BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 8 2005, pages 1550–1558
doi:10.1093/bioinformatics/bti215

Gene expression

Balancing protein similarity and gene co-expression reveals new links between genetic conservation and developmental diversity in invertebrates

Céline Lefebvre[1,2], Jean-Christophe Aude[3], Eric Glémet[2] and Christian Néri[1,*]

[1]INSERM, Avenir Group, Laboratory of Genomic Biology, Centre Paul Broca, 75014 Paris, France, [2]Gene-IT S.A., 92500 Rueil-Malmaison, France and [3]CEA, Département de Biologie Joliot-Curie, Service de Biochimie et de Génétique Moléculaire, Saclay, 91191 Gif-sur-Yvette Cedex, France

Received on September 29, 2004; revised on November 17, 2004; accepted on December 8, 2004
Advance Access publication December 14, 2004

ABSTRACT

Motivation: To identify genetic conservation relative to precise aspects of developmental diversity, an essential question in computational biology, we developed a new comparative method that allows conserved modules for the best balance between protein sequence similarity and gene co-expression to be constructed, in invertebrates.

Results: Our method, referred to as the best-balance constraint procedure (BBCP), yielded 719 functionally conserved modules (FCMs) comprising 2–23 gene pairs. These modules were consistent with the developmental roles of orthologues as inferred from Gene Ontology, RNAi knockouts, InterPro and process-specific microarray data. New relationships were defined between genetic conservation and developmental diversity. Novel gene associations were indeed found in 94% of the FCMs, 150 modules being completely new. A significant proportion of the FCMs (18%, 132 modules) described cell type-specific mechanisms, comprising neuronal, muscle and germ cell signaling, new associations being found in 125 modules. Also found were gene associations for cell fate specification activities previously not highlighted by computational means, e.g. in FCMs containing homeogenes. These data indicate that highly discriminative description of genetic conservation can be deduced using BBCP, and reveal new correlations between cellular and developmental diversity and gene essentiality in invertebrates.

Contact: christian.neri@broca.inserm.fr

Supplementary information: For supplementary information, please refer to *Bioinformatics* online.

INTRODUCTION

To compare the development of an organism with another, it is important to consider the properties of developmental systems (Rudel and Sommer, 2003) and the biological roles of genes and their encoded products in the context of complex gene regulatory networks (Davidson *et al*., 2002; Hinman *et al*., 2003; Rast, 2003). The availability of genome sequences and genome-wide biological attributes provides a large amount of information that can be analyzed *in silico* in order to detect developmental genetic mechanisms that may be conserved among different organisms. In this respect, the analysis of gene co-expression was applied to the comparison of two or more different organisms (Teichmann and Babu, 2002; Stuart *et al*., 2003; van Noort *et al*., 2003; Bergmann *et al*., 2004; McCarroll *et al*., 2004). One of these studies reported a global decomposition of conserved expression from six evolutionarily distant organisms, and was based on modules that do not necessarily contain only homologous proteins (Bergmann *et al*., 2004). The other studies were focused on functional conservation. Two studies compared *Saccharomyces cerevisiae* and *Caenorhabditis elegans*, defining orthologous relationships using either the best reciprocal hit approach (Teichmann and Babu, 2002) or phylogenetic trees (van Noort *et al*., 2003). Stuart *et al*. (2003) performed a global clustering of genetic conservation in yeast, *C.elegans*, *Drosophila melanogaster* and human, using pairs of orthologous proteins as defined by best reciprocal hits. McCarroll *et al*. (2004) analyzed aging in *C.elegans* and *Drosophila* using one-to-one orthologs as generated by combining BLASTP, phylogenetic trees and Smith–Waterman alignments. Considering the best orthologs may, however, not allow conserved modules to be retrieved at high sensitivity for two reasons. First, complex biological processes may involve a strong inter-specific diversity among organisms, which implies that orthologous proteins with rather divergent sequences may be involved in similar biological processes or molecular functions. This is for example the case with Hox genes in regulating morphogenesis (Aboobaker and Blaxter, 2003) or members of the nuclear receptor family in regulating the function of *C.elegans* lateral hypodermal cells (Miyabayashi *et al*., 1999). Second, the search for best orthologs may often return a gene that is not the nearest phylogenetic neighbor of the query sequence (Koski and Golding, 2001). Thus, we hypothesized that the computational analysis of conserved biological processes might gain in informativity and accuracy if performed by applying a new method, referred to here as the 'best-balance constraint procedure' (BBCP), that takes into account the influence of protein similarity and gene co-expression with no a priori. To test our hypothesis, we applied the BBCP method to the comparison of *C.elegans* (The *C.elegans* Sequencing Consortium, 1998; Kim *et al*., 2001) and *Drosophila* (Adams *et al*., 2000; Arbeitman *et al*., 2002). To test for BBCP validity and discriminating power, we used Gene Ontology (GO) annotations (Ashburner *et al*., 2000), InterPro domain annotations (Mulder *et al*., 2003), RNAi phenotypes resulting from genome-wide analyses of *C.elegans*

*To whom correspondence should be addressed.

or *Drosophila* gene function (Ashrafi *et al*., 2003; Kamath *et al*., 2003; Simmer *et al*., 2003; Boutros *et al*., 2004; Nollen *et al*., 2004), process-specific microarray data (De Gregorio *et al*., 2002; Gaudet and Mango, 2002; Klebes *et al*., 2002; Mallo *et al*., 2002; Romagnolo *et al*., 2002; Zhang *et al*., 2002; Lee *et al*., 2003; Wang and Kim, 2003; Roxstrom-Lindquist *et al*., 2004) and information reported in the scientific literature. We conclude that our approach yielded new instructive information on essential genes, a large proportion of them describing new aspects of cellular and physiological diversity in *C.elegans* and *Drosophila*.

METHODS

Construction of functionally conserved modules

We developed a new comparative approach, the BBCP method, as described below. Functional conservation may be associated with several constraints during evolution. These constraints may result in a limited number of alternatives to achieve a given biological process, which may be reflected by conservation at different levels, e.g. at the level of gene sequence, gene regulation or pathway activity. The analysis of these phenomena thus requires methods based on the combination of different information. Previous work in this respect combined sequence information with information about pathway topology (Forst and Schulten, 1999, 2001). Other examples comprise the alignment of protein interaction networks across species through the analysis of sequence similarity (Kelley *et al*., 2004) and co-clustering of biological networks and yeast gene-expression data by combining different distances (Hanisch *et al*., 2002). Here, we considered two factors for defining functional conservation, namely the similarity in protein sequence and the similarity in the regulation of gene expression, the latter being reflected by co-expression. The influence of these two factors was accounted for by combining two dissimilarity measures, a measure for protein sequence similarity (δ_{seq}) and a measure for gene co-expression (δ_{exp}), into a single dissimilarity measure, referred to here as the dissimilarity measure for conservation Δ,

$$\Delta = f(\delta_{\text{seq}}, \delta_{\text{exp}}) \tag{1}$$

that describes functional conservation as modules containing four genes (Fig. 1A). To apply this formula to the comparison of one organism (geneset X) to another (geneset Y), we considered gene pairs showing a similar protein sequence (x, y) and (x', y') and gene pairs showing co-expression (x, x') and (y, y'). The function $\delta_{\text{seq}} : X \times Y \rightarrow [0, 1]$ is defined below:

$$\delta_{\text{seq}}(x, y) = \begin{cases} t \cdot [\text{score}(x, y)]^{-1} & \text{if score} \geq t \\ 1 & \text{otherwise,} \end{cases} \tag{2}$$

where score(x, y) is an arbitrary, positive, sequence pair alignment scoring function, and t is the threshold. Distances between expression patterns are usually calculated with a Pearson correlation coefficient ρ. As we were only interested in similar expression patterns, we introduced a positive threshold t_x. The function $\delta_{\text{exp}} : X \times X \rightarrow [0, 1]$ is defined below:

$$\delta_{\text{exp}}(x, x') = \begin{cases} 1 - \rho(x, x') & \text{if } \rho \geq t_x \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

We then computed Δ using a function that (1) accounts for the influence of protein sequence similarity and that of gene co-expression with no a priori, and (2) results in a dissimilarity measure that may be used as an input for data clustering. One function that fulfills these requirements and is capable of balancing δ_{seq} and δ_{exp} equally, is the sum of the two dissimilarity measures. The function $\Delta : ((X \times Y) \times (X \times Y)) \rightarrow [0, 4]$ is defined below:

$$\Delta((x, y), (x', y')) = \begin{cases} 0 & \text{if } x = x' \text{ and } y = y' \\ \delta_{\text{seq}}(x, y) + \delta_{\text{seq}}(x', y') & \\ \quad + \delta_{\text{exp}}(x, x') + \delta_{\text{exp}}(y, y') & \text{otherwise.} \end{cases} \tag{4}$$

As Δ combines two dissimilarity measures, it is symmetrical and reflexive (the demonstration, not given here, is straightforward), two properties that make it sufficient for use as an input for data clustering.

The focus of the dissimilarity measure combination defined by the function Δ is to compare functional features (here sequences and gene expression) as variables modeled by the evolutionary pressure. This analysis was applied to the comparison of *C.elegans* and *Drosophila* genes as illustrated in Figure 2. First, we defined orthologous relationships between proteins of the two organisms (The *C.elegans* Sequencing Consortium, 1998; Adams *et al*., 2000). Protein sequences for *C.elegans* were retrieved from the Sanger Institute website (wormpep89), and those for *Drosophila* from Flybase release 3 (Flybase Consortium, 2003). The all-by-all protein comparison was performed using the zval algorithm (Comet *et al*., 1999; Aude and Louis, 2002) of the Biofacet™ software (Glemet and Codani, 1997) set to default parameters (gapo 50, gape 3, cutoff 220, matrix Dayhoff 10/3). This algorithm computes the Z-value, which is based on a Monte-Carlo approach to estimate the significance of a Smith–Waterman alignment score (Smith and Waterman, 1981). In contrast to alignment scores, Z-values reduce the biases due to sequence length and composition (Comet *et al*., 1999) and are independent of the size of the database being queried. The zval algorithm has been used for massive analysis of protein families as reported, e.g. in the CluSTr database, an automatic classification of UniProt Knowledgebase proteins into groups of related proteins (Kriventseva *et al*., 2001). Setting the Z-value threshold to 10 corresponds to a statistical alpha-risk (risk of wrongly concluding that there is a difference when really there is none) of 1% (Bastien *et al*., 2004). If using the Z-value as a score function, Equation (2) becomes

$$\delta_{\text{seq}}(x, y) = \begin{cases} 10 \cdot [\text{Zval}(x, y)]^{-1} & \text{if Zval} \geq 10 \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

and allows a group of 177 944 homologous relationships involving 10 625 (52%) *C.elegans* and 9261 (67%) *Drosophila* proteins to be obtained.

Second, we generated gene expression clusters using one expression dataset for *C.elegans* (Kim *et al*., 2001), and one for *Drosophila* (Arbeitman *et al*., 2002). The *C.elegans* dataset was generated using cDNA microarrays that contained 17 817 genes and corresponded to 553 experiments involving whole animal RNA (Kim *et al*., 2001). The *Drosophila* dataset was generated using cDNA microarrays that contained 4028 genes and corresponded to wild-type flies examined during 66 sequential time periods from fertilization to the first 30 days of adulthood using whole animal RNA (Arbeitman *et al*., 2002). We identified 48 088 homologous relationships involving encoded products for which gene-expression data were available, and corresponding to 6222/10 625 (58.5%) *C.elegans* and 2781/9261 (30%) *Drosophila* genes. We applied hierarchical clustering (Johnson, 1967) to these microarray datasets using the amap package from the R statistical language and environment package version 1.8.1 (http://www.r-project.org/). We calculated a distance matrix that contained all possible pairs of genes by applying the Pearson correlation coefficient formula. We then computed a hierarchical clustering on this matrix, using the *Ward*'s minimum variance agglomeration method (Ward, 1963). Final gene expression clusters were constructed in two steps. First, we elected to calculate the optimal cluster partition. This was performed using two different validation indices, namely the Dunn's and Silhouette indices (Bolshakova and Azuaje, 2003). This resulted in 97 clusters containing 35–372 *C.elegans* genes (mean size: 158 genes) and 82 clusters containing 8–185 *Drosophila* genes (mean size: 49 genes). Second, we considered the Pearson correlations (ρ) between genes belonging to the same cluster, and we assessed their statistical significance using Student's t-test $[t = \rho((n-2)/(1-\rho^2))^{1/2})]$ with $(n-2)$ degrees of freedom, where n is the number of genes for each organism. Among the 1 437 118 associations for *C.elegans* and 131 434 associations for *Drosophila*, 21 050 correlations that would occur by chance for *C.elegans* (1.5%) and 767 for *Drosophila* (0.6%) had to be excluded for a t-test significance level of 10^{-4}. The remaining Pearson correlations reflected positively correlated genes, and their value was in the range [0,1], thus allowing δ_{exp} values to be in the same range.

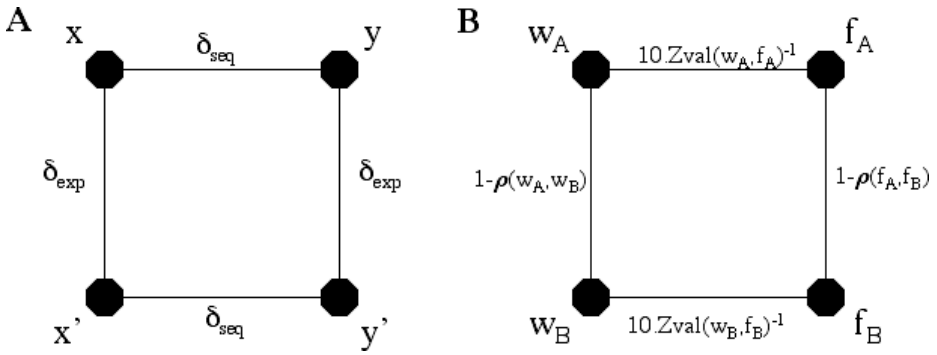

Fig. 1. Formal representation of dissimilarity measure combination in the BBCP procedure. Functional conservation is inferred, with no a priori, from the influence of protein similarity (dissimilarity measure δ_{seq}) and co-expression (dissimilarity measure δ_{exp}) for genes x, y, x' and y' **(A)**. When comparing *C.elegans* (W) and *Drosophila* (F), protein sequence similarity is defined by the Z-value for (w_A, f_A) and (w_B, f_B), while co-expressed genes (w_A, w_B) and (f_A, f_B) share a Pearson correlation (ρ) and are defined by the dissimilarity measure $1 - \rho$ **(B)**.

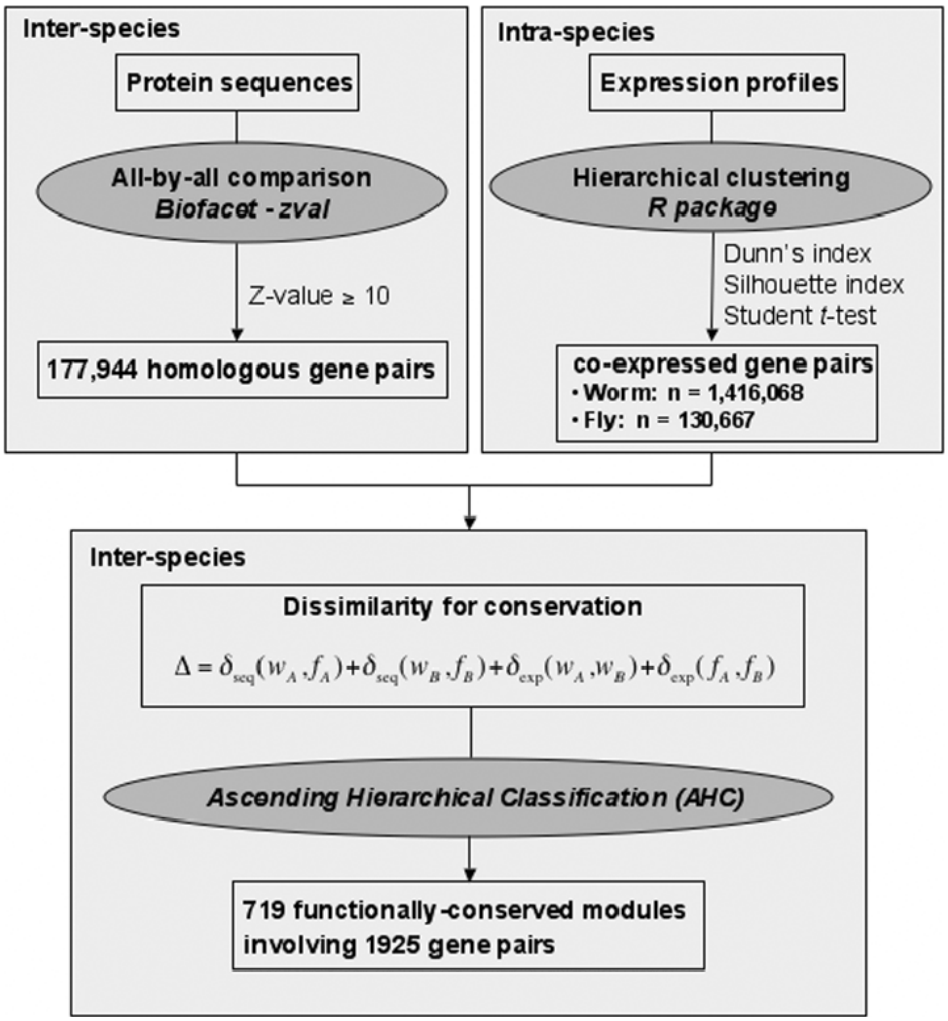

Fig. 2. Flowchart of the BBCP comparative method. Homologous relationships were computed through all-by-all comparison of *C.elegans* and *Drosophila* proteins using the zval algorithm of the Biofacet program (left upper panel). Gene expression clusters were computed from RNA studies using hierarchical clustering, and validated using the Dunn's and silhouette indices, and Student t-test (right upper panel). Functionally conserved modules (FCMs) were computed with no a priori through calculation of the dissimilarity for conservation Δ (lower panel).

Finally, the dissimilarity for conservation Δ was defined between two gene pairs from *C.elegans* (W) and *Drosophila* (F) as

$$\Delta = \delta_{\text{seq}}(w_A, f_A) + \delta_{\text{seq}}(w_B, f_B) + \delta_{\text{exp}}(w_A, w_B) + \delta_{\text{exp}}(f_A, f_B), \quad (6)$$

where $\delta_{\text{seq}}(w_A, f_A)$ applied to proteins encoded by genes w_A and f_A, $\delta_{\text{seq}}(w_B, f_B)$ applied to proteins encoded by genes w_B and f_B, $\delta_{\text{exp}}(w_A, w_B)$ applied to co-expressed genes w_A and w_B, and $\delta_{\text{exp}}(f_A, f_B)$ applied to co-expressed genes f_A and f_B (Fig. 1B). The matrix describing Δ values was used to compute a hierarchical classification (Johnson, 1967) based on the *Ward*'s agglomeration method. This method iteratively merges the two closest clusters, resulting in a binary tree. As the matrix describing Δ values was sparse, we used a modified ascending hierarchical classification method and obtained a set of trees, referred here to as functionally conserved modules (FCMs). Overall, the underlying idea of the BBCP method was thus to extract more functionally relevant data on genetic conservation by considering sequence similarity and gene co-expression at the same level.

Validation analysis of functionally conserved modules

To evaluate the biological significance of FCMs, we used GO annotations (Ashburner *et al*., 2000). Gene ontology data were downloaded from the GO consortium website (http://www.geneontology.org/) in March 2004. We sought to identify statistically significant link(s) between GO terms and FCM gene content. The probability of selecting the observed number of genes from a given GO category by chance was calculated using the hypergeometric distribution. This calculation takes into account the total number of genes in all FCMs (N), the number of genes in the FCM considered (n), the number of genes in a particular GO category (M) and, in the FCM considered, the number of genes defined by the GO category (m):

$$P\text{-value} = \frac{C_m^M * C_{n-m}^{N-M}}{C_n^N}. \quad (7)$$

This probability is strongly dependent on the abundance of annotations available in each of the FCMs. Thus, associations that were not statistically significant may reflect poor annotation for the FCM genes. In addition, we sought to identify statistically significant link(s) between the FCM gene content and RNAi phenotypes from large-scale RNAi screens (Ashrafi *et al*., 2003; Kamath *et al*., 2003; Simmer *et al*., 2003; Boutros *et al*., 2004; Nollen *et al*., 2004) by using the hypergeometric probability.

As a complement to using GO annotations and RNAi phenotypes, we used InterPro protein domain annotations (Mulder *et al*., 2003) downloaded from the European Bioinformatics Institute website (http://www.ebi.ac.uk/interpro/) in June 2004. We sought to identify FCMs that primarily described a single protein family or that comprised two or more different protein domains as carried by different proteins. Additionally, we used functional descriptors for microarray studies of peculiar physiological responses or signaling pathways (De Gregorio *et al*., 2002; Gaudet and Mango, 2002; Klebes *et al*., 2002; Mallo *et al*., 2002; Romagnolo *et al*., 2002; Zhang *et al*., 2002; Lee *et al*., 2003; Wang and Kim, 2003; Roxstrom-Lindquist *et al*., 2004). Finally, we analyzed the scientific literature for the FCMs found.

RESULTS

Balancing protein similarity and gene co-expression reveals new conserved genetic modules

We obtained 719 FCMs (Supplementary Figure 1) that involved 1925 gene pairs. Compared to previous studies (Stuart *et al*., 2003; Bergmann *et al*., 2004), these conserved modules mostly contained less than eight genes (Fig. 3), and their content was analyzed for GO terms. Although some GO terms may not be very informative, GO is recognized as a useful reference system for the biological classification of large datasets (Stuart *et al*., 2003; McCarroll *et al*., 2004).

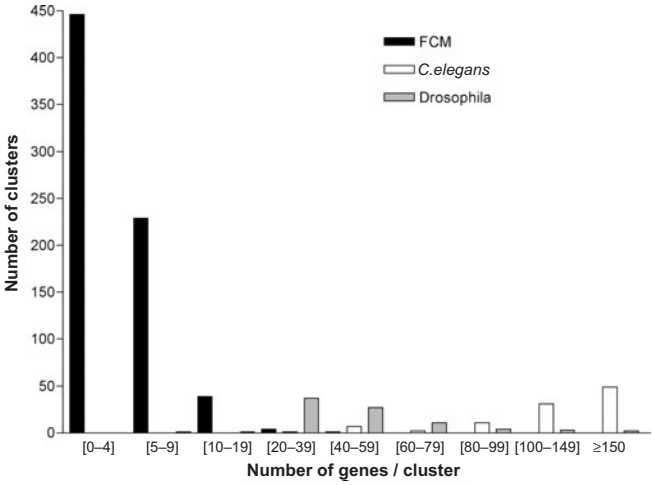

Fig. 3. Distribution of the cluster sizes. This shows the size of *C.elegans* and *Drosophila* gene expression clusters, compared with that of FCMs resulting from the BBCP analysis.

The examination of FCM content for GO terms indicated that 1578 (82%) *C.elegans* and 1662 (83%) *Drosophila* genes were defined by one or more GO annotations. Statistically significant enrichment in GO terms was observed for 518 (72%) FCMs, as indicated by the association to one or more GO annotations at a significance level of 10^{-2} (Supplementary Table 1). This included 65% of the FCMs showing a 'biological process' annotation, 84% showing a 'molecular function' annotation and 28% showing a 'cellular component' annotation (Fig. 4), which indicated a significant coverage of FCMs by GO terms. In each of these ontologies, general-to-precise GO terms were found, suggesting instructive coverage of the FCMs by these GO terms. We observed that FCMs may differ from the modules previously reported to be conserved in *C.elegans* and *Drosophila* (Stuart *et al*., 2003; Bergmann *et al*., 2004) by one or more gene(s), or by one or more pair(s) of homologous and/or co-expressed genes. This observation led to the assumption that FCMs containing at least 25% of previously undescribed gene associations may be considered as providing significantly new information. This feature was present in 94% of the conserved modules, 150 FCMs being completely new (Supplementary Figure 1).

Conserved genetic modules for cell type-specific signaling

Using GO terms, we scored 46 FCMs that were almost or fully described in previous studies (Supplementary Table 2) and that corresponded primarily to housekeeping biological processes. For example, we detected modules associated with the proteasome (FCM 102), respiratory chain complex (FCM 193), ribosome (FCM 320), RNA polymerase II complex (FCM 330) and cell cycle (FCM 650). This provided evidence that our approach was able to replicate previous findings on modules that relate to biological processes known to be highly conserved (Teichmann and Babu, 2002; Stuart *et al*., 2003; van Noort *et al*., 2003; Bergmann *et al*., 2004).

Next, we analyzed GO terms for the 50 FCMs that best-ranked for 'biological process' and 'molecular function' (Supplementary Table 3). While these FCMs appeared to be primarily enriched in housekeeping biological mechanisms, we noticed they may correspond to differentiated functions such as e.g. 'neuropeptide

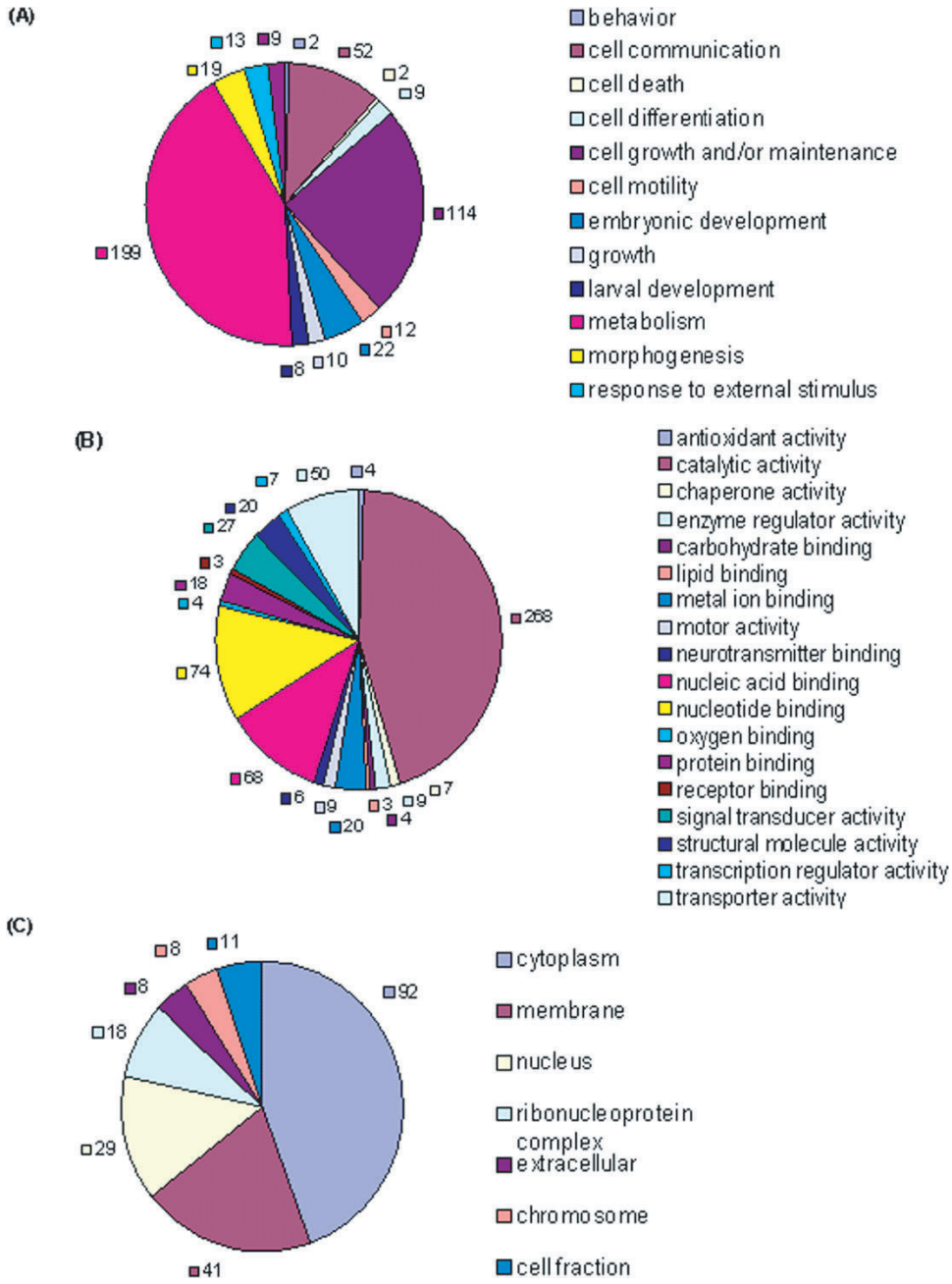

Fig. 4. Distribution of GO terms in FCMs. Shown here are the numbers of FCMs significantly described ($P < 0.01$) by the GO categories 'biological process' **(A)**, 'molecular function' **(B)** and 'cellular component' **(C)**. In this figure, the GO terms shown are the most general terms, corresponding to levels 1–3 of the GO hierarchy (Ashburner *et al*., 2000). The most precise GO term(s) relating to each of the FCMs are indicated in Supplementary Table 1.

receptor activity' (FCM 210), 'synaptic transmission' (FCM 419) or 'gametogenesis' (FCM 622). A more precise analysis of FCM enrichment in GO terms led to the identification of 57 FCMs enriched in differentiated functions (Supplementary Table 4). Our analysis of the scientific literature led to the identification of 75 additional FCMs of this type (Supplementary Table 4). Thus, in addition to detecting FCMs for housekeeping function, the BBCP procedure highlighted a significant proportion (132/719 FCMs: 18%) of conserved modules that described cell type-specific processes, most of these modules (125 FCMs) containing significantly new information. Compared to GO terms describing these modules, we identified a more precise functional and/or developmental annotation for 90/132 FCMs (Supplementary Table 4). Finally, 283 predicted genes were found to be associated with cell type-specific events, illustrating the predictive value of the BBCP method.

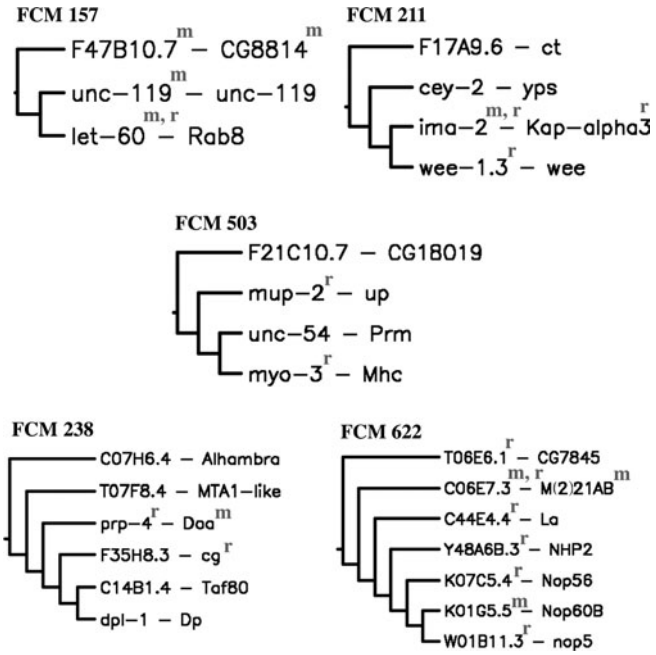

Fig. 5. Examples of FCMs with high biological content. High biological content is assessed based on analysis of the functional attributes and scientific literature. The FCM 157 contained genes involved in Ras neuronal pathway. FCM 211 contained genes involved in gametogenesis. FCM 238 contained transcription factors or regulators active in germ or neuronal cells. FCM 503 was composed of muscle cell components. FCM 622 was enriched for genes involved in RNA processing and embryogenesis or gametogenesis. m, microarray data; r, RNAi phenotype.

Conserved genetic modules for precise cell fate specification and physiological activities

The comparison with genome-wide RNAi knockouts in *C.elegans* (Ashrafi *et al*., 2003; Kamath *et al*., 2003; Simmer *et al*., 2003; Nollen *et al*., 2004) and *Drosophila* (Boutros *et al*., 2004) indicated that 350 FCMs contained at least one gene for which a RNAi phenotype was available. Statistically significant enrichment in RNAi phenotypes (Supplementary Table 5) was observed in 7.2% of the FCMs (52 modules). This small percentage was expected as RNAi phenotypes are not available for a large proportion of *C.elegans* or *Drosophila* genes. The observation of RNAi-enriched FCMs supported the notion defined above that the BBCP method identifies functionally conserved gene pair modules. Modules describing cell type-specific signaling (Supplementary Table 4) were indeed observed among the 52 FCMs, e.g. FCM 354, a module that relates to vulva development in *C.elegans* and formation of imaginal disks in *Drosophila*. The occurrence of vulva-related RNAi phenotypes for all worm genes in this FCM notably indicated that the worm *D2021.1* predicted gene may participate in the regulation of transcription together with the Zn-finger domain protein flectin *flt-1* and deacetylase *dcp-66*. Furthermore, 29 FCMs (Supplementary Table 6) contained at least one RNAi phenotype associated with lethality in *C.elegans* ('embryonic lethal' phenotype) and *Drosophila* ('lethality' phenotype), suggesting common survival functions in this FCM group and several new players (predicted genes) to participate in this conserved function.

Using InterPro annotations (Mulder *et al*., 2003), we detected 29 FCMs strongly enriched for a single protein domain, suggesting the consistent detection of a single family or superfamily of genes by the BBCP method (Supplementary Table 7). For instance, (1) FCM 645 contained three *Drosophila* cytochrome P450 genes classically involved in detoxification, suggesting that the three *C.elegans* predicted proteins may achieve a similar function, (2) FCM 712 corresponded to laminins involved in morphogenesis and (3) FCM 717 contained two *C.elegans* helicases active in the germline that were associated with two *Drosophila* predicted proteins known to contain helicase domains (Supplementary Figure 1). The use of InterPro annotations also indicated how protein domains were distributed in the FCMs, providing a criterion to detect FCMs that may consistently bring together different 'molecular function' categories in terms of signal integration during development. For instance, at the intersect of FCMs showing a 'spectrin' domain (8 FCMs) and those showing a 'myosin' domain (7 FCMs) were FCM 503 (see the next subsection, and Fig. 5) and FCM 32, a module that contains two pairs of homologous genes associated with epithelial cell morphogenesis. Another example related to protein domains found in transcription factors. We observed a 'ligand-binding domain of nuclear hormone receptor' in nine FCMs, a Zn-finger domain in 85 FCMs, and a 'homeobox' domain in 24 FCMs. The latter group comprised FCM 631 that brought together the *C.elegans pal-1–Drosophila Antennapedia* pair of homeobox genes and the *C.elegans mom-2–Drosophila wingless* pair of 'Wnt superfamily' genes, thus describing the segregation of cell fate specification activities along the anteroposterior axis, which may be linked to intestinal cell specification in *C.elegans*. Also comprised in this FCM group were the *C.elegans* homeobox genes *ceh-8* (FCM 359), *ceh-17* (FCM 471), *ceh-22* (FCM 423), *ceh-24* and *ceh-26* (FCM 616), *ceh-27* (FCM 467), *ceh-33* (FCM 421) and *ceh-39* (FCM 258). Notable of these was FCM 471 that associated the *C.elegans* paired-like homeobox gene *ceh-17*, a likely ortholog of mammalian *Phox2a/b* expressed in five head neurons (Pujol *et al*., 2000), with the neuronal

marker *ser-2*, the *C.elegans* tyramine receptor gene (Tsalik *et al*., 2003). The *Drosophila* genes associated in this FCM were the *eyeless* homeobox gene and *Oamb*, the octopamine receptor gene (Han *et al*., 1998). Tyramine is the chemical precursor of octopamine which is believed to be the invertebrate counterpart of norepinephrine, all being neuroactive substances (Tsalik *et al*., 2003). This FCM together with overlap of *ceh-17* and *ser-2* expression in SIAV worm neurons (Pujol *et al*., 2000; Tsalik *et al*., 2003) suggested a previously undetected link between *ceh-17* and the specification of neurons that express biogenic amine receptors. Additionally, this FCM suggested an overlap of *eyeless* and *Oamb* developmental function, consistent with the notion that these two genes regulate the development of the *Drosophila* mushroom body (Han *et al*., 1998; Noveen *et al*., 2000).

To evaluate further the presence of an instructive biological content in the FCMs, we sought to examine the correlation between FCM genes and microarray studies of specific physiological (De Gregorio *et al*., 2002; Mallo *et al*., 2002; Zhang *et al*., 2002; Lee *et al*., 2003; Roxstrom-Lindquist *et al*., 2004) or developmental (Gaudet and Mango, 2002; Klebes *et al*., 2002; Romagnolo *et al*., 2002; Wang and Kim, 2003) processes in *C.elegans* or *Drosophila*. This evaluation highlighted 363 FCMs containing at least one gene characterized in one of these studies. In this FCM subgroup were highlighted conserved genesets that may underlie different physiological responses in *C.elegans* and *Drosophila*. For example, we detected 15 FCMs containing both dauer gene(s) in *C.elegans* (Wang and Kim, 2003) and immune response gene(s) in *Drosophila* (Roxstrom-Lindquist *et al*., 2004). Also highlighted were genesets that may underlie similar physiological responses in *C.elegans* and *Drosophila*. For example, we detected four FCMs (FCMs 141, 356, 408 and 564) containing immune response genes in *C.elegans* (Mallo *et al*., 2002) and *Drosophila* (De Gregorio *et al*., 2002; Roxstrom-Lindquist *et al*., 2004).

Altogether these observations identified BBCP products as informative and functionally conserved gene modules. Furthermore, our data indicated that precise functional categories such as those involved in cell fate mechanisms and physiological responses may be detected by the BBCP procedure. These observations were pertinent to FCMs that did not appear to fall into conserved functional categories previously detected by means of computational analysis (Stuart *et al*., 2003; Bergmann *et al*., 2004).

Predictive value of BBCP products

We observed that previously undescribed biological role(s) were suggested for FCM genes. This applied noticeably to predicted genes in FCMs describing cell type-specific activities (Supplementary Table 4). This also applied to a group of 215 predicted genes in FCMs that contained a significant amount of new gene associations and were strongly associated ($P < 10^{-3}$) with a GO term 'Biological Process' (Supplementary Table 8). Detailed examples of the predictive value of BBCP products are given below. FCM 157 may define Ras activation in neural lineage (Supplementary Table 4, Fig. 5). This FCM contained the *C.elegans* gene *let-60* and two of its transcriptional targets *unc-119* and *F47B10.7* (Romagnolo *et al*., 2002), *let-60* being known to play a role in the differentiation of several *C.elegans* tissues, among which is the neural lineage. Furthermore, *unc-119* is required for nervous system maintenance in *C.elegans* and *Drosophila* (Knobel *et al*., 2001), and predicted *F47B10.7* is known to be expressed in touch receptor neurons (Zhang *et al*., 2002). FCM 211 (Supplementary Table 4, Fig. 5) suggested that the predicted worm gene *F17A9.6* may be involved in gametogenesis. This FCM comprised 7/8 genes known to participate in this process, e.g. *wee* required for oocyte development and maturation (Lamitina and L'Hernault, 2002), and somatically expressed *cut*, required for germ cell integrity (Jackson and Blochlinger, 1997). FCM 238 (Supplementary Table 4, Fig. 5) suggested that the four *C.elegans* predicted genes *C07H6.4*, *T07F8.4*, *F35H8.3* and *C14B1.4* may be transcription factors or regulators active in germ or neuronal cells as all *Drosophila* genes in this cluster [*Alhambra*, *MTA1-like*, *Doa* (Darkener of apricot), *cg* (combgap), *Taf80* (TBP-associated factor 80 kDa), *Dp*, (DP transcription factor)] were known to be involved in these processes, noticeably *Dp* (Cayirlioglu *et al*., 2003). FCM 503 (Supplementary Table 4, Fig. 5) suggested that predicted *C.elegans* F21C10.7 and *Drosophila* CG18019 genes may be associated with muscle organization as they were aggregated with three gene pairs encoding myosin heavy chain, paramyosin and tropomyosin (Honda and Epstein, 1990). Finally, FCM 622 (Supplementary Table 4, Fig. 5) suggested a function for seven *C.elegans* and one *Drosophila* predicted genes, here in RNA processing and embryogenesis or gametogenesis. These molecular function and/or biological processes are indeed known to involve 6/7 *Drosophila* genes part of this FCM, e.g. the *Nop* gene family (Vorbruggen *et al*., 2000).

DISCUSSION

The determination of conserved genesets that may underlie essential developmental and physiological programs using the computational analysis of gene expression is an emerging aspect in comparative biology and, more widely, evolutionary developmental biology (Rast, 2003). One approach in this field aims at studying the conservation of *cis*-regulatory control circuits in development by modeling 'gene regulatory networks' (GRNs) (Bolouri and Davidson, 2002). This approach has noticeably led to the identification of GRNs architecture across 500 million years of echinoderm evolution (Hinman *et al*., 2003). Another approach relies more specifically on inferring functional relationships between conserved genes by exploiting genome-wide data, noticeably microarray data. Pioneer work using microarray data has identified several genes that may act through conserved mechanisms among distantly related organisms (Teichmann and Babu, 2002; Stuart *et al*., 2003; van Noort *et al*., 2003; Bergmann *et al*., 2004; McCarroll *et al*., 2004). From these studies, there appears to be a tendency for genes in a functional class to be co-expressed, especially those in ancient, permanent or stable complexes (Teichmann and Babu, 2002; Stuart *et al*., 2003; van Noort *et al*., 2003; Bergmann *et al*., 2004), a trend also delineated by studies of co-orthologs as based on genome sequence analysis (Koonin *et al*., 2004). By balancing protein similarities and gene co-expression, we obtained modules conserved in *C.elegans* and *Drosophila* that essentially differ from previously described clusters (Stuart *et al*., 2003; Bergmann *et al*., 2004) in their size and gene content. These differences are likely to reflect the best ability of the BBCP method to identify genes involved in common functions compared to previous approaches. We indeed generated a single subspace of homologous proteins above a given threshold, which is more permissive compared to the best reciprocal hit procedure, and a series of gene expression clusters, here derived from stringent clustering. Balancing protein similarities and gene co-expression with no a priori when calculating genetic conservation (Fig. 2) generated the shortest 'functional distance' between genes. Thus, gene association based

on BBCP differed from considering either one of the parameters alone—distance for sequence similarity or co-expression. As a result, we obtained conserved modules that mostly composed of two to four gene pairs, and observed conservation of gene co-expression for proteins that are associated in stable complexes as well as those that do not form such complexes. A significant proportion of conserved modules were previously undetected.

The source data used in our study may influence BBCP informativity. Since all genes were not represented in the microarrays either for *C.elegans* (Kim *et al*., 2001) or for *Drosophila* (Arbeitman *et al*., 2002), FCM construction may not be optimal for some gene families. Nonetheless, this was unlikely to be a significant problem as nearly all genes are represented in the *C.elegans* microarrays (Kim *et al*., 2001). Another factor that may influence FCM gene content is the relatively limited number and diversity of the experimental conditions currently probed with microarrays. Finally, the use of whole animal RNA may reduce the sensitivity for detection of biological processes that involve a small number of cells (Zhang *et al*., 2002), and may generate associations between gene pairs that are co-regulated but expressed in different cell types or tissues. However, microarray experiments based on cell type-specific RNA remain scarce. Despite these limitations and potential biases, the comparison of FCM gene content with the biological data available for *C.elegans* and *Drosophila* genes indicated that the BBCP method consistently detected new elements of genetic conservation. The fact that two or more homologous gene pairs, not necessarily selected as best orthologs, are co-expressed in *C.elegans* and *Drosophila* is a strong indication for the putative participation of these genes in essential developmental programs in these two organisms, whether or not they may ultimately result in a similar physiological function. A large proportion (83.4%) of FCMs contained both predicted and known genes, which, together with moderate FCM sizes, provided an enhanced basis for new hypotheses to be raised on the molecular function and/or biological role(s) of several predicted genes. It is to be noted that, several FCMs fall into 'precise' molecular, cellular or physiological categories, previously not detected by means of computational analysis. Some modules related, for example, to cell fate specification and physiological activities in *C.elegans* and *Drosophila*, which illustrated the potential power of BBCP at discriminating specific gene aggregation events. These features suggested that the BBCP method is able to describe genetic conservation relative to precise aspects of developmental diversity, an essential question in computational biology. It should be noted that (1) the reliability of BBCP products is not dependent on the size of the resulting modules, small FCMs (two to three gene pairs) being as reliable and informative as larger FCMs, (2) genes in FCMs, and most widely in conserved modules (Stuart *et al*., 2003; Bergmann *et al*., 2004), may not necessarily be in the same pathway or belong to the same cellular mechanism, as the expression data currently available do not support this level of precision and (3) as genome-wide functional attributes are gaining in diversity and specificity, BBCP products should be considered as evolutive computational images rather than final figures of genetic conservation. Finally, while the conserved module analysis used herein does not allow convergence to be analyzed directly, it can result in FCMs that contain genes showing weak sequence homology and strong co-expression, which may be pertinent to convergence in pathway connectivity and activity (Dowell *et al*., 2003). Regarding convergence in physiological function, this can be inferred a posteriori only, by analyzing the FCM gene content.

Our study is based on proteins from *C.elegans* and *Drosophila* whose sequences were available as of October 2002, and relies on one set of microarray data comprising a large number of different experimental conditions for each of the organisms studied (Kim *et al*., 2001; Arbeitman *et al*., 2002). More recent data were not used or included due to the time-consuming and labor-intensive nature of the BBCP procedure and its validation. The use of these and newly analyzed biological processes will enable more comprehensive studies into genesets functionally conserved in *C.elegans* and *Drosophila*.

In summary, using a new comparative procedure—BBCP, we identified new conserved genetic modules that may underlie *C.elegans* and *Drosophila* development. We detected a significant proportion of modules that highlighted new relationships between genetic conservation and developmental diversity in these two organisms. Our case study shows the potential of BBCP for analyzing large and diverse datasets in order to detect conserved genetic modules that may trigger specific developmental events in distantly related organisms. This approach together with other frameworks for the comparative analysis of genomes (Koonin *et al*., 2004) and postgenomic data (Ashburner *et al*., 2000; Davidson *et al*., 2002; Alter *et al*., 2003; Shannon *et al*., 2003; Stuart *et al*., 2003; Kyoda *et al*., 2004) may allow evolutionary conservation to be studied on the basis of accurate *in silico* prediction.

ACKNOWLEDGEMENTS

We thank Hamid Bolouri and John Aitchison at the Institute for Systems Biology (Seattle, WA) for critical reading, Gilles Didier at the Genome and Informatics Laboratory, UMR 8116 CNRS/UEVE, for stimulating discussion and Infobiogen (Evry, France) for access to high-speed computing facilities. C.L. is supported by a doctoral fellowship at Gene-IT S.A. under the auspices of the Agence Nationale de la Recherche et de la Technologie (ANRT), Paris, France. This work was supported by the Centre d'Etude du Polymorphisme Humain (CEPH) and the Institut National de la Recherche et de la Santé Médicale (INSERM), Paris, France.

REFERENCES

Aboobaker,A. and Blaxter,M. (2003) Hox gene evolution in nematodes: novelty conserved. *Curr. Opin. Genet. Dev.*, **13**, 593–598.

Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Alter,O., Brown,P.O. and Botstein,D. (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl Acad. Sci. USA*, **100**, 3351–3356.

Arbeitman,M.N., Furlong,E.E., Imam,F., Johnson,E., Null,B.H., Baker,B.S., Krasnow,M.A., Scott,M.P., Davis,R.W. and White,K.P. (2002) Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, **297**, 2270–2275.

Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Ashrafi,K., Chang,F.Y., Watts,J.L., Fraser,A.G., Kamath,R.S., Ahringer,J. and Ruvkun,G. (2003) Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature*, **421**, 268–272.

Aude,J.C. and Louis,A. (2002) An incremental algorithm for *Z*-value computations. *Comput. Chem.*, **26**, 403–411.

Bastien,O., Aude,J.C., Roy,S. and Marechal,E. (2004) Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of *Z*-value statistics. *Bioinformatics*, **20**, 534–537.

Bergmann,S., Ihmels,J. and Barkai,N. (2004) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.*, **2**, E9.

Bolouri,H. and Davidson,E.H. (2002) Modeling transcriptional regulatory networks. *Bioessays*, **24**, 1118–1129.

Bolshakova,N. and Azuaje,F. (2003) Machaon CVE: cluster validation for gene expression data. *Bioinformatics*, **19**, 2494–2495.

Boutros,M., Kiger,A.A., Armknecht,S., Kerr,K., Hild,M., Koch,B., Haas,S.A., Consortium,H.F., Paro,R. and Perrimon,N. (2004) Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science*, **303**, 832–835.

Cayirlioglu,P., Ward,W.O., Silver Key,S.C. and Duronio,R.J. (2003) Transcriptional repressor functions of *Drosophila* E2F1 and E2F2 cooperate to inhibit genomic DNA synthesis in ovarian follicle cells. *Mol. Cell. Biol.*, **23**, 2123–2134.

Comet,J.P., Aude,J.C., Glemet,E., Risler,J.L., Henaut,A., Slonimski,P.P. and Codani,J.J. (1999) Significance of *Z*-value statistics of Smith–Waterman scores for protein alignments. *Comput. Chem.*, **23**, 317–331.

Davidson,E.H., Rast,J.P., Oliveri,P., Ransick,A., Calestani,C., Yuh,C.H., Minokawa,T., Amore,G., Hinman,V., Arenas-Mena,C. *et al*. (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.

De Gregorio,E., Spellman,P.T., Tzou,P., Rubin,G.M. and Lemaitre,B. (2002) The Toll and Imd pathways are the major regulators of the immune response in *Drosophila*. *EMBO J.*, **21**, 2568–2579.

Dowell,P., Otto,T.C., Adi,S. and Lane,M.D. (2003) Convergence of peroxisome proliferator-activated receptor gamma and Foxo1 signaling pathways. *J. Biol. Chem.*, **278**, 45485–45491.

Flybase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.

Forst,C.V. and Schulten,K. (1999) Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information. *J. Comput. Biol.*, **6**, 343–360.

Forst,C.V. and Schulten,K. (2001) Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.*, **52**, 471–489.

Gaudet,J. and Mango,S.E. (2002) Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. *Science*, **295**, 821–825.

Glemet,E. and Codani,J.J. (1997) LASSAP, a LArge Scale Sequence compArison Package. *Comput. Appl. Biosci.*, **13**, 137–143.

Han,K.A., Millar,N.S. and Davis,R.L. (1998) A novel octopamine receptor with preferential expression in *Drosophila* mushroom bodies. *J. Neurosci.*, **18**, 3650–3658.

Hanisch,D., Zien,A., Zimmer,R. and Lengauer,T. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**(Suppl. 1), S145–S154.

Hinman,V.F., Nguyen,A.T., Cameron,R.A. and Davidson,E.H. (2003) Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proc. Natl Acad. Sci. USA*, **100**, 13356–13361.

Honda,S. and Epstein,H.F. (1990) Modulation of muscle gene expression in *Caenorhabditis elegans*: differential levels of transcripts, mRNAs, and polypeptides for thick filament proteins during nematode development. *Proc. Natl Acad. Sci. USA*, **87**, 876–880.

Jackson,S.M. and Blochlinger,K. (1997) Cut interacts with Notch and protein kinase A to regulate egg chamber formation and to maintain germline cyst integrity during *Drosophila* oogenesis. *Development*, **124**, 3663–3672.

Johnson,S.C. (1967) Hierarchical clustering schemes. *Psychometrika*, **2**, 241–254.

Kamath,R.S., Fraser,A.G., Dong,Y., Poulin,G., Durbin,R., Gotta,M., Kanapin,A., Le Bot,N., Moreno,S., Sohrmann,M. *et al*. (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**, 231–237.

Kelley,B.P., Yuan,B., Lewitter,F., Sharan,R., Stockwell,B.R. and Ideker,T. (2004) PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res.*, **32**, W83–W88.

Kim,S.K., Lund,J., Kiraly,M., Duke,K., Jiang,M., Stuart,J.M., Eizinger,A., Wylie,B.N. and Davidson,G.S. (2001) A gene expression map for *Caenorhabditis elegans*. *Science*, **293**, 2087–2092.

Klebes,A., Biehs,B., Cifuentes,F. and Kornberg,T.B. (2002) Expression profiling of *Drosophila* imaginal discs. *Genome Biol.*, **3**, RESEARCH0038.

Knobel,K.M., Davis,W.S., Jorgensen,E.M. and Bastiani,M.J. (2001) UNC-119 suppresses axon branching in *C.elegans*. *Development*, **128**, 4079–4092.

Koonin,E.V., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Krylov,D.M., Makarova,K.S., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N., Rao,B.S. *et al*. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.*, **5**, R7.

Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.*, **52**, 540–542.

Kriventseva,E.V., Fleischmann,W., Zdobnov,E.M. and Apweiler,R. (2001) CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res.*, **29**, 33–36.

Kyoda,K., Baba,K., Onami,S. and Kitano,H. (2004) DBRF-MEGN method: an algorithm for deducing minimum equivalent gene networks from large-scale gene expression profiles of gene deletion mutants. *Bioinformatics*, **20**, 2662–2675.

Lamitina,S.T. and L'Hernault,S.W. (2002) Dominant mutations in the *Caenorhabditis elegans* Myt1 ortholog wee-1.3 reveal a novel domain that controls M-phase entry during spermatogenesis. *Development*, **129**, 5009–5018.

Lee,C.Y., Clough,E.A., Yellon,P., Teslovich,T.M., Stephan,D.A. and Baehrecke,E.H. (2003) Genome-wide analyses of steroid- and radiation-triggered programmed cell death in *Drosophila*. *Curr. Biol.*, **13**, 350–357.

Mallo,G.V., Kurz,C.L., Couillault,C., Pujol,N., Granjeaud,S., Kohara,Y. and Ewbank,J.J. (2002) Inducible antibacterial defense system in *C.elegans*. *Curr. Biol.*, **12**, 1209–1214.

McCarroll,S.A., Murphy,C.T., Zou,S., Pletcher,S.D., Chin,C.S., Jan,Y.N., Kenyon,C., Bargmann,C.I. and Li,H. (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.*, **36**, 197–204.

Miyabayashi,T., Palfreyman,M.T., Sluder,A.E., Slack,F. and Sengupta,P. (1999) Expression and function of members of a divergent nuclear receptor family in *Caenorhabditis elegans*. *Dev. Biol.*, **215**, 314–331.

Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.

Nollen,E.A., Garcia,S.M., van Haaften,G., Kim,S., Chavez,A., Morimoto,R.I. and Plasterk,R.H. (2004) Genome-wide RNA interference screen identifies previously undescribed regulators of polyglutamine aggregation. *Proc. Natl Acad. Sci. USA*, **101**, 6403–6408.

Noveen,A., Daniel,A. and Hartenstein,V. (2000) Early development of the *Drosophila* mushroom body: the roles of eyeless and dachshund. *Development*, **127**, 3475–3488.

Pujol,N., Torregrossa,P., Ewbank,J.J. and Brunet,J.F. (2000) The homeodomain protein CePHOX2/CEH-17 controls antero-posterior axonal growth in *C.elegans*. *Development*, **127**, 3361–3371.

Rast,J.P. (2003) Development gene networks and evolution. *J. Struct. Funct. Genomics*, **3**, 225–234.

Romagnolo,B., Jiang,M., Kiraly,M., Breton,C., Begley,R., Wang,J., Lund,J. and Kim,S.K. (2002) Downstream targets of let-60 Ras in *Caenorhabditis elegans*. *Dev. Biol.*, **247**, 127–136.

Roxstrom-Lindquist,K., Terenius,O. and Faye,I. (2004) Parasite-specific immune response in adult *Drosophila melanogaster*: a genomic study. *EMBO Rep.*, **5**, 207–212.

Rudel,D. and Sommer,R.J. (2003) The evolution of developmental mechanisms. *Dev. Biol.*, **264**, 15–37.

Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Simmer,F., Moorman,C., Van Der Linden,A.M., Kuijk,E., Van Den Berghe,P.V., Kamath,R., Fraser,A.G., Ahringer,J. and Plasterk,R.H. (2003) Genome-wide RNAi of *C.elegans* using the hypersensitive rrf-3 strain reveals novel gene functions. *PLoS Biol.*, **1**, E12.

Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.

Stuart,J.M., Segal,E., Koller,D. and Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Teichmann,S.A. and Babu,M.M. (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol.*, **20**, 407–410, discussion 410.

The *C.elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C.elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.

Tsalik,E.L., Niacaris,T., Wenick,A.S., Pau,K., Avery,L. and Hobert,O. (2003) LIM homeobox gene-dependent expression of biogenic amine receptors in restricted regions of the *C.elegans* nervous system. *Dev. Biol.*, **263**, 81–102.

van Noort,V., Snel,B. and Huynen,M.A. (2003) Predicting gene function by conserved co-expression. *Trends Genet.*, **19**, 238–242.

Vorbruggen,G., Onel,S. and Jackle,H. (2000) Restricted expression and subnuclear localization of the *Drosophila* gene Dnop5, a member of the Nop/Sik family of the conserved rRNA processing factors. *Mech. Dev.*, **90**, 305–308.

Wang,J. and Kim,S.K. (2003) Global analysis of dauer gene expression in *Caenorhabditis elegans*. *Development*, **130**, 1621–1634.

Ward,J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.

Zhang,Y., Ma,C., Delohery,T., Nasipak,B., Foat,B.C., Bounoutas,A., Bussemaker,H.J., Kim,S.K. and Chalfie,M. (2002) Identification of genes expressed in *C.elegans* touch receptor neurons. *Nature*, **418**, 331–335.