Pacific Symposium on Biocomputing 7:247-258 (2002)

EVIDENCE FOR SEQUENCE-INDEPENDENT EVOLUTIONARY TRACES IN GENOMICS DATA

W. VOLKMUTH, N. ALEXANDROV
Ceres Inc., 3007 Malibu Canyon Road,
Malibu, CA 90265, USA

Sequence conservation during evolution is the foundation for the functional classification of the ennormous number of new protein sequences being discovered in the current era of genome sequencing. Conventional methods to detect homologous proteins are not always able to distinguish between true homologs and false positive hits in the twilight zone of sequence similarity. Several different approaches have been proposed to improve the sensitivity of these methods. Among the most successful are sequence profiles, multi-linked alignment, and threading. However, evolution might offer up other clues about a protein's ancestry that are sequence independent. Here we report the discovery of two such traces of evolution that could potentially be used to help infer the fold of a protein and hence improve the ability to predict the biochemical function. The first such evolutionary trace is a conservation of fold along the genome, i.e. nearby genes tend to share a fold more often than expected by chance alone—a not unexpected observation, but one which holds true even when no pair of genes being examined share appreciable homology. The second such evolutionary trace is, surprisingly, present in expression data: genes that are correlated in expression are more apt to share a fold than two randomly chosen genes. This result is surprising because correlations in expression have previously only been considered useful for determining *biological* function (e.g. what pathway a particular gene fits into), yet the observed fold enrichment in the expression data permits us to say something about *biochemical* function since fold corresponds strongly with biochemical function. Again, the fold enrichment observed in the expression data is apparent even when no pair of genes being examined share appreciable homology.

1 Introduction

The evolutionary tool of sequence duplication and subsequent sequence divergence is the means used by Nature to create new biochemical and biological function. The full tapestry on which these evolutionary events are played out is becoming available thanks to the multitude of successful whole-genome sequencing projects. Having the entire tapestry allows us to improve our understanding of evolution, and that improved understanding will in turn lead to better prediction of both the biological and biochemical functions of experimentally uncharacterized proteins.

Many experimentally uncharacterized proteins can be identified through sequence similarity to other proteins that *have* been characterized. However, the identification of distant homologs is a fundamental problem in modern computational biology with enormous potential for practical applications. The vast amount of genome sequencing is dramatically increasing the importance of the problem. The function of about 30% of *Arabidopsis thaliana* genes cannot be predicted by sequence similarity search methods[1-5]. About 40% of the identified genes in human chromosomes 21 and 22 do not have detectable homology to known genes[6,7]. Therefore even a small improvement in our ability to identify distant homologs can help us make functional predictions for a large number of newly discovered genes.

To build proteins, Nature draws on what appears to be a more or less fixed repertoire of *folds* developed during the course of evolution. A fold is a structural motif used as a building block in proteins. Various fold classifications exist, for our purposes we use that of the Structural Classification of Proteins (SCOP)[8]. Structural similarity has been recognized as a solid argument for evolutionary relationship between proteins[9,10] and hence, prediction of their function. Success in protein structure prediction by fold recognition is limited by our ability to identify homologous protein with known three-dimensional structure. Progress in fold recognition can be assessed at the regular CASP conferences (http://predictioncenter.llnl.gov/). The recent CASP4 meeting clearly showed that an expert, knowledge-based approach is superior to a purely computational, fully automated approach. The advantage of experts is that they use not only biochemical or biophysical information on protein sequence and structural properties, but also knowledge of protein function derived from biochemical experiments. In support of the utility of expert knowledge, it has in fact already been demonstrated that even something as simple as a key-word comparison of protein descriptions in the database increases the accuracy of fold recognition programs[11]. Clearly then, incorporation of additional data will improve the accuracy of fold prediction

What other kinds of data could be incorporated? The mechanism of duplication often leads to neighboring genes that share a common ancestor. It seems natural then to expect a fold enrichment among neighboring genes, and that is in fact one of the conclusions we report here for both *Arabidopsis* and *Saccharomyces cerevisiae*, consistent with an analysis on a subset of the yeast data analyzed here[12]. However, we can make a stronger statement. Not only do neighboring genes share fold, the distance between pairs of genes in the genome conveys information about structural homology *over and above* the information from the sequence comparison alone.

We will distinguish *biochemical* function of a protein from the gene's *biological* function, where by biological function we mean what the organism accomplishes with the gene and other co-expressed genes, for example a signal transduction cascade. Just as sequence conservation during evolution can be used to infer structural homology and biochemical function, one might wonder if Nature has left a trace of structural information during its evolution of biological (as opposed to biochemical) function.

If a structural trace can be found in the evolution of pathways, that trace could also be helpful in improving the prediction of protein fold. We looked for and found that trace using correlations in expression data, a method used for inferring biological function since the invention of the Northern blot but only recently performed on the genome-wide scale necessary to detect such a trace. As in the case of the fold enrichment seen between nearby genes on the genome, the fold enrichment between pairs of co-expressed genes conveys structural information over and above that from sequence similarity alone. Hence, *biological* function of a gene says something about the gene's corresponding protein fold, and via the fold something about the *biochemical* function of that protein.

In the following, then, we show that both physical distance on the genome and correlation in expression show traces of evolution manifested in structural homology. Those weak signals could in principle be used to improve prediction of structural homology in the twilight zone of sequence similarity.

2 Methods

2.1 Genomes

To investigate the evolutionary trace of fold enrichment along the genome, we used the genomes of the model organisms *Saccharomyces cerevisiae*[13] and *Arabidopsis thaliana*[1-5]. The yeast genome consists of 16 chromosomes, with 6,310 identified ORFs, available from the *Saccharomyces* Genome Database (SGD) at http://genome-www.stanford.edu/Saccharomyces/. The smallest chromosome, chromosome I, is ~0.23 Megabases and has 107 ORFs. The largest chromosome, chromosome IV, is ~1.53 Mb and encodes 819 ORFs. The recently finished Arabidopsis genome consists of five chromosomes, with 25,498 genes predicted. The smallest chromosome is chromosome IV and is ~17.5 Mb in length, containing 3,825 protein encoding genes. The largest chromosome is chromosome I, approximately 29.1 Mb in length, with 6,543 genes. We downloaded Arabidopsis genes from the NCBI web site (http://www.ncbi.nlm.nih.gov).

Only chromosomes II and IV were available as one contig at the time we made our analysis, so we used only 7,852 protein-coding genes from these two chromosomes in our analysis of fold enrichment in the genome neighborhood.

2.2 Microarray Expression Data

Yeast and arabidopsis expression data were downloaded from the Stanford Microarray Database[14]. A subset of non-biological experiments (e.g. assessing the performance of microarrays) was excluded from our analysis. The resulting dataset contained expression data from 345 yeast and 201 Arabidopsis microarray experiments.

The similarity in expression across all experiments between a pair of genes was measured using the Spearman rank correlation coefficient on the normalized ratios[15]. The Spearman r was chosen because it is a robust statistic that will capture any monotonic relationship between a pair of variables, as opposed to the commonly used Pearson correlation coefficient which is suitable for detecting linear relationships between pairs of variables. Missing data points were handled by pairwise deletion of observations from the Spearman *r* calculation and any pair of genes having fewer than 10 experiments in common were ignored. No special attempt was made to account for the redundancy due to experimental replicates or similarities in subsets of experiments. We note, however, that for the purposes of a global correlation analysis at least, it would be more desirable to have a larger number of distinct, diverse experiments than to have experimental replicates since the correlation coefficient implicitly takes inherent experimental variation into account.

Spurious significant correlations might be introduced between a pair of genes that are not actually co-expressed if the two genes are sufficiently similar that cross-hybridization occurs, where "sufficiently similar" is roughly taken to be in the neighborhood of 80% over 50 nucleotides[16-18]. No large scale, systematic experimental study of cross-hybridization on microarrays has been done, so we assessed the degree of cross-hybridization indirectly as follows. Pairwise similarity was measured using Wash-U BLASTN, version 2.0 with M=2 and all other parameters set to their defaults. We compared the overall distribution of correlation coefficients between pairs of genes to the distribution between non-identical chip features showing similarity of >=85% over >=50 nt. There is a clear shift towards one for the similar sequences as seen in Figure 1. The distribution for clones having 70%-85%, >=50 nt similarity is shifted towards 1 as well, but the shift is less pronounced (data not shown). To be conservative and minimize the possibility of cross-hybridization affecting our results, we discarded any chip feature having a sequence with an HSP showing 70% or greater similarity over at least 50 nt to some other gene in the genome. Applying this approach for yeast is straightforward since the microarray features are PCR fragments of the ORFs, the complete sequence of the features is known[19] and there are essentially no introns in yeast. In the case of Arabidopsis, the full sequence of the clones serving as source material for the microarrays is not available, so we mapped ESTs from the clones to the cDNAs from the annotation of the Arabidopsis genome[1-5] and assumed that the entire cDNA sequence was present in the microarray feature, then screened against all other cDNAs in Arabidopsis. While conservative, our approach cannot guarantee complete exclusion of features that might cross-hyb because the genomic annotations often lack full UTRs or have other errors.

The cross-hybridization filtering resulted in 4,280 yeast features and 3,011 Arabidopsis features. The smaller number of filtered features for Arabidopsis was primarily a consequence of feature redundancy on the chip (more than one clone for a given cDNA) and a higher amount of gene duplication in Arabidopsis. A correlation and cluster analysis was performed with the biological sensibility of results conforming to those from the literature, though the Arabidopsis results were less compelling[20-22].

2.3 Fold Assignment

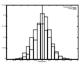
For each gene we assigned fold(s) according to the SCOP-1.55 classification of protein structures[8]. Assignment was done by WU-Blastp[23] search against the Astral database of non-redundant SCOP domains at the 95% identity level [24]. We considered all matches with a P-value < .001. At that level of significance approximately 2% of our assignments are wrong[9]. One protein may consist of more than one domain, in those cases multiple folds were assigned to the corresponding gene. Out of 6,310 yeast genes, we assigned folds to 1,839 genes (29%). Out of 27,469 Arabidopsis genes from the TIGR gene index, we assigned folds to 9,147 genes (33%). The distribution of different SCOP folds in the two genomes is shown in Figure 2, with the most frequent folds summarized in Table 1. This is consistent with the most frequent folds in other organisms[25]. More advanced methods of fold assignments, e.g. PSI-BLAST, the profile-profile technique, and threading, increase the coverage, but overall do not change the statistical observations.

2.4 Non-redundant Set of Proteins

Since our intent is to determine if we can detect distant homologs, we created a non-redundant set of proteins from the overall set of proteins that had folds assigned to them. To create the non-redundant set, the following procedure was applied: for each protein, beginning with the longest, all shorter proteins were removed from the list if they matched the first protein with a P-value < 1.0e-3.

2.5 Fold Enrichment Along the Genome

The relative enrichment of folds along the genome was defined as the ratio of the probability of finding the same fold between pairs of genes a given distance apart in the genome to the probability of finding the same fold between two randomly selected pairs of genes in the genome. The ratio is therefore a function of the distance in nucleotides between gene pairs. At a given distance, a ratio greater than one implies that more similar folds are occurring than one would expect if folds were distributed randomly over that distance. A ratio of one indicates that the folds are distributed randomly.



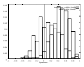


Figure 1. a) Histogram of Spearman *r* distribution for clones showing >=85% similarity over >=50 nt. b) Histogram of Spearman *r* distribution overall for clones included in analysis, i.e. those that show no similarity to any other clone at the 70%, 50 nt level. It should be pointed out that Figure 1a strongly suggests that the Arabidopsis data is of inferior quality to the Yeast data, see comment in the Results section below.

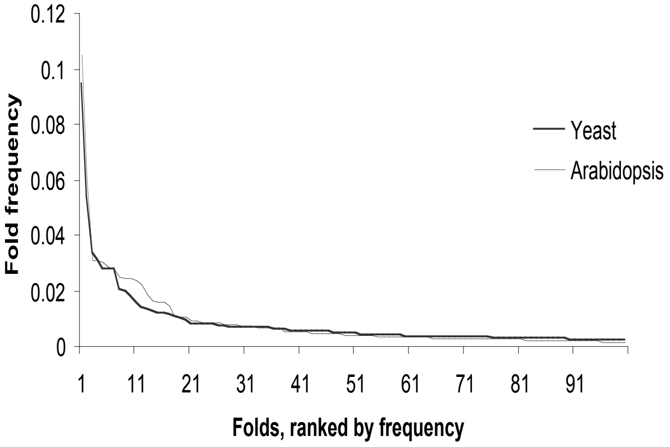


Figure 2. Fold distributions in yeast and Arabidopsis genomes.

SCOP_1.55 fold	Description	Yeast		Arabidopsis	
		rank	frequency	rank	frequency
c.37	P-loop containing nucleotide triphosphate hydrolases	1	0.095	2	0.059
d.144	Protein kinase-like (PK-like)	2	0.054	1	0.105
c.1	TIM beta/alpha-barrel	3	0.033	5	0.030
b.69	7-bladed beta-propeller	4	0.030	10	0.024
a.118	alpha-alpha superhelix	5	0.028	7	0.027
c.2	NAD(P)-binding Rossmann-fold domains	6	0.028	8	0.025
d.58	Ferredoxin-like	7	0.027	3	0.031
g.38	Zn2/Cys6 DNA-binding domain	8	0.021	>279	0
f.2	Membrane all-alpha	9	0.020	18	0.011
c.55	Ribonuclease H-like motif	10	0.018	21	0.009
a.4	DNA/RNA-binding 3-helical bundle	11	0.016	4	0.031
g.44	RING finger domain, C3HC4	22	0.008	6	0.028
a.104	Cytochrome P450	149	0.001	9	0.024

Table1. Ten most frequent folds in the yeast and Arabidopsis genomes. Seven folds belong to the ten most frequent folds in both genomes.

2.6 Fold Enrichment for Genes with Similar Patterns of Expression

The relative enrichment for co-expressed genes was defined by calculating the ratio of the probability of having a matching fold at or above a given level of correlation coefficient to that expected by randomly choosing pairs of genes. Fold enrichment for correlated genes is therefore a function of the Spearman *r*, with a ratio greater than one indicating that a pair of correlated genes is more likely to share fold than expected from chance. Error bars were estimated from counting error.

3 Results

3.1 Fold Enrichment Along the Genome

One of the most frequent evolutionary events is gene duplication, with the Arabidopsis genome being especially rich in tandemly repeated genes[1-5]. Therefore it is not surprising to see enrichment of homologous genes in the chromosomal neighborhood for both organisms. The effect can, however, still be observed even for the set of non-redundant proteins (Figure 3)

3.2 Fold Enrichment for Genes with Similar Patterns of Expression

Figure 4a shows a plot of the fold enrichment in yeast as a function of *r*. Figure 4b shows the corresponding plot for Arabidopsis. Both organisms show enrichment that is significantly elevated from the baseline of 1.0, although the difference is more pronounced for yeast. The enrichment is maintained even when redundant proteins are removed.

In the case of Arabidopsis there is only a weak signal at best, yet we suspect that it is in fact real. We believe that it is weak as compared to yeast at least in part because of dataset size and in part because of overall data quality. A power calculation shows that for the set of yeast data, we can reliably detect correlations down to ~0.4 (significance 0.01, power 80%, Bonferonni correction, power calculation assumed Pearson *r* rather than Spearman *r*). For Arabidopsis the threshold is roughly 0.5. This weaker detection ability is confounded by relatively poor quality data for Arabidopsis. The poorer quality Arabidopsis data means that the threshold for a biologically significant correlation is higher than for the yeast data. The difference in quality is evident from the much smaller overall shift towards a correlation of 1 in the distribution of Figure 1a). With more and better quality data, the peak should presumably become cleaner.

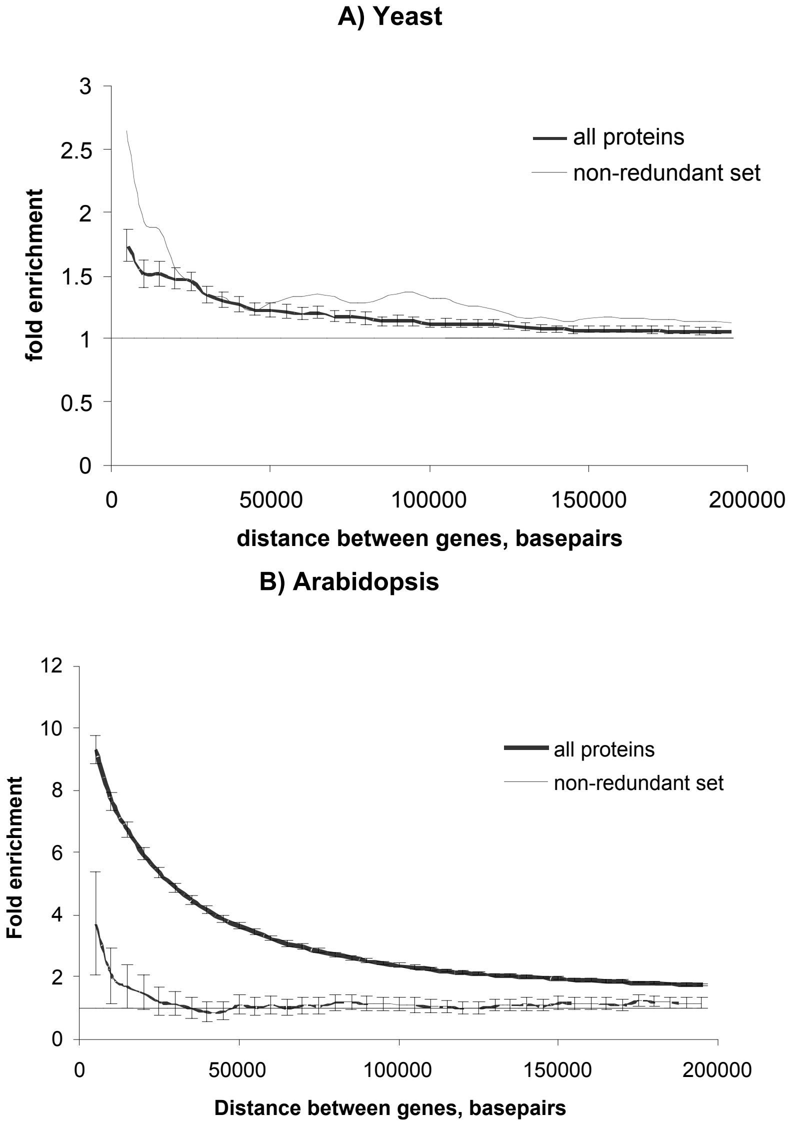


Figure 3. We examined two sets of genes for fold enrichment along the genome. The first set was all genes that had a fold assignment. The second set was a subset of the first consisting of genes whose proteins showed no significant homology to one another. For each set, fold enrichment was measured as the ratio of the frequency of same fold for genes within *d*

nucleotides of each other to the frequency of the same fold in randomly selected genes. The enrichment ratio is then plotted as a function of distance *d*.

4 Summary and Conclusions

Folds among nearby genes in the genome and among co-expressed genes are enriched relative to that expected by chance alone. We examined the distributions of enriched folds and were unable to explain the enrichment through a bias in folds in either case. From this we conclude that the enrichment we see is a more-or-less general feature of folds in organisms; with the large amount of data becoming available for worm[26], human and other organisms we will be able to confirm or rule out this speculation in the near future. Assuming we are correct one should in principle be able to incorporate such information into fold prediction for proteins whose fold is unknown. We are currently evaluating approaches to accomplish that goal.

The mechanism behind the enrichment of folds along the genome seems clear. Gene duplications lead to pairs of genes with similar ancestors, and even after substantial divergence results in no *sequence* homology between nearby genes on the genome, there is still a remnant of structural similarity. It is that remnant which accounts for the observed enrichment.

The mechanism behind the enrichment of folds among co-expressed genes is less clear. One hypothesis is that during the course of evolution of (e.g.) a particular metabolic pathway, a newly duplicated gene is created. For the sake of illustration let us say that the duplicated gene is an enzyme. Since that newly duplicated gene, which includes the promoter region of the original gene, is now redundant, one of two things must happen. Either one of the duplicated genes will disappear or the pair will diverge apart in sequence, with one retaining the original function (by function here we mean both biochemical function as well as biological role), and the other taking on a new function. Since both originally operate on the same substrate there is a structural constraint to how the pair diverge in sequence, and this constraint tends to cause the fold to be maintained.

The extent to which the behavior described above actually explains how Nature evolves pathways remains to be demonstrated. It is interesting to note, however, that we made a correct, blinded prediction of protein fold for a recent target in the CASP4 competition, using in part exactly the above reasoning. The target in question was pectin methylesterase, which is co-expressed with its metabolic pathway neighbor pectate lyase[27]. Both enzymes share exactly the same SCOP fold, the single-stranded right-handed beta-helix.

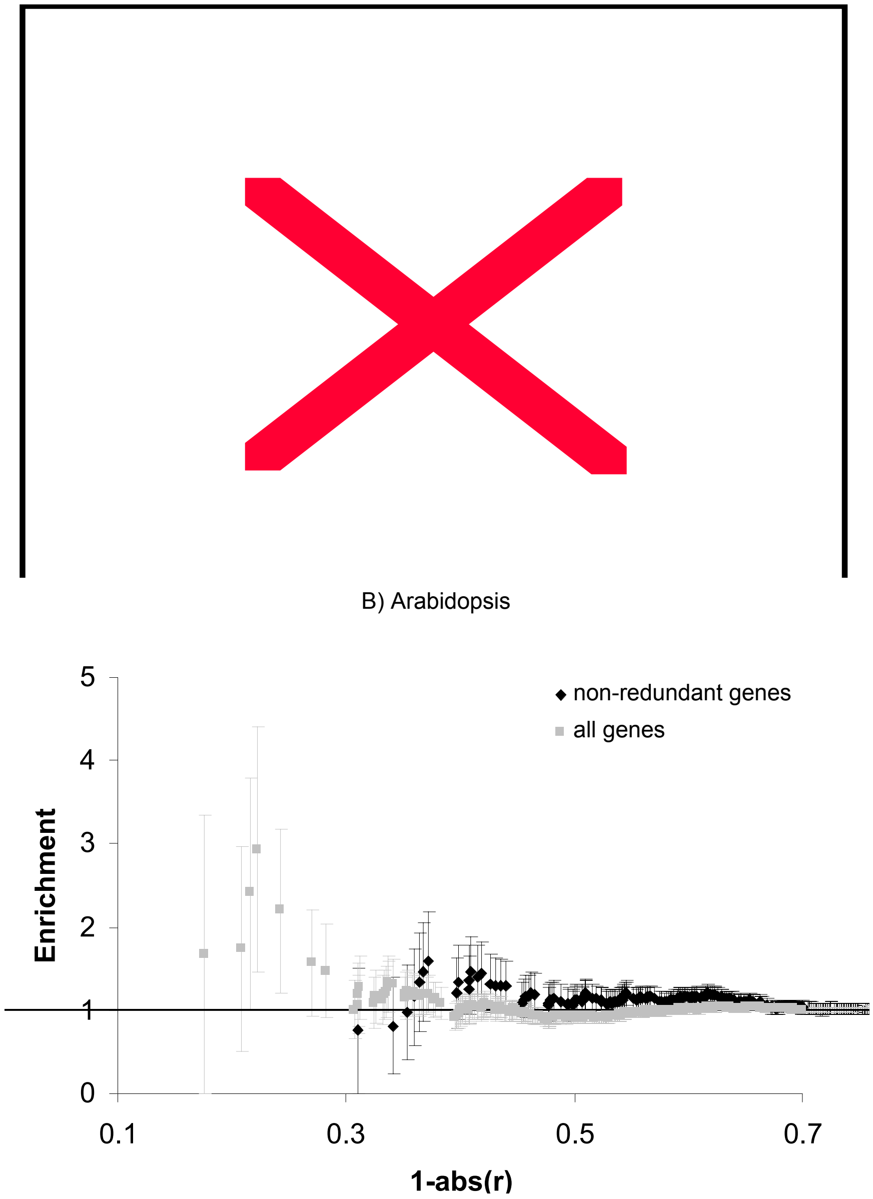


Figure 4. We examined two sets of genes for fold enrichment among co-expressed genes. The first set was all genes that had a fold assignment and showed no significant sequence homology to other genes at the nt level, see methods description for details on selection. The second set was a subset of the first consisting of genes whose proteins showed no significant homology to one another. For each set, fold enrichment was measured as the ratio of the

frequency of same fold for genes correlated at *r* or better to each other, relative to the frequency of the same fold in randomly selected genes. The enrichment ratio is then plotted as a function of *r*. Error bars are counting statistics only.

References

1. Theologis, A. *et al. Nature* **408**, 816-20 (2000).
2. Lin, X. *et al. Nature* **402**, 761-8 (1999).
3. Salanoubat, M. *et al. Nature* **408**, 820-2 (2000).
4. Mayer, K. *et al. Nature* **402**, 769-77 (1999).
5. Tabata, S. *et al. Nature* **408**, 823-6 (2000).
6. Dunham, I. *et al. Nature* **402**, 489-95. (1999).
7. Hattori, M. *et al. Nature* **405**, 311-9. (2000).
8. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. *J Mol Biol* **247**, 536-40. (1995).
9. Brenner, S.E., Chothia, C. & Hubbard, T.J. *Proc Natl Acad Sci U S A* **95**, 6073-8. (1998).
10. Park, J. *et al. J Mol Biol* **284**, 1201-10. (1998).
11. MacCallum, R.M., Kelley, L.A. & Sternberg, M.J. *Bioinformatics* **16**, 125-9. (2000).
12. Cohen, B.A., Mitra, R.D., Hughes, J.D. & Church, G.M. *Nat Genet* **26**, 183-6. (2000).
13. Goffeau, A. *et al. Science* **274**, 546, 563-7. (1996).
14. Sherlock, G. *et al. Nucleic Acids Res* **29**, 152-5 (2001).
15. Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.R. Numerical Recipes in C. 639-640 (Cambridge University Press, Cambridge Cb2 2RU, UK, 1992).
16. Kane, M.D. *et al. Nucleic Acids Res* **28**, 4552-7 (2000).
17. Girke, T. *et al. Plant Physiol* **124**, 1570-81 (2000).
18. Xu, W. *et al. Gene* **272**, 61-74. (2001).
19. DeRisi, J.L., Iyer, V.R. & Brown, P.O. *Science* **278**, 680-6 (1997).
20. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. *Proc Natl Acad Sci U S A* **95**, 14863-8 (1998).
21. Heyer, L.J., Kruglyak, S. & Yooseph, S. *Genome Res* **9**, 1106-15 (1999).
22. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. *Nature* **402**, 83-6 (1999).
23. Gish, W. (1994).
24. Brenner, S.E., Koehl, P. & Levitt, M. *Nucleic Acids Res* **28**, 254-6. (2000).
25. Gerstein, M., Lin, J. & Hegyi, H. *Pac Symp Biocomput* , 30-41. (2000).
26. Kim, S.K. *et al. Science* **293**, 2087-92. (2001).
27. Tierny, Y., Bechet, M., Joncquiert, J.C., Dubourguier, H.C. & Guillaume, J.B. *J Appl Bacteriol* **76**, 592-602. (1994).