BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 3 2009, pages 322–330
doi:10.1093/bioinformatics/btn625

Gene expression

Gene expression trends and protein features effectively complement each other in gene function prediction

Krzysztof Wabnik[1,2,*], Torgeir R. Hvidsten[3,4], Anna Kedzierska[5], Jelle Van Leene[1,2], Geert De Jaeger[1,2], Gerrit T. S. Beemster[1,2], Jan Komorowski[3] and Martin T. R. Kuiper[1,2,6,*]

[1]Department of Plant Systems Biology, VIB Technologiepark 927, [2]Department of Molecular Genetics, Ghent University, Technologiepark 927, 9052 Gent, Belgium, [3]The Linnaeus Centre for Bioinformatics, Uppsala University, BMC Box 598, SE-751 24 Uppsala, [4]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-901 87 Umeå, Sweden, [5]Bioinformatics and Genomics Group, Center for Genomic Regulation (CRG) and Department of Applied Mathematics I, Polytechnic University of Catalonia, Barcelona, Spain and [6]Department of Biology, Norwegian University of Science and Technology, 7491 Trondheim, Norway

Received on April 29, 2008; revised on October 9, 2008; accepted on November 30, 2008

Advance Access publication December 2, 2008

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Genome-scale 'omics' data constitute a potentially rich source of information about biological systems and their function. There is a plethora of tools and methods available to mine omics data. However, the diversity and complexity of different omics data types is a stumbling block for multi-data integration, hence there is a dire need for additional methods to exploit potential synergy from integrated orthogonal data. Rough Sets provide an efficient means to use complex information in classification approaches. Here, we set out to explore the possibilities of Rough Sets to incorporate diverse information sources in a functional classification of unknown genes.

Results: We explored the use of Rough Sets for a novel data integration strategy where gene expression data, protein features and Gene Ontology (GO) annotations were combined to describe general and biologically relevant patterns represented by If-Then rules. The descriptive rules were used to predict the function of unknown genes in *Arabidopsis thaliana* and *Schizosaccharomyces pombe*. The If-Then rule models showed success rates of up to 0.89 (discriminative and predictive power for both modeled organisms); whereas, models built solely of one data type (protein features or gene expression data) yielded success rates varying from 0.68 to 0.78. Our models were applied to generate classifications for many unknown genes, of which a sizeable number were confirmed either by PubMed literature reports or electronically interfered annotations. Finally, we studied cell cycle protein–protein interactions derived from both tandem affinity purification experiments and *in silico* experiments in the BioGRID interactome database and found strong experimental evidence for the predictions generated by our models. The results show that our approach can be used to build very robust models that create synergy from integrating gene expression data and protein features.

Availability: The Rough Set-based method is implemented in the Rosetta toolkit kernel version 1.0.1 available at: http://rosetta.lcb.uu.se/

Contact: kuiper@nt.ntnu.no; krwab@psb.ugent.be

*To whom correspondence should be addressed.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Genome-scale 'omics' data from technologies such as transcriptomics, proteomics and metabolomics are essential for data-driven systems biology approaches (Patterson, 2003; Tikunov *et al*., 2005; Wu *et al*., 2002). Together with other biologically relevant information, 'omics' data analysis remains vitally important for understanding gene function and gene regulatory networks. The type of biological information may vary greatly from regulatory motifs in genome sequences (Aderem and Hood, 2001; Jensen *et al*., 1999) to protein physical/chemical properties (Jensen *et al*., 2003). Whereas many procedures have been devised to mine single data sources, far fewer approaches allow the mining of combined sources of information.

Data integration has previously been used to merge data from several microarray experiments (Schena *et al*., 1995; Yeung *et al*., 2003), merge various types of sequence information (Jensen *et al*., 2002) or to combine completely different data sources such as regulatory binding sites and microarray data (Hvidsten *et al*., 2005), or microrarray data and gene function annotations (Lægreid *et al*., 2003). Amongst the available approaches for analyzing such data, clustering techniques group genes that behave similarly across experiments or in a specific time interval (Quackenbush, 2001). Similarities between gene expression profiles have been used to assign functions to unknown genes based on the concept of 'guilt by association': genes with similar expression characteristics might play a role in the same biological mechanism (Walker *et al*., 1999; Wu *et al*., 2002). Nevertheless, even genes with similar, well-described functions can display highly divergent gene expression behavior. Such behavior may reflect their involvement in various processes, and be the result of a complex regulation of their activity levels that allows their expression levels to meet the requirements of widely different environmental or physiological conditions. Their complex expression patterns contain information that can

be interpreted by more advanced functional inference approaches that exploit a supervision or training step. Supervised techniques aim to integrate biological knowledge into the learning process. Examples include support vector machines (Brown *et al*., 2002), artificial neural networks (Honeyman *et al*., 1998), decision tree learning (Adie *et al*., 2005) and the Rough Set approach (Andersson *et al*., 2005; Komorowski *et al*., 1999, 2002). Machine learning methods (Schlitt *et al*., 2007) are among the approaches able to generate decision-making systems that can utilize input knowledge. In contrast to the clustering techniques their performance can be qualitatively measured using statistical validation on the training set [i.e. cross-validation (CV), boot-strapping].

Here we present a novel data integration pipeline where gene expression data (time profiles), sequence information (protein properties) and gene function annotations are combined and analyzed to find common denominators that can classify functional groups. We used the Gene Ontology (GO) (Ashburner *et al*., 2000) as a source of functional annotation, often used to link functional classes to groups of co-clustered genes (Maere *et al*., 2005). We decided to apply the Rough Sets framework (Komorowski *et al*., 1999; Pawlak, 1992) to study various biological information derived from independent data sources, mainly because of its flexibility for data integration, and its previous application in biology (Hvidsten *et al*., 2003; Komorowski *et al*., 2002; Lægreid *et al*., 2003) or medical research (Dennis *et al*., 2005; Słowiński *et al*., 2002). In general, Rough Set-based approaches can easily deal with occasional data inconsistency and, different from neural networks or support vector machines, can generate models that are readily interpretable (readable If-Then rules). The Rough Set models consist of deterministic (certain) and nondeterministic (probabilistic) rules. Unlike decision tree learning approaches that rank features individually, Rough Set rules result from initially considering all features and then removing those that are least effective or redundant. We demonstrate that the fusion of gene expression and protein sequence-derived information provides a better understanding of cell regulation systems than either of these data sources individually, and that the resulting rule model can be used to provide high-quality hypotheses about the function of unknown genes.

2 METHODS

2.1 Time series data

Gene expression time series data for *Schizosaccharomyces pombe* was from Rustici *et al*. (2004), and consisted of a time course experiment with synchronized cells monitored over two full cell cycles. Normalized signal ratios of synchronized culture sampled at different time point versus unsynchronized cells were downloaded from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/), available at accession number E-MEXP-54. Triplicate measurements were averaged, yielding data for nearly 4600 known probes at 20 time points. The regularized Expectation–Maximization algorithm (EM) (Schneider, 2001) was used to reconstruct missing expression values. For the *Arabidopsis thaliana* gene expression time series data, microarray data for nearly 15 000 gene probes measured at eight stages of leaf development were produced within the CAGE Project (http://www.cagecompendium.org). The entire dataset was log2 transformed and filtered on three criteria (background, foreground and average standard deviation) to retain only the genes with spots above background, yielding nearly 11 000 genes. The regularized EM algorithm was used to fill in 3% of the missing values. The CAGE experiment is available under accession number E-CAGE-198 at Array Express. As a final processing step, both datasets were transformed using a moving average filter (http://www.stats.gla.ac.uk/steps/glossary). This popular method retains the general trends in the time series data while smoothing out spikes that greatly affect the data representation method (Lægreid *et al*., 2003).

2.2 Protein features collection

The protein features were derived directly from amino acid sequence and calculated from (predicted) open reading frames (Lichtenberg *et al*., 2003).

For more detailed information see Supplementary Material files 1 and 2.

2.3 Annotation sources

The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org) provides a comprehensive source for *A.thaliana* GO annotations, including their association to genes. *Schizosaccharomyces pombe* annotations were obtained from the GeneDB project database at the Sanger Institute (http://www.genedb.org). The revised GO database released on July 21, 2007 was used to obtain gene annotations.

2.4 Model overview

2.4.1 Feature mixture Gene expression time series profiles were transformed into sets of trends (templates) consisting of 'increase', 'decrease' and 'constant', denoting the relevant expression difference (defined by parameters) observed in a short time intervals covering at least three time points (Hvidsten *et al*., 2001). To fit templates to the original gene expression data (leaf development and cell cycle), we optimized the template parameters with respect to the performance of the resulting templates in a subsequent classification of genes. This optimization approach showed that the 'constant' template should be assigned to intervals when the log2 deviation from the mean expression did not exceed 0.05. The 'increase/decrease' templates were assigned when the log2 expression ratio increased/decreased, respectively, by at least 0.4 over at least three subsequent time points. In practice, these parameters can be tuned for any type of time series data where trends can be observed. The protein feature data were transformed to the corresponding protein feature ranges using an equal frequency binning algorithm (Boulle, 2005). The combination of gene expression trends and protein features carries descriptive information about the function of a gene.

2.4.2 Model training The rule models were trained using genes with known annotations to GO classes that were part of 'biological process' or 'molecular function'. Genes in the training set were linked with the general trends in gene expression data and discrete protein feature ranges (called attributes). Rough Set theory considers data in terms of sets of genes with the same attribute strings also called equivalence classes. A GO class is said to be 'rough' if it cannot be uniquely defined in terms of the equivalence classes, and thus has to be defined by a lower approximation (all equivalence classes containing genes only belonging to the GO class) and an upper approximation (all equivalence classes containing at least one gene belonging to the GO class). Minimal sets of attributes that uphold the same discriminatory power (the same lower and upper approximation) as the full set of attributes are called 'reducts'. In practice, approximate reducts that uphold most of the discriminatory power are preferred since they tend to result in more general rule models and avoid overfitting. Such approximate reducts were efficiently calculated using a genetic algorithm (Vinterbo and Øhrn, 2000) and used as templates for generating rules. A rule example is provided as follows:

If 1.04-1.06(increase) and 1.07-1.1(decrease) and N-linked glycosylation [-0.641, 0.149) and Positive residues [-0.295, 0.491) **Then** *leaf morphogenesis (GO:9965)*

To facilitate interpretation, we reduced the size of the initial rule model by removing rules that were either too specific (rarely contributing to classification) or too general (attracting high numbers of false positives). To achieve this, we used a pruning algorithm that, for each GO class,

sorted the rules by *P*-value and iteratively selected the rules that covered the highest number of genes not already covered by at least 100 rules. Classification of a gene is obtained by allowing all matching rules to 'cast votes' to assign that gene to their corresponding GO class(es) equivalent to the support that these rules had acquired in the training set (for a complete treatment of rule induction and classification, see Hvidsten *et al*., 2003). Votes were then calibrated by classifying 10 000 randomized examples, fitting an Extreme Value Distribution (Hernandez and Johnson, 1984) to these scores and calculating false discovery rates (*P*-values). All GO classes with a *P*-value lower than a corresponding selection thresholds were finally selected as predictions.

2.4.3 Robustness of the model The statistical significance of the models was assessed by a 40-fold CV over the training examples (Kohavi, 1995). The training set was divided into 40 equally sized subsets. One subset of the training examples was used for testing the model and the remaining 39 subsets to support the training. In 40 repetitions, each subset was once a test set and 39 times part of the training. The CV test estimated the predictive power of the classifier. This was done by plotting, for each GO class, sensitivity against the false positive rate (FPR; 1-specificity) for all possible selection thresholds, and then reporting the area under this so-called ROC curve (AUC). The specificity for each class was calculated as TN/(TN + FP) where TN (true negatives) is the number of genes neither classified nor annotated to a functional class and FP (false positives) refers to the genes wrongly predicted to belong to a class. Class sensitivity was defined as TP/(TP + FN) where TP (true positives) refers to the genes correctly assigned to a class, and FN (false negatives) concerns genes that could not be assigned to the class that they belong to according to their annotation. To classify unknown genes we had to fix the selection thresholds. For each class, we minimized the expression:

$$\min_{\tau} c \cdot (1 - \text{specificity}(\tau)) + (1 - \text{sensitivity}(\tau)) \quad (1)$$

with respect to the threshold τ where c describes the parameter that regulates the FPR. In this article, we visualized the model performance over all classes by plotting precision against coverage for all possible values of c [precision-recall (PR) curve in Fig. 2]. Precision was calculated as the fraction of correct classifications TP/(TP + FP) over all classes, while coverage was the fraction of annotations correctly predicted TP/(TP + FN). Balanced selection thresholds for each class corresponded to the Break-Even Point (BEP) (precision $\approx$ coverage) in the PR curves over all classes. We found that the c value for BEP was equal to 0.3, resulting in precisions of 52% and 51% and coverages of 52% and 45% for the *A.thaliana* and *S.pombe* models, respectively.

3 RESULTS AND DISCUSSION

3.1 Data source and knowledge integration

Our data integration procedure combines gene expression time series data (time profiles), protein properties obtained by bioinformatics analysis of amino acid sequences (protein features) and gene function annotation for *A.thaliana* and *S.pombe*. In *A.thaliana*, the time series data included transcript levels of 10 820 genes at eight different stages of leaf development (see Section 2). For *S.pombe*, we used the detailed dataset of Rustici *et al*. containing expression measurements of 4648 genes monitored during two full cell division cycles. Protein features were calculated for each gene product and comprised protein modification sites, subcellular localization or physical/chemical properties (Jensen *et al*., 2003; more details in Supplementary Material files 1 and 2). From the gene function annotations obtained from GO, we selected only genes that were tagged with the most reliable evidence codes (manual and experimental curation). These most reliable evidence codes included Interferences made by the Curator (IC), Inferred from Direct Assay (IDA), Inferred from Genetic Interactions (IGI) and Traceable Author Statement (TAS). For more information on evidence codes and their ranking reliability, we refer to the Evidence Code Guide (http://www.geneontology.org/GO.evidence.shtm). In *A.thaliana*, 2132 out of 10 698 protein-coding genes had satisfactory evidence codes, while this was true for 1598 of 4387 genes in *S.pombe* (release from July 21, 2007). We used both GO biological process and molecular function annotations.

The structure of the GO hierarchy allows functional classes either to be aggregated into more general classes in order to increase their size or to be broken down into smaller classes of higher biological specificity (children classes). The initial classes associated with the selected annotations (see above) had either very low or very high numbers of members. Based on Lægreid *et al*. (2003), we assumed that associated classes needed to have at least four gene members to ensure an adequate representation of the feature space. From the classes with fewer than four members we built classes large enough to support rule induction in a training step. Similarly, classes with more than 30 genes were broken down into smaller ones. This allowed us to work with more detailed biological functions of the genes. Many genes can be members of multiple classes leading to functional ambiguities in the training process. By introducing the 'singleton' gene concept, meaning that one gene is associated with only one class, we eliminated this problem. The class normalization process is explained in Supplementary Figure S1. In total, we used 375 and 269 genes of *A.thaliana* and of *S.pombe*, comprising 32 and 22 GO classes, respectively. The number of selected classes may seem low, but they represent metabolism, development, cell cycle regulation, protein processing, transcription and translation and many other processes involved for instance in transport of proteins and their processing. These processes cover functions that are essential to every living organism. We used them subsequently to induce the Rough Set If-Then rules.

3.2 Building the rule models

First, general gene expression trends were produced by converting quantitative gene expression time series data into the qualities 'increased', 'decreased' and 'constant' denoting what happens to expression over a short time interval (see Section 2). The reason for such discretization was to convert highly complex temporal gene expression to a string of attributes that could be linked to genes involved in one biological process (Lægreid *et al*., 2003). Discretization allowed us to break up temporal expression profiles into a combination of several discrete subinterval profiles. Essentially this would allow the discovery of similarities in transcription behavior over shorter time frames. To enrich the number of attributes defining specific genes, we also discretized 39 protein features into numerical ranges. These protein feature ranges hold detailed information about primary amino acid sequence variation, especially descriptive for proteins that are far apart in sequence space. Protein features include amino acid composition, local residues and consensus motifs, regions rich in proline, glutamic acid, serine and threonine (PEST) and many other more complex patterns that can be assigned to general or specific types of protein function (Jensen *et al*., 1999, 2002). If-Then rules were constructed that describe minimal set of general trends in expression data and protein feature ranges (conditional attributes) that discern genes

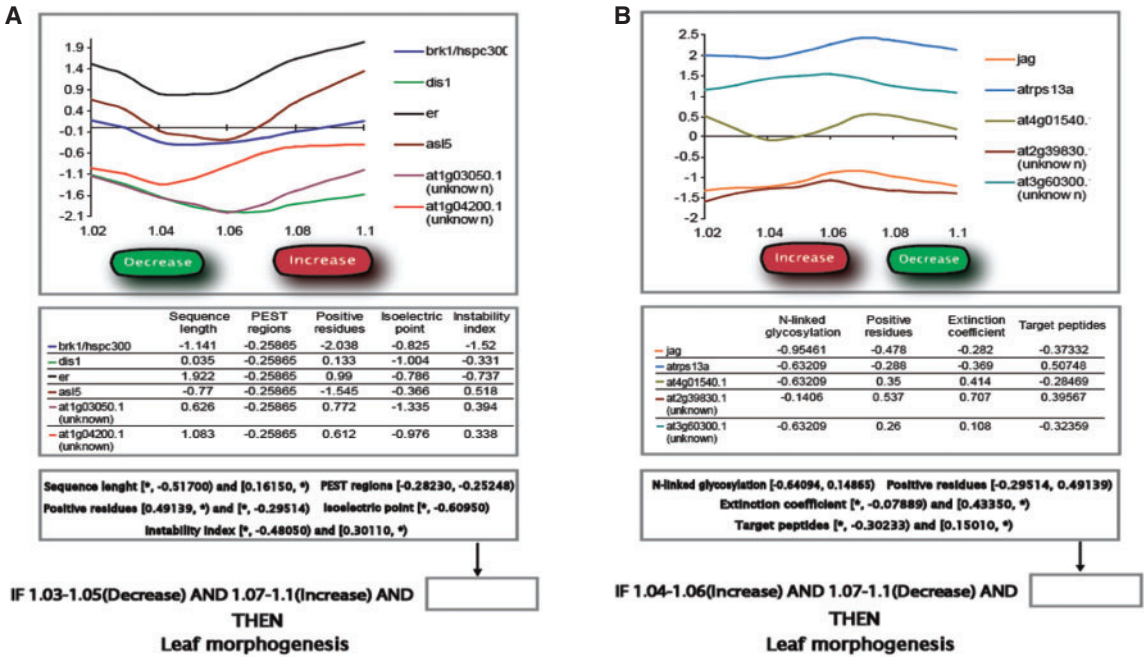

Fig. 1. Examples of biological information used to model *leaf morphogenesis*. Common patterns (expression trends and protein feature ranges that were a part of the rule description) for two subgroups of genes. (**A**) Co-regulated genes with general expression trends (down-regulation and up-regulation) and relevant protein feature ranges. (**B**) Genes showing inverse-regulation and different protein feature ranges compared to (A). These subgroups of genes define various biological patterns that are used to determinate the If-Then rule model for *leaf morphogenesis*.

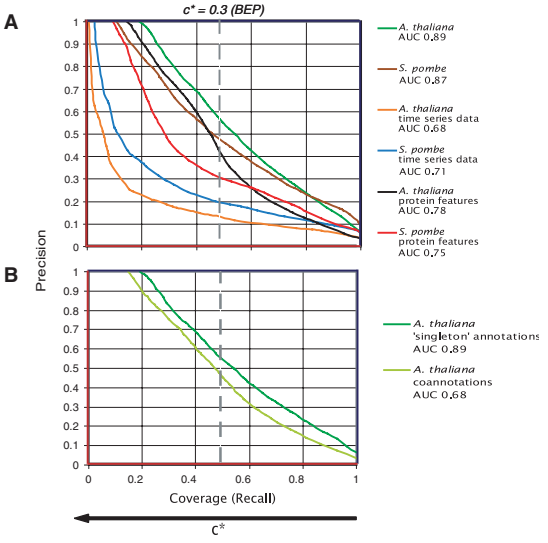

Fig. 2. PR curves for *A.thaliana* and *S.pombe* models. The average AUC values were estimated over all functional classes of known genes. The c* denotes the parameter that regulates the FPR [see Equation (1) in Section 2]. At the BEP, c* was 0.3 (dashed line). Note that a random classifier would on average have precision = 1/no. classes, while coverage would go from 0 to 1 depending on the number of predictions (i.e. guesses). (**A**) Comparison of models trained from both gene expression profiles and protein features, and models trained from either gene expression profiles or protein features. (**B**) Comparison of models trained with either singletons or co-annotated genes.

participating in one process from genes participating in all other processes. The pruned rule sets (see Section 2) for *A.thaliana* and *S.pombe* contained 12 597 and 8326 rules, respectively. Figure 1 illustrates how biological data can be combined and interpreted in the form of If-Then rules for *leaf morphogenesis*. By inspecting the information patterns in these rules, we observed that two different groups of genes share some general trends in regulation of their expression within short time intervals: 1.02–1.05 and 1.07–1.1 (Fig. 1A), 1.04–1.06 and 1.08–1.1 (Fig. 1B). This example nicely illustrates the complex nature of gene regulation—two groups of overall co-regulated genes participating in the same process are inversely regulated over a short interval of the time course. Analysis of the rules that describe this set of genes intuitively suggests a possible explanation for the observed behavior: six genes from one group (Fig. 1A), including two hypothetical genes, are initially down-regulated, possibly resulting in delayed activation of five genes from the other group (Fig. 1B). After some time, an activation of these six genes occurs (Fig. 1A), accompanied by a corresponding inactivation of the five genes from the other group. Also conspicuously, each group of genes presented in Figure 1 seems to share some common protein feature ranges. By assuming that proteins are more likely to interact with each other when they collaborate to perform the same biological task, they are also likely to share similar features that identify them to the cellular machinery for modification and sorting. The numbers behind the protein features represent the enrichment relative to the mean, of the particular features in terms of motifs, local residues, physical/chemical properties applied to each amino acid sequence. For proteins coded by the group of six genes (Fig. 1A), only

Table 1. Model performance

PROCESS (*A.thaliana*)	AUC	SE	THR	Sens.	Spec.	GO
Signal transducer activity X (GO:4871)	1.00	0.00	3.26E-04	1.00	1.00	MF
Protein phosphorylated amino acid binding (GO:45309)	1.00	0.00	1.94E-03	1.00	1.00	MF
Water channel activity (GO:15250)	1.00	0.01	3.50E-06	1.00	0.99	MF
SNAP receptor activity (GO:5484)	0.99	0.02	1.45E-02	0.91	0.99	MF
Cellulose and pectin-containing secondary cell wall biogenesis (GO:9834)	0.98	0.04	3.06E-03	0.60	0.99	BP
Amino acid transmembrane transporter activity (GO:15171)	0.97	0.04	7.72E-06	0.44	0.99	MF
Transcription initiation (GO:6352)	0.97	0.05	3.01E-04	0.67	1.00	BP
Nucleobase, nucleoside, nucleotide and nucleic acid transmembrane transporter activity (GO:15932)	0.96	0.04	2.50E-05	0.83	0.98	MF
Branched chain family amino acid biosynthetic process (GO:9082)	0.95	0.06	1.91E-04	0.33	1.00	BP
Shoot development X (GO:48367)	0.91	0.05	3.93E-03	0.59	0.98	BP
Brassinosteroid metabolic process (GO:16131)	0.90	0.07	3.76E-04	0.63	1.00	BP
Aromatic amino acid family biosynthetic process (GO:9073)	0.90	0.07	2.02E-03	0.10	0.99	BP
RNA metabolic process X (GO:16070)	0.90	0.06	4.61E-03	0.36	0.98	BP
Sulfur metabolic process (GO:6790)	0.90	0.04	3.49E-02	0.59	0.92	BP
Cation transport X (GO:6812)	0.90	0.06	2.91E-04	0.57	0.99	BP
Sequence-specific DNA binding (GO:43565)	0.90	0.07	9.60E-04	0.22	1.00	MF
Response to auxin stimulus X (GO:9733)	0.89	0.05	2.71E-02	0.80	0.97	BP
Transferase activity, transferring glycosyl groups X (GO:16757)	0.89	0.08	1.10E-02	0.86	0.97	MF
Positive regulation of enzyme activity (GO:43085)	0.88	0.10	1.00E-08	0.10	1.00	BP
Structural molecule activity (GO:5198)	0.86	0.07	2.15E-02	0.50	0.96	MF
Carotene metabolic process (GO:16119)	0.85	0.09	4.18E-03	0.29	1.00	BP
RNA splicing (GO:8380)	0.85	0.07	2.10E-04	0.64	1.00	BP
Auxin binding (GO:10011)	0.85	0.11	1.58E-04	0.80	0.99	MF
Regulation of cell cycle (GO:51726)	0.85	0.06	1.74E-02	0.25	0.98	BP
Red, far-red light phototransduction (GO:9585)	0.85	0.09	2.79E-04	0.43	1.00	BP
Histidine biosynthetic process (GO:105)	0.84	0.10	5.97E-04	0.33	1.00	BP
Pectin biosynthetic process (GO:45489)	0.83	0.11	3.35E-04	0.20	1.00	BP
Peroxisome organization and biogenesis (GO:7031)	0.83	0.07	5.21E-02	0.64	0.94	BP
Response to starvation (GO:42594)	0.81	0.12	1.62E-04	0.40	0.99	BP
Ethylene mediated signaling pathway (GO:9873)	0.80	0.05	1.22E-02	0.43	0.97	BP
Photosynthesis (GO:15979)	0.78	0.06	3.44E-02	0.46	0.95	BP
Leaf morphogenesis (GO:9965)	0.72	0.06	6.25E-02	0.46	0.92	BP
Average[a]	**0.89**	0.06	1.03E-02	0.51	0.98	
c[b]	**0.30**					
Precision[c]	**0.52**					
Coverage[d]	**0.52**					

The 40-fold CV estimates of the area under ROC curve (AUC), the standard error (SE) for AUC, specificity and sensitivity for a fixed selection thresholds (THR) for the *A.thaliana* model (Supplementary Table S3 for *S.pombe* model). For all genes in the training set the correct class assignment is known from their GO annotation. The threshold (THR) refers to the *P*-value threshold (selection threshold). The last column named Role describes whether the class originates from GO Biological Process (BP) or Molecular Function (MF). The letter X in the process column indicates classes that resulted from aggregating small subclasses. [a]Average AUC over all classes in bold. [b]Cost on false positives to balance the model (precision = coverage) and 'c' describes the parameter in Equation 1 (see Section 2) resulting in the BEP. [c]The overall precision of the model. [d]Model coverage for the selected thresholds.

five of the most discriminative protein features from a total of 39 features occurred in their rules. We also observed that the combination of relevant features retained in the rules can be slightly different for other proteins involved in the same biological process or function (Fig. 1B). This example shows how a biological process or molecular function can be explicitly modeled using expression and sequence-derived data.

By applying a 40-fold CV to the training set (randomly dividing data into training and test sets, see Section 2), we assessed the classification performance of the If-Then rule models. These models showed a high classification quality over all classes (average AUC values 0.87–0.89; Table 1 and Supplementary Table S3), meaning that the data integration pipeline (data discretization, rule building and pruning) was quite successful in combining essential functional information contained in gene expression data and protein sequence features with GO annotation classes. CV estimates were considered to indicate the expected quality of the classification of unknown genes. PR curves (Davis and Goadrich, 2006) were then used as an alternative to ROCs to visualize the model performance in terms of precision and coverage. The advantage of PR over ROC curves is that they show the model performance over all classes and not only for each class separately. We used these PR plots to rate selection thresholds for the models (see Section 2). It resulted in a balanced classifier (precision $\approx$ coverage) where 52% of the predictions were correct and 52% of the GO annotations were predicted for *A.thaliana* model. For the *S.pombe* model, precision was 51% and

coverage 45% (Fig. 2). These models could be adjusted to satisfy more strict requirements of higher precision (fewer predictions, but of higher quality) or higher coverage (more predictions, but of lower quality) by shifting the selection thresholds (see Section 2).

3.3 Model validation

To test the importance of data integration, we rebuilt the models with either only protein features or gene expression time series data. These altered models were then subjected to a 40-fold CV (described above) and the resulting AUC scores and PR curves were compared with those obtained from the original models. We found that models performed better when both protein features and gene expression data were integrated (higher average AUC value, better precision/coverage; Fig. 2A). Separately, the gene expression data nicely contributed to the interpretation of various transcriptional and cellular processes, including transcription initiation, regulation of enzyme activity, cell cycle regulation and pectin biosynthesis. In *A.thaliana,* for instance, the predictions of cell cycle regulation were more robust using gene expression trends (AUC 0.73) than solely based on protein features (AUC 0.64). That was also true for the class *leaf development* (AUC value of 0.71 using gene expression data), however, here the performance of protein features was also high (AUC 0.68) suggesting that protein modifications and sorting might be especially important for this process. Because protein features represent functional parameters of proteins, they efficiently capture sequence information essential to molecular functions (i.e. molecule binding, protein transport and enzyme activity) and some post-transcriptional processes (i.e. RNA splicing). An example would be the *auxin binding* class that obtained an AUC of 0.51 using expression data, and an AUC of 0.87 using protein features. The high overall prediction quality of the models (i.e. AUC of 0.85 for the cell cycle class, Table 1) suggests that gene expression, protein features and GO annotations can reinforce and complement each other in descriptive general rules. Compared to the recent study of Chua *et al*. (2007) on different data integration scenarios including various data sources, and including a training method comparison, we found that the If-Then rules models presented here provide well-balanced (precision $\approx$ coverage) and powerful (AUC up to 0.89) classifiers for both GO biological processes and GO molecular functions. Moreover, here we used low level GO classes (high specificity) that are essential for better understanding of protein expression and function. Besides, the inspection of If-Then rule models led us to interesting findings on proteome dynamics (see above). We also demonstrated the advantage of using 'singleton' class information (each gene assigned only to one class) by applying our CV test to the *A.thaliana* model with a set of 375 genes with a maximum of two GO annotations per gene (giving rise to 473 training examples). Figure 2B illustrates the results of this analysis. It is evident from the lower AUC score that the 'singleton' genes provide more precise and powerful models. Reclassification of known genes (i.e. using a model induced from all known genes to classify those same genes) of *A.thaliana* showed 420 classifications for 359 out of 375 genes, 85% of these were in agreement with the available annotations. For *S.pombe*, 316 classifications were generated for 251 out of 275 known genes with 84% in agreement with the annotations. However, both models also generated classifications that were not part of the training input set. We assessed the possibility that they might represent augmented knowledge of gene function annotations not previously incorporated into the GO. We tested all these additional classifications of known *A.thaliana* and *S.pombe* genes by searching the PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) database for high impact citations where gene function was recently reported based on experimental evidence. For the *A.thaliana* genes, we could confirm 51% of these additional annotations and for *S.pombe* we found that literature offered supporting information for 36% of our 'FP' classifications (Table 2). Examples of confirmed predictions for *A.thaliana* include AGO1 (encoding RNA splicing enzyme, Vaucheret *et al*., 2006) predicted to be an RNA splicer by our method, and at1g30210.2, controlling the morphology of shoot lateral organs via negative regulation of the expression of boundary-specific genes in *Arabidopsi*s (Koyama *et al*., 2007). Examples for *S.pombe* include kap111, previously known to be involved in mitotic cell cycle (Chen *et al*., 2004) (predicted by us to participate in mitotic metaphase/anaphase transition) and ppc89, required for proper coordination between the nuclear-division cycle and cytokinesis (Rosenberg *et al*., 2006) (predicted to be involved in cytokinesis). This literature examination revealed the ability of the models to accurately assign documented, but presently missing functions to known genes, further demonstrating the usefulness of the data integration scheme.

3.4 Classification of unknown genes

In total, we obtained one or more classifications for almost 40% of the genes in the entire dataset (Supplementary Table S1). Of the 4058 *A.thaliana* genes, 1964 were unknowns (not previously functionally annotated). As discussed above, 359 represented training examples and the rest (1735) had insufficient evidence codes (only electronically curated annotations). For *S.pombe*, 269 genes were unknowns and 251 were training examples, whereas 1197 genes had less reliable evidence for their function. For the genes tagged with insufficient evidence codes, we extracted the homology information (NCBI) and their recent GO annotations. For 29% and 26% of these genes in *A.thaliana* and in *S.pombe*, our classifications were confirmed by homology information or GO annotations that result from computational protein or motif analysis. We found that their predicted classes represented the same GO hierarchy level, or higher levels, in the same regulatory pathway. Most importantly, consensus predictions where our classifications agree with those of other computational approaches signify more reliable functional annotation. An example of such a detailed annotation is given by *A.thaliana* protein at2g27970.1 that was suspected to be a cyclin-dependent kinase linked to cell cycle, and predicted by our method to be a regulator of cell cycle (Table 2).

We present a general overview of the classification results in Supplementary Table S2, including the processes and functions that are crucial to living organisms (i.e. cell cycle regulation, activities, binding and development). For full lists of generated classifications, we refer to Supplementary Material file 3.

3.5 Predicted cell cycle protein interactions

Many cellular processes are known to depend on stable protein–protein interactions, indicating that protein complexes are crucial for many aspects of cell biology. Identification of such complexes can be a first step to assess their function. These complexes can be formed from highly diverse proteins, with little sequence similarity. Cell cycle-related proteins may undergo post-translational

Table 2. Model validation

Gene alias	Gene description (NCBI)	GO biological process or molecular function annotations	Predictions of biological process or molecular function	Comments/Evidence codes*
Fat1g30210.2, AT1G30210 *A.thaliana*	tcp24; transcription factor; similar to tcp family transcription factor. putative (gb:cad91129.1); contains interpro domain tcp transcription factor; (inter-pro:ipr005333)	(bp) RNA metabolic process	(bp) RNA metabolic process; (bp) shoot development	TCP transcription factors control the morphology of shoot lateral organs via negative regulation of the expression of boundary-specific genes in Arabidopsis (PMID: 17307931)
ago1, AT1G48410, *A.thaliana*	Encodes an RNA slicer that selec-tively recruits microrna and sirnas. there is currently no evidence that ago1 slicer is in a high molecular weight RNA-induced silencing complex (RISC)	(bp) Leaf morphogenesis	(bp) Leaf morphogenesis; (bp) RNA splicing	ago1 slicer is in a high molecular weight RNA-induced silencing complex (risc) (PMID: 16600876)
at2g27970.1, AT2G27970, A.thaliana	cks2 (cdk-subunit 2); cyclin-dependent protein kinase; similar to cks1 (cdk-subunit 1). cyclin-dependent protein kinase (gb:aas79576.1)	(bp) cell cycle; (mf) cyclin-dependent protein kinase activity	(bp) regulation of cell cycle	rca;nd;rca*
ppc89, mug127, SPAC4H3.11C, *S.pombe*	Spindle pole body protein ppc89	(bp) Recombinational repair	(bp) recombinational re-pair; (bp) regulation of cytokinesis	Nuclear division and cytokinesis checkpoint (PMID: 16775007)
kap111, SPAC22G7.02, *S.pombe*	karyopherin kap111	(mf) Protein transporter activity	(bp) mitotic metaphase/ anaphase transition; (mf) protein transporter activity	Mitotic cell cycle spindle assembly checkpoint (PMID: 15116432)
mbx1, SPBC19G7.06, *S.pombe*	mads-box transcription factor mbx1	(bp) Cytokinesis; (bp) g2/m transition of mitotic cell cycle; (bp) regulation of transcription. mitotic; (mf) DNA binding	(bp) Regulation of cytokinesis	imp;imp;iea;ida;tas;tas; iss;iea;tas*

Examples of FP reported in the literature or deduced from homology indicating that the classifications might yet be correct and thus reflects new knowledge or missing annotation (white fields). Examples of classifications of 'test-set' genes (gray fields) supported by recent GO annotations and homology information. These annotations had no evidence codes of the highest confidence level, but confirmed the class prediction. The full lists of FP and 'test-set' classifications we refer to Supplementary Material files 6 and 7.

modification, and rapid degradation. They can also co-localize or interact with each other or modulate the expression of other genes of interest. We speculate that proteins involved in the formation of complexes show similarities in expression patterns and undergo common modification and translocation mechanisms. Our predictions suggest many new cell cycle proteins that could be involved in protein–protein interactions and therefore be part of complexes. To test the significance of these predictions, we used the BioGRID protein–protein interaction database (Breitkreutz *et al*., 2008) to extract known interactors for 24 documented *S.pombe* core cell cycle proteins through 'in silico' experiments. Such experiments represent the mix of confirmed and predicted interactions between 24 core cell cycle proteins (baits) and 265 'pull-down' proteins (preys). We assumed that the preys associated with a bait in a protein complex would have a function similar to the bait gene (Hishigaki *et al*., 2001; Hollunder *et al*., 2005). We found that 49% of our predictions concerning 113 of the 265 pull-down proteins in *S.pombe* were reported as cell cycle interactors (*P*-value of 1.64E-08) (Supplementary Table S2; Supplementary Material file 5). This provided additional support that our model could provide correct predictions for unknown genes. Finally, we applied a similar analysis on the results from tandem affinity purification experiments (TAP-tag) targeted toward documented cell cycle-related proteins (six baits) and carried out in *A.thaliana* (Van Leene *et al*., 2007). The list of preys contained 218 proteins, but unfortunately only 46 of these obtained predictions by our method. However, 13 of these 46 predictions were previously confirmed as stable cell cycle bait interactors in more than one experimental repeat (Van Leene *et al*., 2007), and 9 of them had been classified as cell cycle regulators (*P*-value of 1.62E-10) (Supplementary Table S2; Supplementary Material file 4). Among the rest (an additional 33 non-confirmed interactors), 18 proteins were classified to the plant development branch (*P*-value of 2.10E-03), many of them being 'bona fide' interactors, including kinesins. We hypothesize that these kinesins may include crucial substrates of cyclin-dependent kinease/cyclin complexes, thus linking cell cycle regulation with plant development.

4 CONCLUSIONS

In this study, we applied a novel data integration scheme for modeling a broad range of biological processes and molecular functions. Our approach aims to integrate gene expression data, protein features and GO annotations in the form of interpretable If-Then rules rather than providing a new framework for supervised learning in gene function prediction. The empirical studies of Tan (Tan and Gilbert, 2003) and Chua (Chua *et al*., 2007) showed that none of the existing machine learning methods is consistently better

on all types of biological data. Thus, the choice of tool for biological data analysis should mainly be dictated by the type of research problem at hand, and to what degree interpretability is important. We decided to use the Rough Sets framework as a tool for the analysis because it is recognized as a robust and reliable technique in various biological data integration scenarios (Hvidsten *et al*., 2005; Lægreid *et al*., 2003). We showed that the information gathered in such If-Then rules can be processed and used to model and interpret biological processes and functions. Previously, Lichtenberg and collaborators (Lichtenberg *et al*., 2003) proposed that periodic, co-expressed genes encode cell cycle proteins that might share combinations of features, together providing an overview of dynamics of the cell cycle proteome. Here, we showed that this notion can be applied to any other biological process or molecular function, essentially constituting a global annotation approach. If-Then rule models dynamically connect genes with common trends in gene expression (co-regulation, inverse regulation) and various combinations of protein features (similar protein modification machinery) to their GO annotations representing crucial processes and functions. These models were statistically validated showing a high classification performance (Fig. 2). However, the major part of our study was directed toward rigorous verification of classification results. For this we used PubMed literature reports, homology information (NCBI) and recent GO annotations to confirm or negate hypotheses generated by our models. Many of the additional classifications for known genes (FP reclassifications) represented existing knowledge. For some of the unknown genes we could find an agreement between our predictions, homology information and GO annotations obtained using other computational analyses (not considered for training of our models) (Table 2). Finally, a closer look at the classification results revealed significant fractions of proteins that could be involved in the formation of complexes governed by core cell cycle proteins. The validity of these predictions was based on the protein–protein interactions extracted from the BioGRID (Breitkreutz *et al*., 2008) database, and on results from our own TAP-tag interactome experiments (Van Leene *et al*., 2007).

In our present study, the microarray datasets included gene expression measurements for leaf development and cell cycle. Thus, genes involved in these processes were, as expected, particularly well predicted using expression data. Our approach allows the use of multiple sets of expression data sources (i.e. stress response, metabolism and developmental processes) each focused to specific sets of biological processes to increase the information content supporting specific functional GO classes. We think that integrating more diverse expression data would lower the potential experimental bias in the functional predictions. An intriguing extension to the data integration procedure presented here would be to include functional genomic information, such as regulatory binding sites motifs as lately proposed in the work of Hvidsten *et al*. (2005).

ACKNOWLEDGEMENTS

The authors thank Thomas Skøt Jensen for providing processed protein feature datasets, Erick Antezana for assistance with the Gene Ontology knowledge integration and Jens Hollunder for interpreting the results of the TAP-tag experiments. Finally, Astrid Leagreid for helpful discussions.

Funding: Institute for the Promotion of Innovation through Science and technology in Flanders ('Generisch Basisonderzoek aan de Universiteiten', grant no. 20193); European Union-Human Resources and Mobility for an Early Stage Training grant (MEST-CT-2004-514632 to K.W. and A.K.); Swedish Research Council (VR) and the Swedish Governmental Agency for Innovation Systems (VINNOVA) (to T.R.H).

Conflict of Interest: none declared.

REFERENCES

Aderem,A. and Hood,L. (2001) Immunology in the post-genomic era. *Nat. Immunol.*, **2**, 373–375.

Adie,A.E. *et al*. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, **6**, 55.

Andersson,R. *et al*. (2005) A rough knowledge base system. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing,* Vol. 3642 (Edited by Science LNC). Springer, Berlin, Heidelberg, pp. 48–58.

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Boulle,M. (2005) Optimal bin number for equal frequency discretizations in supervised learning. *Int. Data Anal.*, **9**, 175–188.

Breitkreutz,B. *et al*. (2008) The BioGRID interaction database: 2008 update. *Nucleic Acids Res.*, **36**, 637–640.

Brown,M.P. *et al*. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Genetics*, **97**, 262–267.

Chen,X. *et al*. (2004) Identification of genes encoding putative nucleoporins and transport factors in the fission yeast *Schizosaccharomyces pombe*: a deletion analysis. *Yeast*, **21**, 495–509.

Chua,H.N. *et al*. (2007) An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics*, **23**, 3364–3373.

Davis,J. and Goadrich,M. (2006) The relationship between precision-recall and ROC curves. In *ICML: Proceedings of the 23rd international conference on Machine learning*. ACM, New York, NY, USA, Pittsburgh, Pennsylvania, pp. 233–240.

Dennis,J. *et al*. (2005) Markers of adenocarcinoma characteristic of the site of origin: development of a diagnostic algorithm. *Clin. Cancer Res.*, **11**, 3766–3772.

Hernandez,F. and Johnson,R.A. (1984) Selecting an extreme-value distribution and the transforming to a specified distribution. *Oper. Res.*, **32**, 715–725.

Hishigaki,H. *et al*. (2001) Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, **18**, 523–531.

Hollunder,J. *et al*. (2005) Identification and characterization of protein subcomplexes in yeast. *Proteomics*, **5**, 2082–2089.

Honeyman,M.C. *et al*. (1998) Neural network-based prediction of candidate T-cell epitopes. *Nat. Biotechnol.*, **16**, 966–969.

Hvidsten,T. *et al*. (2001) Predicting gene function from gene expressions and ontologies. *Pac. Symp. Biocomput.*, 299–310.

Hvidsten,T. *et al*. (2003) Learning rule-based models of biological process from gene expression time profiles using gene ontology. *Bioinformatics*, **19**, 1116–1123.

Hvidsten,T. *et al*. (2005) Discovering regulatory binding site modules using rule-based learning. *Genome Res.*, **15**, 856–866.

Jensen,L. *et al*. (1999) A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nat. Genet.,* **22**, 271–275.

Jensen,L. *et al*. (2002) Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.*, **316**, 1257–1265.

Jensen,L. *et al*. (2003) Functionality of system components: conservation of protein function in protein feature space. *Genome Res.*, **13**, 2444–2449.

Kohavi,R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)*, Vol. 2. Morgan Kaufmann, pp. 1137–1143.

Komorowski,J. (1999) Rough Sets—a tutorial. In Pal,S.K. and Skowron,A. (eds), *Rough-fuzzy Hybridization—A New Trend in Decision Making*. Springer, Singapore, pp. 3–98.

Komorowski,H.J. (2002) Modelling biological phenomena with rough sets. In Alpigini,J.J. *et al*., (eds), *Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing, Lecture Notes In Computer Science, Malvern, PA, USA, October 14–16*. Vol. 2475, Springer, Berlin, Heidelberg, p. 949.

Koyama,T. *et al*. (2007) TCP transcription factors control the morphology of shoot lateral organs via negative regulation of the expression of boundary-specific genes in *Arabidopsis*. *Plant Cell*, **19,** 473–484.

Lægreid,A. *et al*. (2003) Predicting Gene Ontology biological process from temporal gene expression patterns. *Genome Res.*, **13**, 965–979.

Lichtenberg,U. *et al*. (2003) Protein feature based identification of cell cycle regulated proteins in yeast. *J. Mol. Biol.*, **329**, 149–170.

Maere,S. *et al*. (2005) A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Patterson,S. (2003) Data analysis: the Achilles heel of proteomics. *Nat. Biotechnol.*, **21**, 221–222.

Pawlak,Z. (1992) Rough Sets: theoretical aspects of reasoning about data. In *Theory Decision Lib*. Vol. 9, 1st edn, Kluwer Academic Publishers, Norwell, MA, USA, pp. 1–229.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.*, **2**, 418–428.

Rosenberg,J. *et al*. (2006) Ppc89 links multiple proteins, including the septation initiation network, to the core of the fission yeast spindle-pole body. *Mol. Biol. Cell*, **17**, 3793–3805.

Rustici,G. *et al*. (2004) Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.*, **36**, 809–817.

Schena,M. *et al*. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schlitt,T. and Brazma,A. (2007) Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, **8**, s9.

Schneider,T. (2001) Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J. Clim.*, **14**, 853–871.

Skowron,A. and Suraj,Z. (1996) A parallel algorithm for real-time decision making: a Rough Set approach. *J. Int. Inform. Syst.*, **7**, 15–28.

Słowiński,K. *et al*. (2002) Application of rule induction and Rough Sets to verification of magnetic resonance diagnosis. *Fundam. Inf.*, **53**, 345–363.

Tan,A.C. and Gilbert,D. (2003) An empirical comparison of supervised machine learning techniques in bioinformatics. In *Proceedings of the First Asia Pacific Bioinformatics conference,* Vol. 19, Adelaide, Australia, Australian Computer Science Inc., Darlinghurst, Australia, pp. 219–222.

Tikunov,Y. *et al*. (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol.*, **139**, 1125–1137.

Van Leene,J. *et al*. (2007) A tandem affinity purification-based technology platform to study the cell cycle interactome in *Arabidopsis thaliana*. *Mol. Cell Proteomics*, **6**, 1226–1238.

Vaucheret,H. *et al*. (2006) AGO1 homeostasis entails coexpression of MIR168 and AGO1 and preferential stabilization of miR168 by AGO1. *Mol. Cell*, **22**, 129–136.

Vinterbo,S. and Øhrn,A. (2000) Minimal approximate hitting sets and rule templates. *Int. J. Approx. Reason*, **25**, 123–143.

Walker,M. *et al*. (1999) Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res.*, **9**, 1198–1203.

Wu,L. *et al*. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.*, **31**, 255–265.

Yeung,K. *et al*. (2003) Clustering gene expression data with repeated measurements. *Genome Biol.*, **4**, s34.