthology relationship and expression data, we modified the superimposition of the two 3D-gene expression networks shown in (A). Orthologous gene pairs are connected with red segments. (C) Result of the superimposition after the optimization procedure described in Methods. Only the \textit{S. cerevisiae} genome (blue) was displaced, with a \( \omega \)-value = 2. During the optimization, the \textit{S. cerevisiae} gene expression network was completely turned over. For instance, the genes YJR152W and YIR029W, initially located on the right side were finally found on the left side (vertical arrows). Likewise, the gene YHR087W, initially located at the top of the gene expression network was finally found at the bottom (horizontal arrow). Enlargements of each picture are available online at http://www.biologie.ens.fr/lgmgml/publication/comp3d/figure2.php.

Fig. 2. Cross-species comparison of orthologous gene pairs. (A) Orthologous genes between the budding and fission yeasts were independently represented in a 3D space according to their expression measurements during the time-course experiments described in Mata \textit{et al.} (2002) and Chu \textit{et al.} (1998). Points represent the 98 and 85 genes belonging to the \textit{S. pombe} (green) and \textit{S. cerevisiae} (blue) genomes respectively. White segments link pairs of genes for which the inter-gene distance between their expression measurements is less than the threshold \( D_0 = 0.3 \) (a conventional correlation >0.95). (B) Initial superimposition of the 3D-gene expression networks shown in (A). Orthologous gene pairs are connected with red segments. (C) Result of the superimposition after the optimization procedure described in Methods. Only the \textit{S. cerevisiae} genome (blue) was displaced, with a \( \omega \)-value = 2. During the optimization, the \textit{S. cerevisiae} gene expression network was completely turned over. For instance, the genes YJR152W and YIR029W, initially located on the right side were finally found on the left side (vertical arrows). Likewise, the gene YHR087W, initially located at the top of the gene expression network was finally found at the bottom (horizontal arrow). Enlargements of each picture are available online at http://www.biologie.ens.fr/lgmgml/publication/comp3d/figure2.php.

Expression in the two yeasts and sequence conservation are significantly correlated. The superimposition of the two 3D-gene expression networks was optimized. Figure 2C gives a graphical representation of the cross-species comparison result: expression data and sequence conservation (i.e. orthologous links) are connected in the same space, allowing a quantitative assessment of the similarities and differences between the transcriptional programs of the two yeasts. Note that the highly connected regions of the two networks are superimposed. This is because of the numerous attractor pairs that linked them. It demonstrates that these orthologs occupy similar positions in the gene expression networks.
of the two components was performed with different <i>e</i>-values, <i>e</i> is a weighing required to maintain a balance between the levels of the two components <i>E</i> and (<i>C</i> – <i>C</i>0) (see main text). Grey lines represent the <i>E</i>(random) and (<i>C</i> – <i>C</i>0)(random) values, obtained by reshuffling the ortholog gene list and are expressed similarly in the two species. Then, to assess the significance of overall 3D-proximity between orthologs after optimization, we used a statistical analysis comparing the criterion <i>E</i>. This method is based on a bootstrap procedure which consists of comparing the final proximity of orthologous gene pairs, <i>E</i>(real), to random controls, <i>E</i>(random), obtained by superposing the two gene-expression networks using randomly rearranged orthologous links (Fig. 3). For all values of <i>e</i>, <i>E</i>(real) was >3 SD lower than the mean of the <i>E</i>(random) distributions. More precisely, for a <i>e</i>-value of 2, the calculated Z-score was −4.7 (Fig. 4) meaning that orthology relationships and expression during the sporulation process have significant concordance between <i>S</i>.<i>cerevisiae</i> and <i>S</i>.<i>pombe</i>.<br><br>Fig. 3. <i>E</i> and (<i>C</i> – <i>C</i>0) versus <i>e</i>. The optimal superimposition of the 3D-gene expression networks shown in Figure 2 was performed with different <i>e</i>-values. <i>e</i> is a weighting required to maintain a balance between the levels of the two components <i>E</i> and (<i>C</i> – <i>C</i>0) (see main text). Grey lines represent the <i>E</i>(random) and (<i>C</i> – <i>C</i>0)(random) values, obtained by reshuffling the ortholog gene list and are expressed similarly in the two species. Then, to assess the significance of overall 3D-proximity between orthologs after optimization, we used a statistical analysis comparing the criterion <i>E</i>. This method is based on a bootstrap procedure which consists of comparing the final proximity of orthologous gene pairs, <i>E</i>(real), to random controls, <i>E</i>(random), obtained by superposing the two gene-expression networks using randomly rearranged orthologous links (Fig. 3). For all values of <i>e</i>, <i>E</i>(real) was >3 SD lower than the mean of the <i>E</i>(random) distributions. More precisely, for a <i>e</i>-value of 2, the calculated Z-score was −4.7 (Fig. 4) meaning that orthology relationships and expression during the sporulation process have significant concordance between <i>S</i>.<i>cerevisiae</i> and <i>S</i>.<i>pombe</i>.<br><br>Fig. 4. Statistical analysis comparing the criterion <i>E</i> to random controls. The final <i>E</i>-value (with <i>e</i> = 2), was decomposed into its four components <i>E</i><sub>Nitrogen</sub>, <i>E</i><sub>Early</sub>, <i>E</i><sub>Middle</sub> and <i>E</i><sub>Late</sub>. Each value (vertical line) was compared with random controls obtained by reshuffling the lists of orthologous genes. Means <i>m</i> of the random distributions are indicated, with the calculated Z-scores in square brackets.<br><br>3.3 Budding and fission yeasts show different degrees of cross-species conservation over the sporulation process<br><br>Comparative analysis of the temporal classes: particularities of the ‘Early’ genes. In the final superimposition (Fig. 2C) several inter-organism associated genes are distant from one another after the optimization procedure. This suggests that some genes with sequence conservation have different locations in the 3D-gene expression networks in the two species, i.e. different roles during the sporulation process. We first examined whether the overall concordance between sequence conservation and expression (see previous paragraph) is similar for the four temporal classes of genes: ‘Nitrogen’, ‘Early’, ‘Middle’ and ‘Late’. The criterion <i>E</i> was decomposed into four components: <i>E</i><sub>Nitrogen</sub>, <i>E</i><sub>Early</sub>, <i>E</i><sub>Middle</sub> and <i>E</i><sub>Late</sub>, representing the contribution of genes belonging to the same temporal class, in the displaced genome <i>G</i><sub>1</sub> (<i>E</i> = <i>E</i><sub>Nitrogen</sub> + <i>E</i><sub>Early</sub> + <i>E</i><sub>Middle</sub> + <i>E</i><sub>Late</sub>). Each component was then compared with random controls obtained by reshuffling the orthologous gene list (Fig. 4). For the classes ‘Nitrogen’, ‘Middle’ and ‘Late’, the calculated values were >3 SD lower than the mean of the corresponding random distributions. Findings for the ‘Early’ class were, however, different. The <i>E</i><sub>Early</sub> value was +0.5 SD higher than the mean of the random distribution, indicating that there was no significant concordance between sequence conservation and expression in the two yeasts. Interestingly, this finding agrees with a previous study (Mata and Bahler, 2003), which described the important role in speciation for genes induced during meiotic prophase, when homologous chromosomes pair and recombine. Indeed, organism-specific proteins with roles in homologous chromosome pairing during meiotic prophase may help to prevent fruitful meiosis between closely related organisms, and they may thus favor speciation.<br><br>Similarities and differences in ortholog expression. Next, we scrutinized each orthologous gene pair. To find systematically the most interesting patterns (similar or dissimilar), we classified the <i>S</i>.<i>cerevisiae</i> genes according to their individual <i>E</i><sub>i</sub> values in <i>E</i>. A large <i>E</i><sub>i</sub> value means that the genes <i>i</i> and <i>j</i> are distant in the final 3D space, because of the constraint of limiting the distortion of the displaced gene expression network. In these cases, the two orthologs (<i>i</i>,<i>j</i>) exhibit different locations in the two gene expression networks. A low <i>E</i><sub>i</sub> value means that the genes <i>i</i> and <i>j</i> are close and thus presumably exhibit similar expression profiles. To avoid false positive results, it is necessary to verify that the proximity of two genes does not result from major distortion of the displaced gene expression network (<i>S</i>.<i>cerevisiae</i>). For this reason we defined an index, <i>Δ</i>, (see Methods and Supplementary Materials S6), which quantifies for each gene <i>i</i> (attracted to the area where its related gene <i>j</i> is located) the resulting local deformation of the initial gene expression network after the superimposition procedure. Each of the 85 genes from the <i>S</i>.<i>cerevisiae</i> genome were plotted according to their <i>E</i> and <i>Δ</i> values (Fig. 5A). Average <i>E</i><sub>i</sub> values were computed in cases of multiple links. On the resulting graph, the lower left group includes <i>S</i>.<i>cerevisiae</i> genes with expression profiles similar to those of their <i>S</i>.<i>pombe</i> orthologs (Fig 5B and 1–3); other genes,
two species, despite conserved functions, revealing the plasticity of the functional modules. Comparing genomic properties of different organisms is of fundamental importance and our approach yielded new information on essential genes. Several examples, which exhibit characteristic situations and illustrate the potential of combining sequence and expression data to address particular evolutionary issues, are presented in Supplementary Materials S8 and S9 (http://www.biologie.ens.fr/lgmgml/publication/comp3dSupData.php).

4 DISCUSSION

Direct comparison of gene expression networks involving similar biological processes in two different organisms is still in infancy. One of the main obstacles is the difficulty of normalizing data that have been obtained in different experimental conditions. In this paper we present a new methodology based on MDS techniques which allows graphical visualization as an integral part of statistical modelling and data analysis. We believe that this multi-dimensional space representation of whole-genome expression data can be considered as a useful topology model to stimulate both intuitive and rational comparison of orthologous sets of data. Simulations, performed to evaluate the ability of our approach to identify orthologous pairs of genes whose expression is conserved have led to promising results (see Supplementary Materials S10). The procedure for optimal superimposition has a good tolerance to noise and is able to precisely identify classes of genes whose expression is well-conserved between two species. Our method relies, of course, on the availability of accurate gene expression data. Therefore, unlike previous cross-species comparative analyses of large datasets of expression profiles (Bergmann et al., 2004; Stuart et al., 2003), we decided in common with Alter et al. (2003) and McCarroll et al. (2004) to focus our approach on strictly comparable datasets. The choice of the sporulation data in the fission (Mata et al., 2002) and budding (Chu et al., 1998) yeasts was favored because the same nutritional stress triggers the two sporulation programs and accurate time-series experiments describe the connected transcriptional events. We also compared the gene expression networks of the yeasts S. cerevisiae and S. pombe, using the cell cycle data described in Spellman et al. (1998) and Rustici et al. (2004) (see Supplementary Materials S11).

The purpose of this work was to present a rational approach to the scrutiny of the differences and similarities in these orthologous gene expression programs, taking into account all the intra- and inter-species relationships between genes. Genes contained in each temporal class are shown in a different grey scale on the plot. Genes belonging to the ‘Middle class’ (located in the highly connected regions) and implicated in the meiotic division, for example B-type cyclins and components of the Anaphase Promoting Complex (Hartwell et al., 1974), are over-represented in the set of genes with similar expression in the two yeasts. This may reflect a particular conservation for cell-cycle genes, involved in a core eukaryotic process (Rustici et al., 2004; Spellman et al., 1998). This analysis also reveals genes that have sequence conservation, but display different expression pattern. It demonstrates that despite an overall conservation, the sporulation process is not strictly identical in the two yeasts.

Integration of sequence and expression data provides new insight into cross-species comparisons. Using our methodology, we were able to identify differences between the sporulation processes of the

---

**Fig. 5.** Extraction of similarities and differences from the final 3D superimposition. (A) Biplot representing each S. cerevisiae gene according to its final \( E_i \) (x-axis) and \( \Delta_i \) (y-axis) values (see Methods). In cases of multiple links, the average \( E_i \) value was calculated. Genes contained in each temporal classes are shown in different grey scale on the plot. (B) Expression profiles of four S. cerevisiae genes (ATG8, DAL5, A2R1 and ADY2), underlined on the biplot shown in (A). The expression profiles of their orthologous genes in S. pombe are also represented to allow direct visual comparison, with the correlation measure between the expression values that are common to the S. pombe and S. cerevisiae time series in square brackets (see Supplementary Figure S4).

those with high \( E_i \) or \( \Delta_i \) values, exhibit gene expression profiles different from those of their associated genes in S. pombe (Fig 5B, 4 and 5). Note that the case of the orthologous gene pair number 3 (YHR087W and SPBC21C3.19) deserves a special mention. Even if the expression profiles are not strictly identical (the correlation value is only 0.36), their corresponding final \( E_i \) and \( \Delta_i \) values are low. Indeed, these orthologous genes are both induced at the end of the time course and thus occupy a similar external position in the S. pombe and S. cerevisiae gene expression networks (Fig 2, horizontal arrow). This demonstrates the usefulness of our approach, which goes beyond a simple direct comparison of the gene expression profiles for orthologous pairs (Supplementary Figure S7). It integrates expression data with gene orthology to compare the global organization of two gene expression networks, thus taking into account all the intra- and inter-species relationships between genes.

Genes contained in each temporal class are shown in a different grey scale on the plot. Genes belonging to the ‘Middle class’ (located in the highly connected regions) and implicated in the meiotic division, for example B-type cyclins and components of the Anaphase Promoting Complex (Hartwell et al., 1974), are over-represented in the set of genes with similar expression in the two yeasts. This may reflect a particular conservation for cell-cycle genes, involved in a core eukaryotic process (Rustici et al., 2004; Spellman et al., 1998). This analysis also reveals genes that have sequence conservation, but display different expression pattern. It demonstrates that despite an overall conservation, the sporulation process is not strictly identical in the two yeasts.

Integrating sequence and expression data provides new insight into cross-species comparisons. Using our methodology, we were able to identify differences between the sporulation processes of the

---
orthologous relationships between sporulation-induced genes (see Supplementary Figure S12). In addition it is important to emphasize that the 3D space, used here for the sake of simplicity, can easily be extended to a multi-dimensional space, opening the way to deeper analyses. Nevertheless, the 3D space representation will remain a useful visualization tool to point out the most interesting features.

One of the long-term objectives of this work is to develop, from gene expression profile data, automatic approaches for characterizing groups of genes which are involved in similar biological functions in different organisms. In that respect our 3D transcriptome representation can be considered as analogous to 3D protein structure in which we try to identify specific structural motifs (Chou, 2004). Whether rigorous modeling of expression profile comparisons between genetic programs of multiple organisms is possible on a genomic scale remains to be seen. Designing strictly comparable transcriptome analyses to start from homogeneous data is certainly an important prerequisite to achieve this goal. Nevertheless, the data will not provide new insights automatically. They will require new tools to be rigorously scrutinized. Thus, for instance, identification of orthologous groups of expression will have to be connected with transcriptional sub-networks. Our approach may represent one of the first steps in this direction.

ACKNOWLEDGEMENTS

The authors wish to thank Catherine Etchebest for helpful advice and comments on the manuscript. Frédéric Devaux and Alexandre de Brevens for discussions and Boris Barbour for English correction. This work was supported by ACI-IMPBIO-2004-45. This paper is dedicated to the Prof. Serge Hazout, who died in April 2005.

Conflict of Interest: none declared.

REFERENCES