Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W445–W448***
DOI: 10.1093/nar/gkh407

ArrayProspector: a web resource of functional associations inferred from microarray expression data

Lars Juhl Jensen[1,2], Julien Lagarde[1], Christian von Mering[1,2] and Peer Bork[1,2,*]

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany and
[2]Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Strasse 10, D-13092 Berlin, Germany

Received February 14, 2004; Revised and Accepted March 29, 2004

ABSTRACT

DNA microarray experiments have provided vast amounts of data which can be used for inferring gene function. However, most methods for predicting functional associations between genes from expression data are not suited to simultaneous analysis of multiple datasets, and a comprehensive resource of coexpression-based predictions is currently lacking. Here, we present an interactive web resource of gene associations predicted by applying a novel algorithm to all expression data in the Stanford Microarray Database. The underlying pre-computed database currently contains more than 200 000 high-confidence gene associations in 12 different species sampled from a broad taxonomic range. The resource allows every association to be inspected visually and can be accessed at http://www.bork.embl.de/ArrayProspector.

INTRODUCTION

Over less than a decade, vast amounts of microarray expression data have been collected. The data have from very early on been used for inferring gene function based on unsupervised clustering of coexpressed genes (1), although it is unclear how correlated the expression profiles must be in order for function to be safely inferred. Also, if one wants to make use of expression data from many very different experiments, these cannot all be assumed to be equally important—nor will correlations necessarily hold across all microarrays. Traditional clustering methods are thus not applicable.

We instead opt for a combination of singular value decomposition and kernel density estimation to calculate a log-odds score for each pair of genes. These scores are further refined to obtain a confidence value for each interaction. Through this procedure, evidence from related arrays is combined and different arrays will contribute differently to the final score depending on how well they correlate with functional annotation. The ArrayProspector web server will be closely integrated with the next version of STRING (2), a resource for the prediction of protein interaction networks, to allow microarray-based predictions to be viewed as an additional evidence type for functional association between genes.

Usage and visualization of predictions

Users can submit queries to the ArrayProspector server using a web interface (http://www.bork.embl.de/ArrayProspector). The gene of interest can be specified using either one of a number of database identifiers or the name of the gene/protein. Alternatively, the user can submit two gene names to specifically check for association of the two.

The predicted functional associations for the query gene are shown as a clickable two-panel view, which by default shows the top 10 genes ranked by confidence score of the association (Figure 1). The left panel displays a traditional blue/yellow representation of the expression log-ratios of each gene on each array; a contrast button allows the user to adjust the color scale used. In addition to the usual features of this visualization, the spot size is used to denote the importance of each experiment for each predicted association. In cases of missing measurements of gene expression, a small white spot is shown. To start with, the arrays contributing most to the most important principal component are shown.

The right panel shows a more compact view of coexpression evidence in which gene expression has been projected onto principal components. The principal components are ranked by importance according to their total contribution to the log-odds scores of the genes shown. The view is very similar to the left panel except that each column represents a principal component and thus a combination of correlated arrays rather than a single array. Also, a different color scale is used to avoid confusion of the two panels. Clicking a column in the right panel will cause the left panel to display the arrays most important for the principal component in question. This allows the user both to quickly get an overview of the global expression patterns of a set of genes across hundreds of arrays and to look at individual arrays in more detail.

*To whom correspondence should be addressed. Tel: +49 6221 387 526; Fax: +49 6221 387 519; Email: Peer.Bork@embl-Heidelberg.De

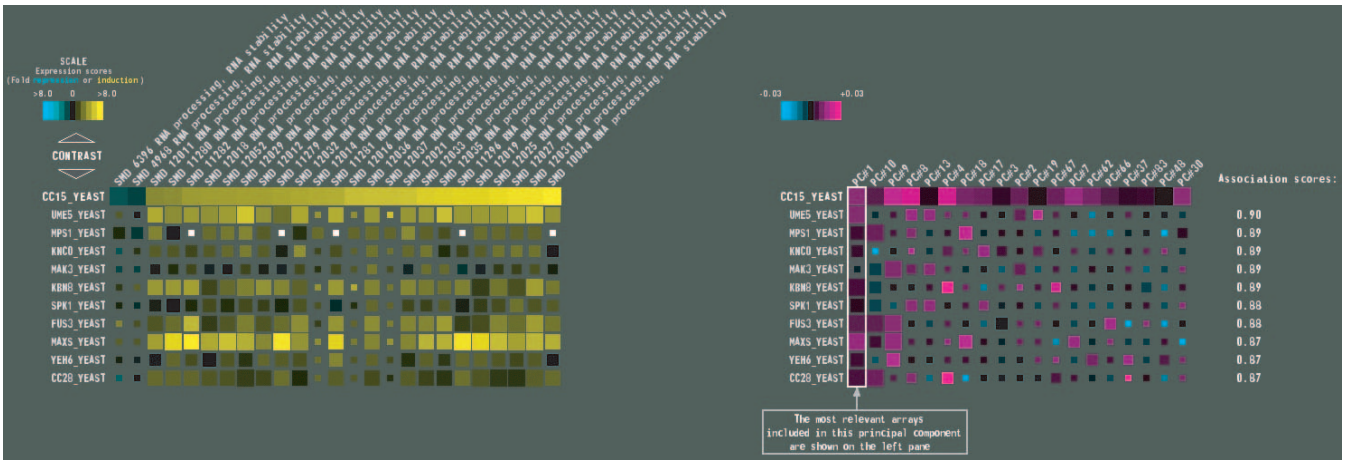

Figure 1. Sample output using CDC15 as query gene. Querying the ArrayProspector resource with the *Saccharomyces cerevisiae* cell cycle gene CDC15 correctly associates CDC15 with other cell cycle-related genes. This is remarkable considering that the CDC15 transcript level has not been suggested to vary periodically through the cell cycle (3–5). Consistent with this, the cell cycle experiments contribute little to the most important principal components for these predictions. For an explanation of how to interpret the visualization, see the section 'Usage and visualization of predictions'.

Below the two panels, a table summarizes the most important information about the arrays currently shown in the left panel, namely the experiment description from the Standford Microarray Database (SMD) and contact information for the person who performed the experiment.

DATA TRANSFORMATION, SCORING AND BENCHMARKING

The spot intensities of the two channels (Cy3 and Cy5) on each microarray were individually normalized using the Qspline method (6) with a log-normal distribution as target (M = ln 1000, S = ln 1000). For the majority of the arrays, namely those where the relative location of the spots was provided by SMD, the channels were further normalized to correct for spatial biases using a Gaussian smoother with $\sigma = 0.8$ (6). After adding a regularization background intensity of 100 to the normalized intensities, a log-ratio was calculated for each gene on each spotted array. This value was semi-empirically chosen to make the spread of log-ratios independent of the spot intensities.

Currently, arrays have been included from six eukaryotes, namely *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae*, as well as six prokaryotes, namely *Bacillus subtilis*, *Campylobacter jejuni*, *Escherichia coli*, *Helicobacter pylori*, *Salmonella typhi* and *Vibrio cholerae*. For each, the log-ratios for all arrays were combined into a matrix, assigning a log-ratio of zero in the case of missing values. We used our gene synonyms resource for solving the problem that the same gene is not always referred to by the same name/identifier on all arrays (http://www.bork.embl.de/synonyms/).

One problem when analyzing microarray data is that different arrays may be strongly correlated, e.g. replicate arrays, adjacent time points in time series or similar experiments performed by different laboratories. Singular value decomposition, a powerful method for dealing with such correlations, was used to obtain a new basis (the principal components) for which the covariance matrix of the data is diagonal. The principal components can be interpreted in a biological context by studying the loading factors and the classification of the arrays contributing most to each principal component. As the log-ratios of each array are already guaranteed to be centered at 0 due to the normalization procedure described above, the missing values have a minimal influence on the components. For the subsequent analysis, each gene was represented by its projection onto the principal component basis.

Raw log-odds scores for a functional association between any two genes were calculated with reference to KEGG maps (7). Two genes are thus considered to be functionally related if their protein products co-occur in at least one KEGG map. Similarly, non-related genes were defined as genes with KEGG assignment but no shared assignment.

For each separate principal component, a two-dimensional Gaussian kernel density estimate was calculated for pairs of related genes [$f_{\text{related}}(x_1, x_2)$]. Similarly, a one-dimensional Gaussian kernel density estimate was calculated for all genes ($f_{\text{all}}(x)$), allowing a log-odds score for the projection of a gene pair onto a principal component to be calculated as follows:

$$\text{logodds}(x_1, x_2) = \log \frac{f_{\text{related}}(x_1, x_2)}{f_{\text{all}}(x_1) \cdot f_{\text{all}}(x_2)}.$$

A total log-odds score for a given pair of genes is calculated as the sum of log-odds scores from the first N principal components. To include only components with a good signal-to-noise ratio, for each species N was determined by visually inspecting a logarithmic plot of the singular values as a function of component number.

Using raw log-odds scores causes certain genes, in particular those encoding cell cycle proteins, to have high log-odds scores to hundreds of other genes, many of which are not functionally related. We therefore down-weighted the log-odds score between two genes by the number of higher-scoring links for the most highly connected of the two genes. This way, we penalize links to the most highly connected genes, which improves the overall accuracy of the predicted associations considerably.

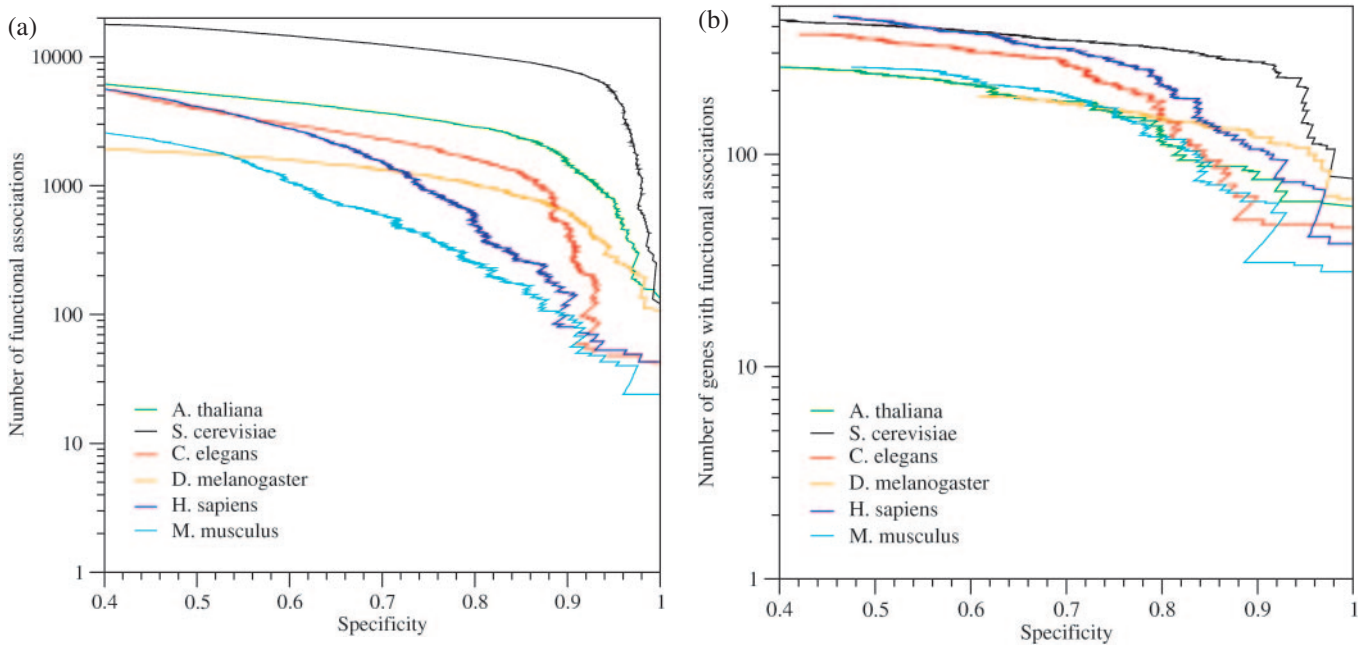

Figure 2. Performance of the method. (**A**) shows for each species the number of predicted functional associations known to be correct versus specificity (also known as accuracy, i.e. the fraction of predictions being correct). Predictors are considered to be better when their curves are higher and further to the right. To show that the functional associations are not due to a small number of highly connected genes encoding large complexes (e.g. the ribosome), a similar plot was constructed showing the number of genes for which the function can be correctly predicted based on the highest-scoring link to another gene of known function (**B**). The performance for a species does not correlate trivially with the number of arrays available, as is exemplified by the yeast predictor (based on 591 arrays) outperforming the human one (2413 arrays).

For each species, we benchmarked the final predicted associations against co-occurrence on KEGG maps. Figure 2 shows the number and accuracy of our predictions for six model eukaryotes. At a specificity of 80%, more than 1000 true positive associations are predicted for 4 of the 6 eukaryotes and more than 10 000 true positives are predicted for *S.cerevisiae* alone. Assuming that this performance is representative of the performance functionally uncharacterized proteins, ArrayProspector contains in excess of 200 000 correct functional associations.

Calibration curves for converting the down-weighted log-odds scores to probabilistic confidence scores were obtained by fitting sigmoid functions to plots of specificity versus score. This calibration strategy is consistent with the one used for genomic context evidence in the STRING server (2).

DATA SOURCES AND STORAGE

All spotted microarray data available from SMD were downloaded for *A.thaliana* (591 arrays), *C.elegans* (291), *D.melanogaster* (170), *H.sapiens* (2413), *M.musculus* (113), *S.cerevisiae* (718) and six bacteria (465) (8).

We use a relational database system (PostgreSQL) for storing all data required for the web interface, namely expression log-ratios, their projections onto principal components, raw log-odds scores (both the total and the contribution from each component), confidence scores as well as text descriptions of both genes and arrays. Storing these many intermediary values allows the user to quickly navigate the large quantity of data.

CONCLUDING REMARKS

We make available a resource, ArrayProspector, that enables biologists to mine spotted array data for functional associations in a number of organisms. Arrays have been re-normalized to correct for both non-linear intensity biases and spatial effects. Subsequently, we have linked expression data for each gene across arrays (using our synonyms resource) and applied a novel algorithm for predicting functional associations between genes. In addition to making available these pre-computed associations, ArrayProspector enables the user to manually investigate the source of each prediction.

ACKNOWLEDGEMENTS

The authors wish to thank Christoffer Workman for help with normalization of expression data. L.J.J. is funded by the Bundesministerium für Forschung und Bildung, BMBF-01-GG-9817.

REFERENCES

1. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.
2. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
3. Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

4. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *S. cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
5. Zhao,L.P., Prentice,R. and Breeden,L. (2001) Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl Acad. Sci., USA*, **98**, 5631–5636.
6. Workman,C., Jensen,L.J., Jarmer,H., Berka,R., Gautier,L., Saxild,H.-H., Nielsen,C., Brunak,S. and Knudsen,S. (2002) A new non-linear normalization method to reduce variability in DNA microarray experiments. *Genome Biol.*, **3**, research0048.1–research0048.16.
7. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
8. Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B. and Hebert,J. (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.*, **31**, 94–96.