36 Boutell, J.M. *et al.* (1999) Aberrant interactions of transcriptional repressor proteins with the Huntington's disease gene product, huntingtin. *Hum. Mol. Genet.* 8, 1647–1655
37 Takano, H. and Gusella, J.F. (2002) The predominantly HEAT-like motif structure of huntingtin and its association and coincident nuclear entry with dorsal, an NF-κB/Rel/dorsal family transcription factor. *BMC Neurosci.* 3, 15–27
38 Holbert, S. *et al.* (2001) The Gln–Ala repeat transcriptional activitor CA150 interacts with huntingtin: neuropathologic and genetic eveidence for a role in Huntington's disease pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.* 98, 1811–1816
39 Kim, S. *et al.* (2002) Polyglutamine protein aggregates are dynamic. *Nat. Cell Biol.* 4, 826–831
40 Carmichael, J. *et al.* (2002) Glycogen synthase kinase-3b inhibitors prevent cellular polyglutamine toxicity caused by the Huntington's disease mutation. *J. Biol. Chem.* 277, 33791–33798

0168-9525/03/$ - see front matter
doi:10.1016/S0168-9525(03)00074-X

Genome Analysis

Predicting gene function by conserved co-expression

Vera van Noort, Berend Snel and Martijn A. Huynen

Nijmegen Center for Molecular Life Sciences, Center for Molecular and Biomolecular Informatics, PO Box 9010, 6500 GL Nijmegen, The Netherlands

We show that gene co-expression, which generally provides only a very weak signal for the prediction of functional interactions, can provide a reliable signal by exploiting evolutionary conservation. The encoded proteins of conserved co-expressed gene pairs are highly likely to be part of the same pathway not only after speciation (98%), but also after parallel gene duplication (97%). Conserved co-expression combined with homology data enables us to predict specific gene functions. The use of conservation between parallel duplicated gene pairs to predict function is especially promising given that gene duplication is common in eukaryotes, and that data from only a single organism can be used.

One of the major goals of the post-genomic era is the elucidation of gene function. Correlations between expression patterns [1] from hundreds of experiments for both *Saccharomyces cerevisiae* [2] and *Caenorhabditis elegans* [3] can predict only general functional interactions [4,5]. As the evolutionary conservation of weak signals (like gene order), has been used successfully to predict gene function [6,7], here we examine whether the conservation of co-expression can be used to improve function prediction. We use conservation between pairs of orthologs in two species, as well as conservation of co-expression between parallel duplicated gene pairs in one species to predict functional interactions. We combine these predicted interactions with homology data to predict specific functions for uncharacterized genes.

Co-expression provides a weak signal for pathway prediction

Two large-scale expression datasets were obtained, one from *S. cerevisiae* [2] and one from *C. elegans* [3]. Uncentered correlation [1] was calculated between the expression profiles of all *S. cerevisiae* genes and between the expression profiles of all *C. elegans* genes. The higher the correlation (R) between two genes, the more probable it is that they act in the same pathway (Fig. 1). However, at a significant correlation threshold of 0.6 ($P < 0.005$, Table 1),

Table 1. Significant levels of co-expression conservation after gene duplication or speciation

	Total pairs[a]	Number of pairs > 0.6[b]	Observed fraction > 0.6[c]	Expected fraction > 0.6[d]	Observed/expected
Gene pairs with an orthologous gene-pair > 0.6					
C. elegans	18161	803	**0.0442***	0.00379	12
S. cerevisiae	36548	1215	**0.0332***	0.00216	15
Gene pairs with a paralogous gene-pair > 0.6					
C. elegans	207214	29031	**0.1401***	0.00379	37
S. cerevisiae	38253	2167	**0.0566***	0.00216	26
Gene pairs with a diverged paralogous gene-pair > 0.6					
C. elegans	125852	1299	**0.0103***	0.00379	3
S. cerevisiae	26941	174	**0.0065***	0.00216	3

[a]The number of gene pairs, regardless of their co-expression, with a co-expressed, orthologous gene pair in the other species or a co-expressed paralogous gene pair in the same species.
[b]The number of co-expressed gene pairs with a co-expressed, orthologous gene pair in the other species or a co-expressed paralogous gene pair in the same species.
[c]Observed fraction of conserved co-expressed pairs. Asterisk shows $P < 0.001$, determined by 1000 Monte Carlo simulations; that is, such high levels of conservation were not observed when randomly distributing the correlations over the gene pairs 1000 times.
[d]Expected fraction assuming no conservation of co-expression, determined by the total fraction of co-expressed gene pairs from the total number of gene pairs in that species.

Corresponding author: Martijn A. Huynen (M.Huynen@cmbi.kun.nl).

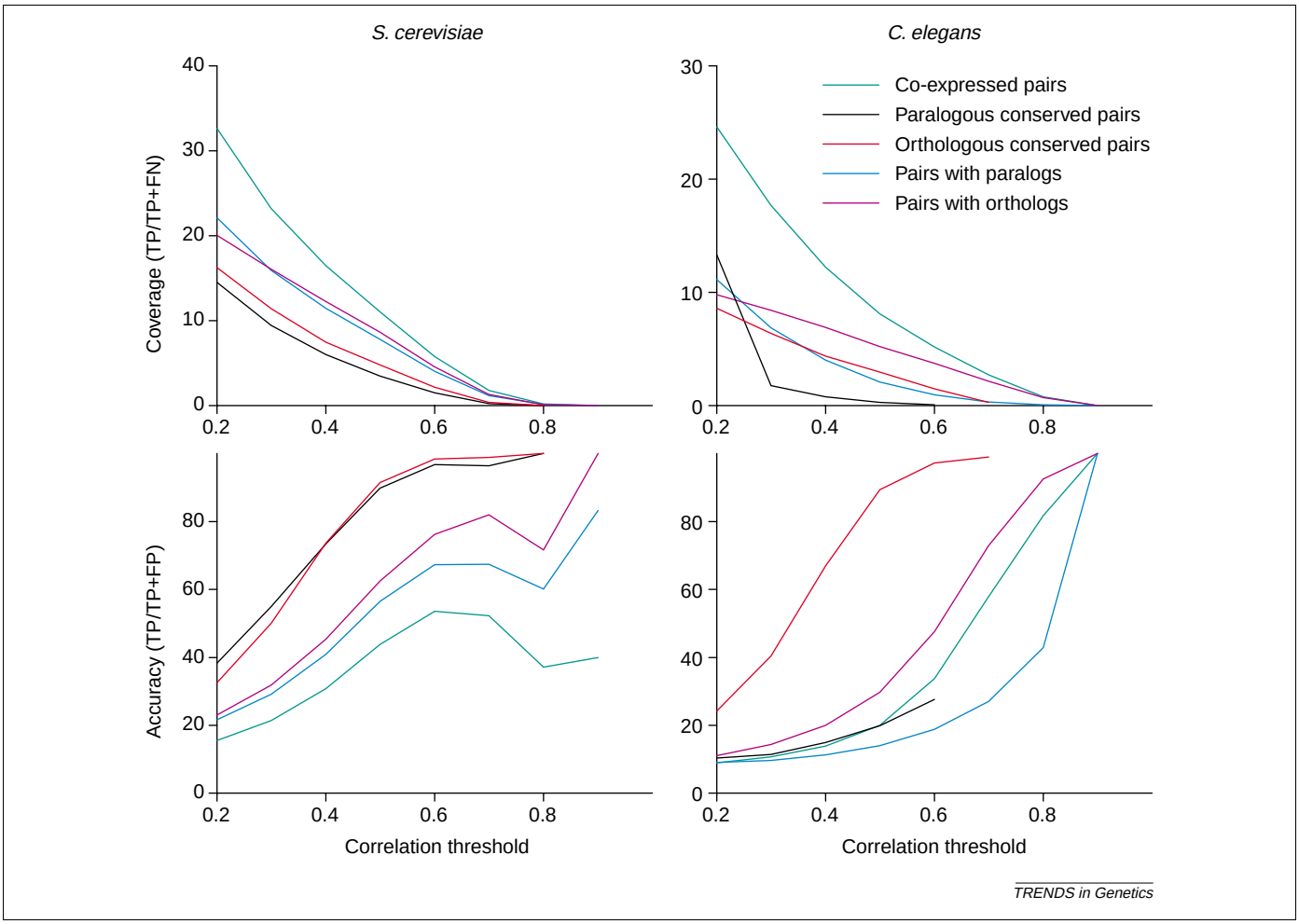

Fig. 1. Accuracy and coverage of functional interaction prediction. Accuracy (bottom) and coverage (top) at varying correlation thresholds for detection of co-expression. The accuracy is obtained by the number of predicted pairs that are on the same map in the PATHWAY database of KEGG (release 23) [23] (true positives) divided by the total number of predicted pairs (true positives plus false positives). The coverage is obtained by dividing the true positives by the total number of gene pairs that can be found on the same map in the PATHWAY database (true positives plus false negatives). Left, co-expressed gene pairs in *Saccharomyces cerevisiae*. Right, co-expressed gene pairs in *Caenorhabditis elegans*. Green, all gene pairs with expression correlation above the threshold; blue, gene pairs with expression correlation above the threshold and a pair of paralogs in the same species; purple, gene pairs with expression correlation above the threshold and a pair of orthologs in the other species; red, co-expressed gene pairs with expression correlation above the threshold and orthologs with an expression correlation above the threshold; black, co-expressed gene pairs with expression correlation above the threshold and paralogs with an expression correlation above the threshold. The increased accuracy of conserved co-expression is partly due to the requirement that both genes to have an ortholog in the other species or a paralog in the same species: the accuracies for gene pairs with orthologs or paralogs are slightly higher than the accuracies for all co-expressed gene pairs, although they fall well below the accuracies attained for conserved co-expressed gene pairs.

the fraction of annotated proteins that are part of the same pathway is only 54% in *S. cerevisiae* and 34% in *C. elegans*.

Significant levels of evolutionary conservation of co-expression

To evaluate whether evolutionary conservation (Fig. 2) can improve upon these limits in the use of co-expression for function prediction, we first established whether there is significant conservation, potentially reflecting selection pressure on maintaining functional interactions. To determine conservation between *S. cerevisiae* and *C. elegans*, we first need to define which genes are orthologs of each other, which we do based on phylogenetic trees allowing for multi to multi orthology relations (Fig. 3). We found 18161 *C. elegans* gene pairs that have an orthologous pair in *S. cerevisiae* with a co-expression correlation higher than 0.6. Of these, 803 also have a correlation higher than 0.6 in *C. elegans* itself (Table 1). Defined this way, 4.4% of the co-expression is conserved, which is 12 times higher than expected assuming no conservation of gene co-expression. Vice versa, of the *S. cerevisiae* gene pairs that have an orthologous pair in *C. elegans* with a correlation higher than 0.6, 1215 also have a correlation higher than 0.6 in *S. cerevisiae* itself, which is 15 times higher than expected (Table 1).

Although significant ($P < 0.001$, determined by 1000 Monte Carlo simulations), the observed level of conservation of co-expression between *S. cerevisiae* and *C. elegans* is quite low (Table 1) as already reported [8]. However, given that at correlations higher than 0.6 in a single species there are still many false positive predictions, this apparent lack of conservation might be due to spuriously detected co-expressed genes. Consistent with this, genes with a high co-expression correlation in *C. elegans* ($R > 0.9$), which we expect to be truly co-regulated, are often co-expressed in *S. cerevisiae* (55%, $R > 0.6$). Interestingly, a considerable fraction (50%) of the gene pairs that have co-expression correlation higher than 0.9, but are not conserved ($R < 0$ in the other species), encode

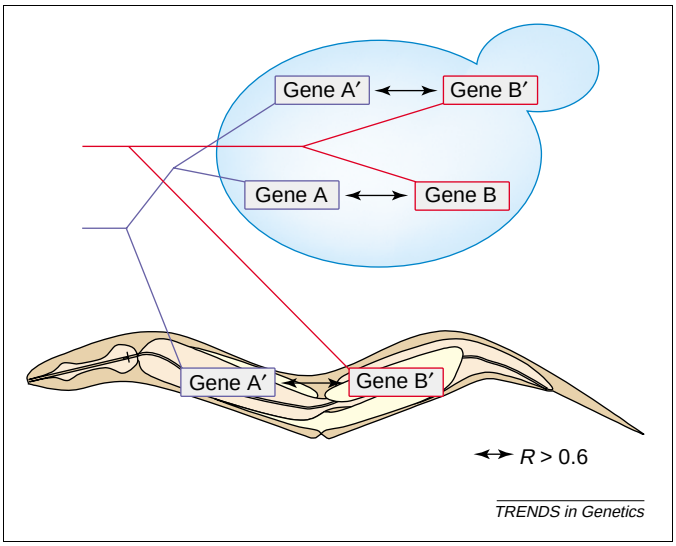

Fig. 2. Conservation of co-expression after gene duplication or speciation. Gene A′ and B′ in *Caenorhabditis elegans* are orthologs of gene A and B in *Saccharomyces cerevisiae*. Gene A′ and B′ in *S. cerevisiae* are paralogs of gene A and B in *S. cerevisiae*. Co-expression is defined by a correlation in the expression profile higher than 0.6, indicated by the arrow. We define a gene pair A-B to be a duplicated pair with conserved co-expression when the genes are co-expressed and their closest relatives (lowest, significant *E*-value in Smith–Waterman searches) A′ and B′ are also co-expressed. Orthologous conservation is the conservation of co-expression between A–B and A′–B′ between two species.

regulatory proteins. They include a TATA-binding protein (T20B12.2) that is co-expressed in *C. elegans* with a ring-type zinc-finger protein (EEED8.9), and in *S. cerevisiae* an RNA-binding protein (YOR319W) with a protein containing a BAF60b domain (YOR295W) that facilitates the function of transcriptional activators. The lack of conservation appears therefore to depend both on spurious co-expression and on rapidly evolving, regulatory interactions.

Next we determined conservation of co-expression between gene pairs within a species after parallel gene duplication (Fig. 2). The number of such pairs is actually higher than the number of pairs with co-expression conserved between species: 29031 in *C. elegans* and 2167 in *S. cerevisiae* (respectively 37 and 26 times higher than expected; $P < 0.001$). Conservation of co-expression within duplicated gene pairs coupled to divergence between the pairs, was studied by selecting the pairs A–B and A′–B′ where the correlations between A and B, and between A′ and B′ are both higher than between A and A′, and between B and B′. This conservation is lower than without divergence, but still higher than expected (Table 1). Thus, even after differentiation in expression pattern, there is significant conservation of co-expression.

Conserved co-expression improves accuracy of pathway prediction

Does the conservation of co-expression after gene duplication or speciation increase the likelihood of a functional relationship between co-expressed genes? Conservation after duplication in *S. cerevisiae* does indeed increase the accuracy levels for prediction of functional interactions, albeit at the expense of coverage of known interactions (Fig. 1). The results for *C. elegans* are similar, but there are not enough genes annotated in the PATHWAY database to establish the accuracy for conserved co-expression above 0.6. Higher accuracy is also achieved for the genes that are co-expressed in both species (Fig. 1). A similar result was described by Teichmann and co-workers [8], who found that 89% of the conserved co-expressed pairs between *S. cerevisiae* and *C. elegans* for which functional annotation was available were part of the same protein complex. However, in this analysis co-expression was defined in

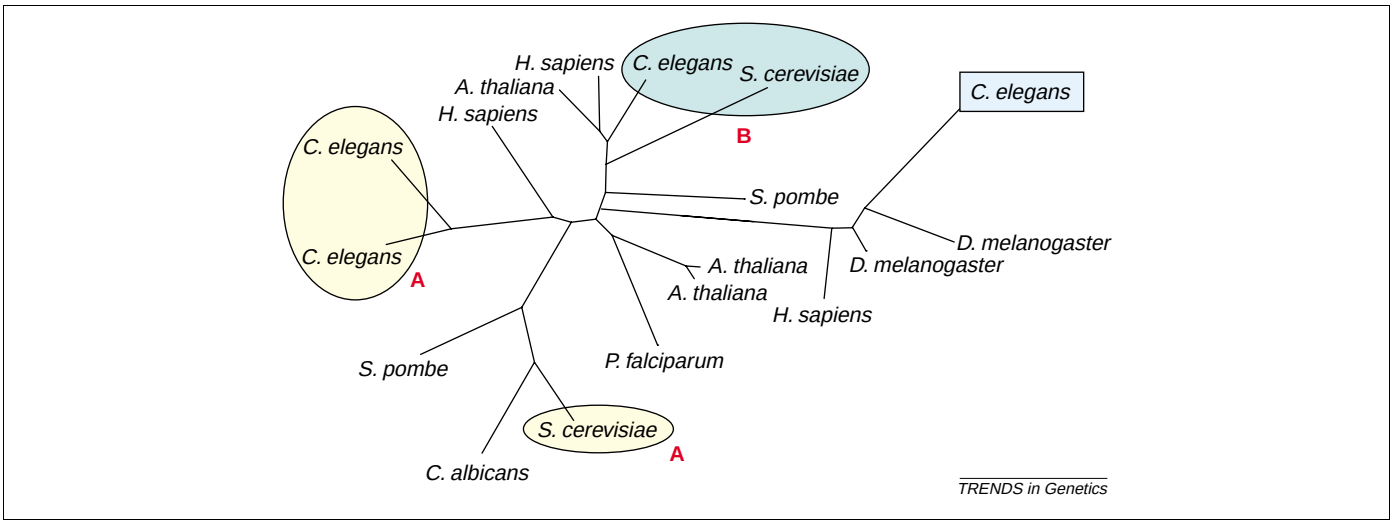

Fig. 3. Orthology prediction using an unrooted phylogenetic tree. Large-scale orthology prediction is generally done by the Best Bi-directional Hit approach or extensions thereof like COGs [24]. As orthology is an evolutionary relation we determine it using phylogenetic trees. Our method includes also inparalogs and is conceptually similar to INPARANOID [25]. All predicted protein-coding genes were obtained for both *Saccharomyces cerevisiae* [26] and *Caenorhabditis elegans* [27], as well as predicted genes of other complete genomes (to improve the quality of calculated phylogenies). Each *S. cerevisiae* gene is considered in turn to find orthologs in *C. elegans*. First we find homologies between all predicted genes and the gene under consideration by Smith–Waterman searches [28,29]. We include all genes with an E-value smaller than 0.01 and of which the region of homology is larger than the 50% of the length of the query. Groups of more than 250 proteins are reduced in size by applying a lower E-value cut-off. A multiple alignment is made with ClustalW [30] from the protein sequences of the gene and its homologs and a Neighbor-Joining [31] tree is calculated. For every query gene, we first select the largest branch containing the query gene and possible paralogs in *S. cerevisiae*, but no *C. elegans* genes. And after that the smallest branch that contains this branch as well as *C. elegans* genes, but no extra *S. cerevisiae* genes is selected. Orthology is assigned between all *S. cerevisiae*–*C. elegans* pairs in this branch. This results in the assignment of orthologous relations to the genes in the yellow circles, indicated with A, and to the genes in the blue circle, indicated with B. The boxed *C. elegans* gene has no orthologs in *S. cerevisiae*.

such a strict way that 93% of the conserved pairs already had a functional annotation and hardly any new predictions could be made. Note that our orthology prediction based on phylogenetic trees, instead of the Bidirectional Best Hit [7] method, allows ~50% more predictions to be made at a correlation higher than 0.6: instead of 799 there are 1215 predicted interactions in *S. cerevisiae*, and instead of 607 there are 803 predicted interactions in *C. elegans*.

Predicted interactions of *S. cerevisiae* genes were verified not only by the PATHWAY database, but also by using gene ontology (GO) annotations [9,10]. When involvement in the same biological process is defined as a common GO process category at the fourth level of specification, the accuracy achieved at a correlation threshold of 0.6 is 93% using orthologous conservation and 82% using paralogous conservation, compared with only 31% for all co-expressed pairs. There are insufficient reliable GO annotations on *C. elegans* genes (most are inferred by electronic annotation) to confirm their predicted interactions using GO.

Conserved co-expressed gene pairs for which only one of the genes is assigned to a pathway form a pool of genes to which we can now assign a pathway. From interspecies or intraspecies conservation, we predict a pathway for 55 and 95 *S. cerevisiae* genes, and for 54 and 596 *C. elegans* genes, respectively. For the vast majority of genes found by paralogous conservation (282 in *S. cerevisiae*, 2216 in *C. elegans*) and by orthologous conservation (91 in *S. cerevisiae*, 143 in *C. elegans*), neither gene in the pair is present in the PATHWAY database.

New predictions from old data

Co-expression conserved between *S. cerevisiae* and *C. elegans* of the hypothetical gene CAT5 (YOR125C, ZC395.2) and COQ2 (YNR041C, F57B9.4) confirms earlier predictions based on knock-out experiments [11] and homology relations [12] that CAT5 is 2-polyprenyl-3-methyl-6-methoxy-1,4-benzoquinone mono-oxygenase, which is involved in ubiquinone synthesis, as COQ2 encodes para-hydroxybenzoate: polyprenyl transferase, which is also involved in ubiquinone synthesis.

A prediction based on conservation of co-expression after duplication concerns the link between YBR052C and YDR074W. The gene YBR052C probably catalyzes a redox reaction, because it belongs to the WrbA family, which is homologous to flavodoxins. The gene YDR074W encodes trehalose-6-phosphatase [13], which is involved in starch and sucrose metabolism. For one redox enzyme in this pathway, glucoside 3-dehydrogenase, no gene has been described yet. This enzyme binds the co-factor flavin mononucleotide (FMN) [14] and has a molecular mass of 85 kDa [15] in *Agrobacterium tumefaciens*, where an ortholog of YBR052C is also present. The *Escherichia coli* ortholog, WrbA, forms multimers and also binds FMN [16]. The amino acid sequence of WrbA indicates a molecular mass of 22 kDa, implying a tetrameric organization consistent with the formation of multimers and the determined molecular mass of 85 kDa. We thus propose that YBR052C encodes the enzyme glucoside 3-dehydrogenase.

A more speculative prediction is that YKL033W-A (R151.8), whose co-expression with the endonuclease APN1 (T05H10.2) is conserved between species, is a 3′ phosphatase involved in DNA repair. The gene YKL033W-A contains a frameshift in the sequence of the published *S. cerevisiae* genome, but has also been sequenced without a frameshift [17] (accession number X71622) and has full-length orthologs in all sequenced eukaryotes. The human ortholog, GS1, is particularly interesting as it is an X-chromosome gene that escapes X inactivation [18]. The protein is homologous to haloacid dehalogenase-like hydrolases, a domain that has phosphatase activity, and is among others found as a 3′ phosphatase in T4 tRNA-repair enzyme, polynucleotide kinase [19]. DNA 3′ phosphatase reactions do have a role in repairing lesions in the DNA. This process involves APN1, which exhibits 3′ phosphodiesterase activity [20].

Modularity in pathway evolution

Of particular evolutionary importance is our finding of a substantial number of cases where, although the expression pattern of A′ and B′ has changed relative to their ancestors A and B, the co-expression of A′ and B′ is conserved. This seemingly contradicts the finding by Wagner that after duplication events, mRNA expression patterns diverge very quickly relative to amino acid sequence [21]. Yet, both results complement each other as we show that the co-expression is often conserved even when the expression patterns are not. However, a real contradiction with our results is apparent in a study of small molecule metabolism pathways in *E. coli* that showed that modular recruitment occurs very rarely [22]. Our observation of co-duplicated, diverged but still co-expressed genes suggests a substantial role for modularity in pathway evolution.

Outlook

Correlations between expression profiles do not necessarily imply co-regulation, and co-regulation does not always indicate functional interaction. Thus, it is important for function prediction to increase the reliability of co-expression data. Overlapping transcriptional clusters from different clustering methods have led to the prediction of functional categories for many genes [5]. Here we show that both intraspecies and interspecies conservation make expression data useful for the reliable prediction of specific functions.

Both types of conservation differ in their future applicability. Paralogous co-expression conservation has great advantages, because it relies on experimentation in only a single organism. Moreover, gene duplications are rampant in eukaryotes. The resulting noise in orthology prediction possibly distorts the usage of conservation of co-expression between species. However it is the very same gene duplication that increases the applicability of co-expression for function prediction.

Acknowledgements

This work was supported in part by a grant from the Netherlands Organization for Scientific Research (NWO).

References

1 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
2 Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126
3 Kim, S.K. *et al.* (2001) A gene expression map for *Caenorhabditis elegans*. *Science* 293, 2087–2092
4 Noordewier, M.O. and Warren, P.V. (2001) Gene expression microarrays and the integration of biological knowledge. *Trends Biotechnol.* 19, 412–415
5 Wu, L.F. *et al.* (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* 31, 255–265
6 Huynen, M. *et al.* (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* 10, 1204–1210
7 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901
8 Teichmann, S. and Babu, M. (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol.* 20, 407–410
9 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29
10 Dwight, S.S. *et al.* (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* 30, 69–72
11 Marbois, B.N. and Clarke, C.F. (1996) The COQ7 gene encodes a protein in *Saccharomyces cerevisiae* necessary for ubiquinone biosynthesis. *J. Biol. Chem.* 271, 2995–3004
12 Rea, S. (2001) CLK-1/Coq7p is a DMQ mono-oxygenase and a new member of the di-iron carboxylate protein family. *FEBS Lett.* 509, 389–394
13 De Virgilio, C. *et al.* (1993) Disruption of TPS2, the gene encoding the 100-kDa subunit of the trehalose-6-phosphate synthase/phosphatase complex in *Saccharomyces cerevisiae*, causes accumulation of trehalose-6-phosphate and loss of trehalose-6-phosphate phosphatase activity. *Eur. J. Biochem.* 212, 315–323
14 Hayano, K. and Fukui, S. (1967) Purification and properties of 3-ketosucrose-forming enzyme from the cells of *Agrobacterium tumefaciens*. *J. Biol. Chem.* 242, 3655–3672
15 van Beeumen, J. and de Ley, J. (1975) A ferrodoxin from *Agrobacterium tumefaciens*. *FEBS Lett.* 59, 146–148
16 Grandori, R. *et al.* (1998) Biochemical characterization of WrbA, founding member of a new family of multimeric flavodoxin-like proteins. *J. Biol. Chem.* 273, 20960–20966
17 Purnelle, B. *et al.* (1994) Analysis of an 11.7 kb DNA fragment of chromosome XI reveals a new tRNA gene and four new open reading frames including a leucine zipper protein and a homologue to the yeast mitochondrial regulator ABF2. *Yeast* 10, 125–130
18 Yen, P.H. *et al.* (1992) Isolation of a new gene from the distal short arm of the human X chromosome that escapes X-inactivation. *Hum Mol Genet* 1, 47–52
19 Galburt, E. *et al.* (2002) Structure of a tRNA repair enzyme and molecular biology workhorse T4 polynucleotide kinase. *Structure* 10, 1249–1260
20 Vance, J.R. and Wilson, T.E. (2001) Repair of DNA strand breaks by the overlapping functions of lesion-specific and non-lesion-specific DNA 3′ phosphatases. *Mol. Cell. Biol.* 21, 7191–7198
21 Wagner, A. (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist–selectionist debate. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6579–6584
22 Teichmann, S.A. *et al.* (2001) The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *J. Mol. Biol.* 311, 693–708
23 Ogata, H. *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34
24 Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28
25 Remm, M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052
26 Goffeau, A. *et al.* (1996) Life with 6000 genes. *Science* 274, 563–567
27 The *C. elegans* Sequencing Consortium, (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018
28 Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197
29 Pearson, W.R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* 276, 71–84
30 Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680
31 Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425

0168-9525/03/$ - see front matter
doi:10.1016/S0168-9525(03)00056-8