Bioinformatics and the discovery of gene function

GEORG CASARI, ANTOINE DE DARUVAR, CHRIS SANDER* AND REINHARD SCHNEIDER
casari@embl-heidelberg.de
daruvar@embl-heidelberg.de
sander@embl-ebi.ac.uk
schneider@embl-heidelberg.de

EUROPEAN MOLECULAR BIOLOGY LABORATORY, EMBL-HD, D-69012, HEIDELBERG, GERMANY.
*EUROPEAN BIOINFORMATICS INSTITUTE, EMBL-EBI, HINXTON HALL, CAMBRIDGE, UK CB10 1RQ.

Scientific history was made in completing the yeast genome sequence, yet its 13 Mb are a mere starting point. Two challenges loom large: to decipher the function of all genes and to describe the workings of the eukaryotic cell in full molecular detail. A combination of experimental and theoretical approaches will be brought to bear on these challenges. What will be next in yeast genome analysis from the point of view of bioinformatics?

Current information status

Functional knowledge about yeast genes is already more advanced than one might have expected at the outset of the sequencing effort. For an amazing 65% of the approximately 6000 protein-encoding genes, we already have some functional information. For about 30% of the total, according to B. Dujon (this issue), functional knowledge is the result of direct experiment, but for a large fraction, about 35% of the total, functional information was derived by homology transfer. Homology transfer of information exploits the evolutionary continuity in protein function and structure over very long time spans, apparent in our world as the presence of similar genes in viruses, bacteria and eukaryotes. Technically, homology transfer becomes possible because of the two pillars of current bioinformatics: databases that capture the experimental knowledge about gene function in different organisms and search algorithms that permit the detailed comparison of a new gene with all available database sequences. Whenever sequence similarity is detected at a level that clearly indicates functional or structural homology, information can be transferred from a gene of known function in one species to one of unknown function in another species (or in the same species). In some cases, the transferred information exquisitely describes the detailed biochemical and/or cellular function of a new gene, for example, that of the yeast open reading frame (ORF) YCR14c on chromosome III (SWISS-PROT Accession No. P25615) as the functional cousin of the mammalian DNA polymerase β (Ref. 1). In other cases, the power of prediction is very limited, because of strong functional divergence or because the homology is limited to a sequence fragment; an example is the prediction of nucleic acid binding properties based on the presence of a zinc finger motif.

In many cases, the prediction of gene function by homology transfer can be easily achieved using standard database search tools. However, trained experts are needed to achieve a high level, in terms of quantity and quality, of derived functional information. The unexpectedly high cumulative value of 65% of functionally annotated protein sequences (Fig. 1) is the result of applying an advanced software system, called GeneQuiz (Ref. 2), that encapsulates expert knowledge, combines a variety of analysis tools and relies on daily updates of the latest database information.

The approach to 100%

How much is left to discover about gene function in yeast? The good news is that some information about gene function is available for an

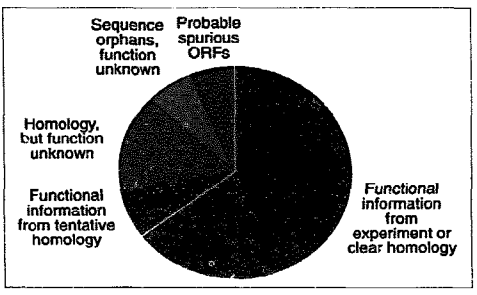

FIGURE 1. How much information is available to date about yeast proteins? For as many as 65% of protein genes more-or-less detailed 'functional information' is available, either from direct experiment and/or from information transfer via 'clear' evolutionary relationships (homology). For more than 7% (contained in the 65%) a three-dimensional model ('3D model') is available, most built by homology modeling. Another 7% have 'tentative homology' to proteins with functional information (of which we expect about 2–3% to survive scrutiny). 'Homology without functional information' is clearly detected for 14% of all proteins (these are called orphan pairs or families by B. Dujon, this issue), and only 12% are 'sequence orphans of unknown function' (called single orphans by B. Dujon), that is, have no sequence relative in the databases, of which 6–7% can be spurious (i.e. are not expressed as messenger nor translated into proteins). The numbers were derived in a GeneQuiz analysis of all yeast protein genes available in March 1996, about 85% of the total number of yeast proteins. More than 6000 genes (partially redundant, depending on the state of the March 1996 publicly available data, with 100% identities removed) were compared with 185688 protein sequences and 438305 ESTs on a 64 processor computer configuration. Differences in the amount of derived functional information and the number of clear homologies relative to the accompanying article by B. Dujon are probably due to (1) differences in the yeast and search databases used, and (2) differences in processing the results of sequence analysis methods. The main future bioinformatics challenges are to increase the quality and completeness of the deduced information, and to develop methods to discover more evolutionary relationships. The derived functional information and 3D models are available (http://www.sander.embl-heidelberg.de/genequiz).

 0168-9525/96/$15.00
PII: S0168-9525(96)30057-7

estimated 70% of all yeast protein genes (deduced from 65%, as above, corrected for 6–7% questionable ORFs (B. Dujon, this issue), none of which are expected to have clear homology to a database protein). The bad news is that this information is incomplete, to varying degrees, and so a sizeable fraction of these 70% deserve further study, experimentally and with bioinformatics tools, as do the remaining 30% of genes of completely unknown function. On second thoughts, the news is actually not so bad for those planning major efforts in functional analysis (S. Oliver, this issue)!

Interestingly, the number of sequence orphans (those without a sequence family of homologs) is already very small – about 10% of the total (number from Fig. 1, corrected for questionable ORFs and tentative homologies) – and will soon reach zero, as a result of a large influx of sequence data in the near future, especially expressed sequence tags (ESTs), as well as worm and human genomic sequences. This is excellent news for bioinformatics, as similarity searches using multiple sequence alignments[3] can reach significantly further into the twilight zone of homology, with an increased probability of linking otherwise unavailable functional information.

Whatever one's point of view, there is a long way to go before we have a fairly complete description of the function of all yeast genes. Bioinformatics is faced with a number of technical challenges in the process: to improve the methods for homology modeling in three dimensions [currently, models have been built for 7% of yeast genes with homology to proteins of known three-dimensional structure (G. Vriend and R. Schneider, unpublished)]; to improve the quality of homology transfer by detecting errors in databases, by developing quantitative methods for assessing functional divergence based on sequence divergence, and by more accurately extracting functional information from sequence families; to go beyond the current level of homology detection, by improved methods of database searches (e.g. those based on profiles derived from sequence families or known three-dimensional structures[3]); and to develop further the direct prediction of function from sequence information, or at least the prediction of functional class (e.g. A. Goffeau's data about transmembrane segments in yeast proteins: see poster in this issue).

In meeting these challenges, bioinformatics will help reduce the experimental effort needed to discover gene function. Each bit of information gained, each extra gene function suggested by homology will either avoid redundant experimental effort or, ideally, focus the experimental strategy on to a smaller set of possibilities. Conversely, each new experimental characterization of gene function will, through the databases, spread information to an increasing number of homologs. In this way, we can look forward to considerable synergy between bioinformatics and experimental functional analysis.

Far beyond 100%

Suppose the projects for functional analysis of yeast proteins were already completed in July 1996 and that we had an accurate functional assignment for each protein. Yes, we would marvel at the detailed part list of the complicated machinery of the eukaryotic cell. But we would also ask many new questions. Can we have a complete list of all specific protein-pair interactions and higher complexes? What is the complete set of cycles of interaction in the metabolic, regulatory and developmental pathways? How can functional concepts of molecular genetics and cell biology best be mapped on to the properties of participating molecules? What are the stability properties of the cell when parts are damaged or added? Which molecules facilitate evolutionary adaptation as external conditions change? And many more.

In the long run, bioinformatics will contribute to the emergence of a new quantitative and predictive molecular biology of the cell. The challenge will be to combine the knowledge of all parts and interactions with new ideas, and to develop methods to simulate, on a computer, the detailed behavior of cells. One way or the other, the genome sequence will revolutionize cell biology.

Acknowledgements

The GeneQuiz software system for large scale sequence analysis was developed in collaboration with M.A. Andrade, P. Bork, C. Ouzounis, M. Scharf, J. Tamames and A. Valencia. The order of authors is alphabetical. We are grateful for the use of SRS by T. Etzold, PHD by B. Rost, and WHATIF by G. Vriend. The March 1996 analysis of yeast ORFs was performed at Silicon Graphics, Supercomputing Technology Centre, Cortaillod, Switzerland, with the support of P. Bremer, M. Schlenkrich and colleagues.

References

1 Bork, P. *et al.* (1992) *Protein Sci.* 1, 1677–1690
2 Casari, G. *et al.* (1995) *Nature* 376, 647–648
3 Holm, L. and Sander, C. (1995) *Trends Biochem. Sci.* 20, 345–347