Overview of the yeast genome

H. W. Mewes, K. Albermann, M. Bähr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S. G. Oliver[1], F. Pfeiffer & A. Zollner

Max-Planck-Institut für Biochemie, D-82152 Martinsried, Germany
[1]University of Manchester Institute of Science And Technology (UMIST), Sackville Street, Manchester M60 1QD, UK

The collaboration of more than 600 scientists from over 100 laboratories to sequence the *Saccharomyces cerevisiae* genome was the largest decentralised experiment in modern molecular biology and resulted in a unique data resource representing the first complete set of genes from a eukaryotic organism. 12 million bases were sequenced in a truly international effort involving European, US, Canadian and Japanese laboratories. While the yeast genome represents only a small fraction of the information in today's public sequence databases, the complete, ordered and non-redundant sequence provides an invaluable resource for the detailed analysis of cellular gene function and genome architecture. In terms of throughput, completeness and information content, yeast has always been the lead eukaryotic organism in genomics; it is still the largest genome to be completely sequenced.

The *Yeast Genome Directory* presents the basic features of this sequence: the arrangement of the 6,000 genes on 16 chromosomes; a summary of the function of the encoded proteins; and a view of the genome's architecture, based on an exhaustive intra-genomic sequence comparison[1]. The complete yeast sequence can be retrieved from a number of public databases, as well as from specialized World-Wide Web sites, which provide sophisticated query interfaces (see Box). These data are maintained and updated continuously. Although the form of the genome directory shown in this printed volume must present static information, its intention is to document the interpretation of the yeast genome shortly after its completion. To present the genome in a printable form, we have had to include a very limited selection of fact-oriented data. The *Yeast Genome Directory* cannot answer the question "What's in the yeast genome?" exhaustively, but summarizes what is known to be in it.

The sequence The final sequence was assembled from roughly 300,000 independent sequence reads, with error rates from 0.5 to 1%, resulting in an estimated error rate of the final sequence of less than 3 errors in 10,000 bases (0.03%). For the European Union part of the sequencing effort, a central database and informatics used to assemble, verify and analyse contiguous sequences submitted by the participating laboratories was provided by the Martinsrieder Institut für Protein Sequenzen (MIPS). The complete sequence was made available to the public on 24 April 1996. The first map represents the open reading frames (ORFs) of each of the successive chromosomes, at a uniform scale of 5 mm per kilobase. ORFs are named according to the location of the gene, using the convention Y (for yeast) followed by a letter denoting the chromosome (A for I, B for II, and so forth), a letter denoting the arm (R or L), a three-digit code ordering the ORFs from the centromere, and a letter denoting the coding strand (w or c).

Duplications The availability of the complete sequence of yeast allows us for the first time to examine the evolution of a eukaryotic species in a truly comprehensive manner. The footprints that indicate the evolutionary path taken by the yeast genome may be recognized by internal similarities between distinct regions of the present-day genome. Our approach to the inspection of these relationships is based on an all-against-all comparison of the genomic sequence data, applying local sequence alignments. Investigation of the yeast genome involved more than 24,000 blocks of 500 nucleotides. For each block, the six-frame translation into protein sequences was also generated, to allow for the concurrent comparison of DNA and amino-acid sequences[2].

Once an all-against-all matching of the yeast genome had been accomplished, duplication patterns within the genome could be investigated in a systematic way. The frequent, collinear block duplications found by our method seem to be an important consequence of the evolutionary development of *S. cerevisiae*. We have systematically inspected the genome for clusters of genes that have been produced by local duplication events. This involved evaluating the parameters that describe a cluster: it's size; the degree of similarity of the duplicated ORFs; and the compactness of the cluster. Scanning this parameter-space at a sensitivity just above noise level (50% identity on the DNA level), we found that a window size of 25 kb and an enforced compactness of 10% of the coding region in that window generated an optimal representation of gene clusters. These criteria allowed us to identify 53 regions of clustered gene duplications, not including the well-known high level of similarity in the telomeric and subtelomeric regions. The second map shows the set of all collinear clusters of genes in the genome as a two-page fold-out. The significant number of gene duplications in yeast must reflect an evolutionarily successful strategy: gene duplications allow for evolutionary modifications in one of the copies without disturbing possibly vital functions of the other.

Gene function The computational analysis of the yeast genome is a challenging task[3]. The scientific assessment of raw sequence data aims to connect genetic entities to biological knowledge, either by computational analysis or by linking sequence-deduced information to other experimental evidence (such as a genetic locus). Sophisticated data modelling is required to reflect the correct relationship between the sequence and any associated information and to allow for the consistent integration of complex, heterogeneous biological data. For example, the precursor of the ADP/ATP carrier protein AAC2p is represented five times in the EMBL Nucleic Acid Data Library, although four of the coding sequences reported are identical. The current nucleic acid databases (EBI/NCBI/DDBJ) contain 2,678 entries covering the yeast genome with a high degree of redundancy and inconsistency. Thus the traditional model of collecting

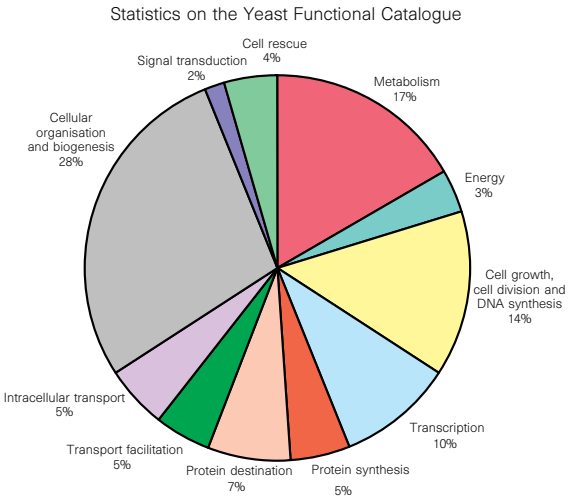

Figure 1 This shows the relative number of ORFs assigned to individual categories in the Gazetteer. There are eleven functional categories. Only proteins with a known function, or similarity or strong similarity to known proteins were assigned to one of the categories (similarities were measured by FASTA scores). In total, 3167 ORFs were assigned to at least one category. A single ORF can be assigned to more than one category.

database entries and presenting them as individual reports does not seem to be suitable for coping with the data analysis of complete genomes. Like the printed page, the traditional layout of the sequence databases is static, incorporating information from scientific annotation at the time of publication. Additional information extracted from the literature must be translated manually into the formal framework of a database entry. The knowledge of the functional properties of an uncharacterized gene may be published independently, and family members found in other organisms may allow for characterization by homology. Suitable models must thus be developed to cope with the data analysis of a complete genome, and to enable integrated views of the genomic sequence and associated information.

Only 43.3 % of the yeast genes are currently classified as 'functionally characterized', having experimentally well-investigated properties, being members of well-defined protein families, or displaying strong homology to proteins with known biochemical functions. The systematic functional analysis of these genes, currently in progress[4], will identify the functions of many of the uncharacterized 'orphans'. Various experimental methods, including improved '*in silico*' analysis, will also increase dramatically the information content of the biological databases. Previously, individual attempts have been made to analyse specific chromosomes systematically by sequence analysis[5,6], and the results of an automated software system, 'Genecrunch', to 'crunch' the complete yeast genome have been offered as an Internet service (http://genecrunch.sgi.com).

To provide information on the biochemical and physiological context of protein function, we have compiled a gazetteer listing all the ORFs that can be related to well-understood functions. Similarities between biological sequences were measured by FASTA scores[7]: a FASTA score between 100 and 200 was defined as a 'weak similarity'; between 200 and a third of the self-score (the score of the protein when aligned with itself) for the protein was defined as a 'similarity'; and higher than a third of the self-score was defined as a 'strong similarity'. Weak similarities were not considered. In addition to the similarity scores, we have used pattern data[8], including experimental data from the literature combined with genetic data, to characterize ORFs.

Each entry in the gazetteer lists the ORF name (defined as above), the gene name (if any), and the name or a short description of the protein. Entries are divided into 11 categories representing the cellular function of the individual ORFs (such as metabolism, energy or transcription), and each category is divided into sub-categories[9]. The yeast genome encodes about 5,800 proteins, less than half of which are 'known' in the sense that they have been genetically and biochemically well characterized. For about 20% of the remaining proteins, the experimental data are heterogeneous and provide only some indication of their functions *in vivo*. The remaining 38% either show similarities to other uncharacterized proteins or show no similarities at all. Defined categories provoke redundant entries, such as the classification of multifunctional enzymes. As gene names are used elsewhere in a very inconsistent way, we have used the name from the *Saccharomyces* genome database (see Box) register whenever possible. Gene names that are used twice for different ORF names are written in brackets.

Sequences of common evolutionary origin (homologous *sensu stricto*) reflect their relatedness by sequence similarities, and the organization of related primary, secondary and tertiary structures into groups ('families') remains the most powerful principle in sequence data analysis. We have used this principle to cluster proteins into families and superfamilies, following a previous scheme[10], allowing us to cope with the many taxonomic complications inherent in protein evolution.

The World-Wide Web site of the Martinsrieder Institut für Protein Sequenzen yeast resource combines information generated by automated procedures with the results of systematic analysis by yeast researchers. Users can: (1) visualize chromosomes and selected regions to inspect genetic elements, such as ORFs, Ty's, tRNAs etc.; (2) receive detailed information on a yeast gene by searching with accession numbers, systematic codes, or gene names; (3) browse yeast genes according to their functional classification; (4) search for human homologues; (5) obtain information on functional properties; (6) download nucleic-acid or protein sequence data; (7) inspect up-to-date sequence homologies and alignments (FASTA database); (8) browse the family and superfamily classification of yeast proteins; (9) search the yeast genome interactively for sequence patterns and sequence similarities; and (10) inspect the yeast genome for gene redundancy.

As a complement to the printed information in the *Yeast Genome Directory*, a CD-ROM compiled by Martinsrieder Institut für Protein Sequenzen, is available on request to subscribers of *Nature* and *Science* for the exploration of the yeast genome on a local, network-independent installation. The basic functionality of the retrieval software and the databases incorporated is equivalent to our World-Wide Web resource. The CD-ROM can be installed on Windows95, WindowsNT and Power-Macintosh operating systems, and is accompanied by a detailed description of its functionality, installation procedures and system requirements. □

1. Goffeau, A. *et al. Science* 274, 546–567 (1996).
2. Heumann, K., Harris, C. & Mewes, H. W. in *Proceedings of the Fourth Int. Conf. Intelligent Systems for Molecular Biology* 98–108(AAAI Press, Menlo Park, 1996). (1996).
3. Casari, G. *et al. Nature* 376, 647–648 (1995).
4. Oliver, S. G. *Trends Genet.* 12, 241–242 (1996).
5. Bork, P. *et al. Protein Sci.* 1, 1677–1690 (1992).
6. Koonin, E. V. *et al. EMBO J.* 13, 493–503 (1994).
7. Pearson, B. & Lipman, D. *Proc. Natl Acad Sci. USA* 85, 2444–2448 (1988).
8. Bairoch A., Bucher, P. & Hoffman K. *Nucleic Acids Res.* 24, 189–196 (1996).
9. Riley, M. *Microbiol. Rev.* 57, 862–869 (1993).
10. Dayhoff, M. O. *Fed. Proc.* 35, 2132–2138 (1976).

Acknowledgements. We thank A. Goffeau for support and encouragement; B. Jasny for suggestions; and Netscape Communications Inc. for allowing the limited distribution of Netscape 3.01. This work was supported the European Commission, the Max-Planck-Society, the Forschungszentrum für Umwelt und Gesundheit, and the Yeast Industry Platform.

Correspondence and requests for materials should be addressed to H.W.M. (e-mail: mewes@mips. embnet.org).

BOX Useful World-Wide Web addresses

Yeast databases

Munich Information Centre for Protein Sequences (MIPS)
http://www.mips.biochem.mpg.de/mips/yeast/

Yeast Protein Database (YPD)
http://quest7.proteome.com/YPDhome.html

Saccharomyces Genome Database (SGD)
http://genome-www.stanford.edu/Saccharomyces/

Specialized Yeast Databases

Sacch3D – Structural information for yeast proteins
http://genome-www.stanford.edu/Sacch3D/

Yeast Gene Duplications
http://acer.gen.tcd.ie/~khwolfe/yeast/topmenu.html

Related human disease genes (NIH XREFdb)
http://www.ncbi.nlm.nih.gov/XREFdb/

Genetic and physical maps (hyperlinked to biological information)
http://genome-www.stanford.edu/cgi-bin/SGD/pgMAP/pgMap

NIH yeast information page
http://www.ncbi.nlm.nih.gov/Yeast/budding.html

Schizosaccharomyces pombe
http://www2.bio.uva.nl/pombe/

Candida albicans information
http://alces.med.umn.edu/Candida.html

Common DNA and protein databases

European Bioinformatics Institute (EBI)
http://www.ebi.ac.uk/services/services.html

National Center for Biotechnology Information (NCBI)
http://www.ncbi.nlm.nih.gov/

DNA Data Bank of Japan (DDBJ)
http://www.ddbj.nig.ac.jp/

PIR-International
http://www.mips.biochem.mpg.de

SwissProt
http://expasy.hcuge.ch/sprot/sprot-top.html