Acta Biochimica Polonica

Vol. 45 No. 3/1998

627–643

QUARTERLY

Review

Sequencing and functional analysis of the yeast genome[o]

Marek Zagulski[1], Christopher J. Herbert[2] and Joanna Rytka[1✉]

[1]*Institute of Biochemistry and Biophysics, Polish Academy of Sciences, A. Pawińskiego 5a, 02-106 Warszawa, Poland;* [2]*Centre de Génétique Moleculaire, Laboratoire propre du CNRS associaté a l'Université Pierre et Marie Curie, 91198 Gif-sur-Yvette, France*

Received: 15 July, 1998

Key words: yeast, *Saccharomyces cerevisiae*, genomics, genome sequence, gene function

The genome of the yeast *Saccharomyces cerevisiae* was sequenced by an international consortium of laboratories from Europe, Canada, the U.S.A. and Japan. This project is now finished and the complete sequence of the first eukaryotic genome was released to the public data bases in April 1996. An overview and preliminary analysis of the entire genome sequence was presented in a special issue of *Nature* in May 1997, entitled "The yeast genome directory". At its origin the Yeast Genome Sequencing Project provoked much debate and controversy; however, the final results obtained and the insights this has given us into the organisation and content of a eukaryotic genome have more than justified the expectations of the supporters of the project. The importance of genomic sequencing and analysis, especially of model organisms, is now widely accepted and this has resulted in the birth of the new science of genomics (Botstein & Cherry, 1997, *Proc. Natl. Acad. Sci. U.S.A.* 94, 5506). The information from gene and protein sequences ultimately lead to functional description of all genes. The main strategies describing possible ways to analyse the function of new genes that have been identified by systematic sequencing of *Saccharomyces cerevisiae* genome are described.

THE UNIQUE POSITION OF *SACCHAROMYCES CEREVISIAE* AMONG MODEL ORGANISMS

For decades, the yeast *S. cerevisiae* has had a prominent position as a model, single-celled, eukaryotic organism. It shares the overall cellular organisation and metabolic processes of higher eukaryotes, but is more easily manipulated. Because of the ease of culture, significant amounts of material can be obtained for biochemical studies. Also the well developed

[o]This work was supported by the Polish State Committee for Scientific Research, grant No. 6PO4A 05014 and the Franco-Polonais Plant Biotechnology Center.

[✉]Corresponding author: Tel: (48 22) 658 4701, e-mail: Rytka@slc.ibb.waw.pl

Abbreviations: ARS, autonomously replicating sequence; 5FOA, 5-fluoroorotic acid; GFP, green fluorescent protein; ORF, open reading frame.

genetic systems of yeast, both classical and molecular, facilitate the dissection of molecular interactions. One of the most powerful advantages of the yeast system is the very efficient homologous recombination and/or gene conversion which enables genes to be manipulated *in vitro* and then reinserted into their normal chromosomal location. One of the benefits of the earlier genetic and molecular studies of *S. cerevisiae* was the availability of both genetic and physical maps which provided a wealth of information on the organisation of yeast nuclear and mitochondrial genomes. Also many structural *cis*-acting elements of the genome, essential for chromosome function, such as replication origins (ARS), centromeres and telomeres had been identified. The haploid nuclear genome of *S. cerevisiae* was estimated to be 13.5 ± 3.0 Mb organised in 16 chromosomes. The examination of global transcription by R-loop electron microscopy (Kaback *et al.*, 1979) and the transcription maps of chromosomes III and I (Yoshikawa & Isono, 1990; Barton & Kaback, 1994) predicted that about 80% of the genome is transcribed and it contains 5000–6000 genes with few introns or repeat sequences. The compactness of the yeast genome and its high information content (about 70% of the sequence are **O**pen **R**eading **F**rames, ORFs) were both strong arguments in favour of the sequencing project.

The first part of the Yeast Genome Sequencing Project, the sequencing of chromosome III of *S. cerevisiae* was launched in 1989 within the framework of the European Communities (EC) research programme BAP (Biotechnology Action Programme). Chromosome III was chosen for the pilot project for several reasons; the chromosome is relatively small, 315 kb with 29 known loci, also ordered lambda clone and plasmid libraries already existed and were provided by M. Olson (Washington, University of St. Louis, U.S.A.) and C. Newlon (New Jersey Medical School, U.S.A.). The DNA libraries were divided between a consortium of 35 European laboratories each being responsible for part of the chromosome III sequence (Słonimski & Brouillet, 1993). Before the sequencing of chromosome III was finished independent programmes to sequence other yeast chromosomes were started in other parts of the world. The complete sequence of the genome of *Saccharomyces cerevisiae* strain S288C was the result of 7 years collaboration between more than 600 scientists from over 100 laboratories in Europe, Canada, the U.S.A. and Japan. The final sequence was assembled from roughly 300000 independent sequence reads, with error rates from 0.5% to 1%, resulting in an estimated error rate in the final sequence of 0.03%.

THE FINAL RESULT OF THE YEAST GENOME PROJECT

The analysis of the total nucleotide sequence revealed 12057500 bp of chromosomal DNA, excluding repeated sequences, for the standard *S. cerevisiae* strain S288C.

The yeast genome contains several regions with extensive repeated sequences. These repeats have not been sequenced in their entirety but at least one and normally two copies have been included in the final data. The total of 969000 bp of missing repeated sequence consists of: about 100 copies of rDNA (chr XII) each about 9137 bp, 2 copies of *ENA2* (chr IV) each about 3 885 bp, about 10 copies of the *CUP1* (chr VIII) each about 1998 bp, about 1–3 copies of the Y' element each 6700 bp (chr IV telomere R), about 1–2 copies of the Y' element each 6700 bp (chr XII telomere R), about 750 bp of the telomeric sequence of chr VI. Therefore in total, the nuclear genome of *S. cerevisiae* strain S288C contains about 13026500 bp.

The *S. cerevisiae* genome displays significant redundancy, 55 duplicated chromosomal regions (or "blocks") have been identified, which together occupy about 50% of the whole genome. Clusters of genes with conserved gene order and transcriptional orientation

(with insertions and/or deletions), are on different chromosomes, or located in the same chromosome in tandem or in inverted orientation. The large, duplicated DNA segments in the subtelomeric regions of several chromosomes may indicate that a continuous exchange of genetic material occurs at these sites.

For most purposes, the ORFs deduced from the sequence are given a lower size limit of 100 amino acids; however, it should be remembered that shorter ORFs have been identified experimentally. Setting a lower limit of 100 amino acids results in the prediction of about 6200 ORFs in the entire yeast genome. On the basis of a computational analysis of the genome together with experimental data, all yeast ORFs have been placed into six classes depending on their similarity to other proteins of known, or unknown function (see Fig. 1).

The yeast genome project divides naturally into two parts: the first is a structural analysis and the second, a functional analysis. With the completion of the sequence of the yeast genome the first part of the project came to an end. The interpretation and exploitation of the structural data is the major challenge for the yeast community in the years to come. This functional analysis of the genome sequence is far more complex than the sequencing project and many different equally valid approaches may be used.

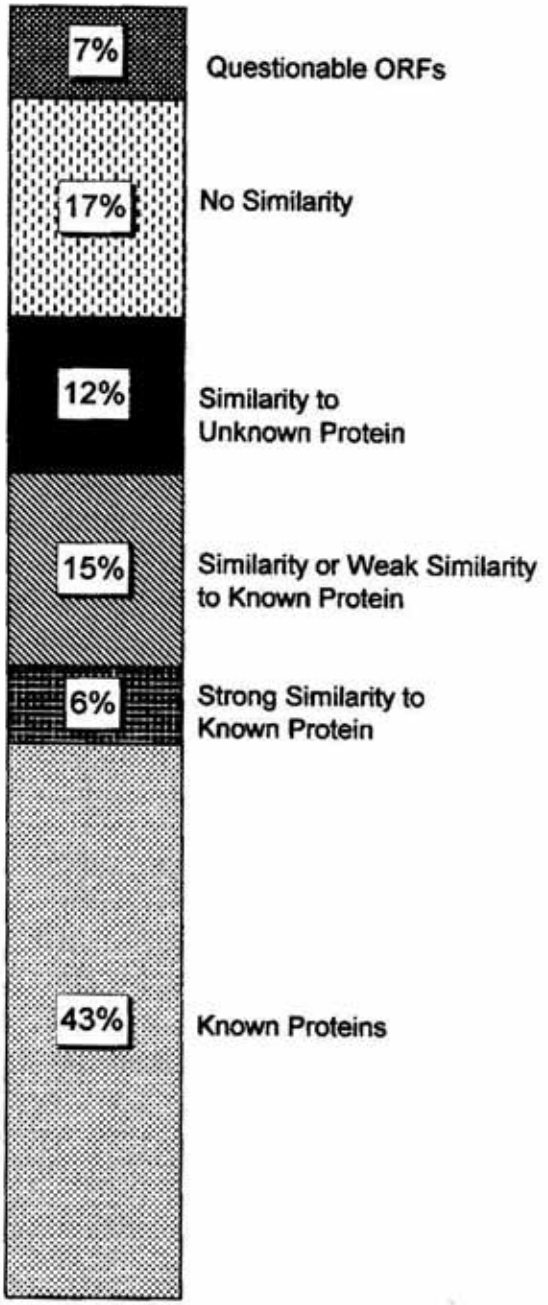

Figure 1. Classification of yeast ORFs according to their similarity to other proteins.

Data taken from MIPS server: http://speedy.mips.biochem.mpg.de/mips/yeast/class.htlm.

Similarities have been measured by FASTA scores. A questionable ORFs are defined by a combination of the following attributes: low codon biasI value, partial overlap with a longer or known ORF, no similarity to other ORFs.

FUNCTIONAL ANALYSIS OF THE GENOME OF *S. CEREVISIAE*

The second phase of the yeast genome project financed by the EEC is called EUROFAN (EUROpean Functional Analysis Network) and was launched in January 1996. The general task of the EUROFAN project is to elucidate the biological function of 1000 novel *S. cerevisiae* genes. In the course of this a large database of gene functions and a genetic archive and stock centre of yeast strains con-

taining specific deletion mutations, plasmids carrying individual genes and disruption cassettes will be created. Eventually this resource will be made available to the general scientific community. EUROFAN involves 144 European laboratories from universities, research institutes and industry and is organised in three levels:

- ***service consortia*** responsible for scientific co-ordination, finance, informatics, genetic archive and stock centre, collaboration with industry;
- ***research consortia*** creating specific deletant strains and their corresponding plasmids and also performing basic phenotypic tests which should allow the individual genes to be passed to the next, more specific level of investigation performed in „nodes";
- ***specific functional analysis nodes*** are groups of laboratories with expertise in specific areas of yeast biology (cell cycle, DNA synthesis, protein synthesis etc.). They will use the deletion strains and the phenotypic information provided by the consortia and to start to more precisely define the function of particular ORFs.

An alternative, but not mutually exclusive approach has been presented by Mark Johnston from the Washington University Medical School in St. Louis (U.S.A.). He suggested that the expertise and organisational skills of the scientists involved in the genome sequencing project should be employed to rapidly produce a complete set of 6000 yeast strains, each with a single ORF deleted. Subsequently this complete set of deletion mutants would be freely distributed to yeast laboratories (possibly 1000 world-wide) with the request that they perform their "favourite test" on each mutant. This could rapidly generate a lot of information and would save individual workers the effort and expense of disrupting genes, also it might stimulate investigators to pursue new approaches to genetic analysis.

A refinement to this approach was developed in the laboratory of Ron W. Davis at Stanford University (U.S.A.) (Shoemaker *et al.*, 1996). This method incorporates unique 20 bp tags into each deletion strain when it is constructed. Thus, each strain carries its characteristic tag serving as a mutant identifier. This method allows many mutants to be grown together, for example in competition experiments, because they can be distinguished from each other by hybridisation to the oligonucleotide tag.

BIOINFORMATICS IN THE FUNCTIONAL ANALYSIS OF GENOMES

Bioinformatics has been defined as "The use of computers in solving information problems in the life sciences..., in genome projects, informatics includes the development of methods to search databases quickly, to analyse DNA sequence information, and to predict protein structure from DNA sequence data". As the amount of sequence data available is growing rapidly, the necessity to develop bioinformatic resources is self-evident.

Most of the useful genomic data such as genetic maps, physical maps, DNA and protein sequences, are available on the World Wide Web and can be maintained only in electronic form. These data are of very little use in print. As a consequence of the rapidly increasing volume of information and the number of people involved in biological data analysis, new generations of computers and the constant development of software is essential. The data on organisation of the yeast chromosomes and analysis of ORFs and their predicted protein products presented in the papers describing the results of the complete sequence of individual chromosomes and the entire yeast genome are good examples of the use of computational analysis.

Of the numerous classes of computing methods one of the most informative in the analysis of gene function is homology transfer. It exploits the evolutionary conservation in pro-

tein function and structure. Several programmes using different algorithms permit the detailed comparison of a new gene with all available database sequences. If a predicted protein sequence, the product of newly discovered ORF of unknown function, clearly shows homology to a protein of known function from another species the transferred information indicates the function of the new gene. The power of the prediction is limited and in most cases we can only get a general description of the biochemical function of the predicted protein such as DNA binding protein, protein kinase, helicase etc. However, each piece of information obtained by homology search, suggesting protein properties reduces the experimental effort needed to discover the function of a new gene. The results of the systematic *in silico* analysis performed on the yeast genome sequence (membership of well-defined protein families, strong homology to a protein with a known function, presence of motifs, prediction of 3D structure) compiled with the sequence and the available experimental data in the literature are on the Internet Service as an integrated system (http://genecrunch.sgi.com).

A global computational analysis of the yeast genome allowed 3167 ORFs out of 6275 to be assigned to one of eleven functional categories (Table 1). Similarities were measured by FASTA scores. In addition to the similarity scores, the experimental data from the literature combined with genetic data were used. In total 3167 ORFs were assigned to at least one category. A single ORF can be assigned to more than one category.

Although an informatic analysis can yield much information concerning a new gene or protein sequence, it is important not to loose sight of the difference between biochemical activity and biological function. For example, an informatic analysis can with a high degree of confidence identify a protein kinase. However, this typical activity strongly suggests that the protein participates in a signal transduction pathway, but at present the informatic analysis cannot predict which signal transduction pathway the kinase participates in.

Table 1. The assigment of predicted yeast proteins to functional categories*

Protein function	Percent of proteins assigned to given function
Protein synthesis	5
Transcription	10
Cell growth, cell division, DNA synthesis	14
Metabolism	17
Energy	3
Cell rescue	4
Signal transduction	2
Cellular organisation and biogenesis	28
Intracellular transport	5
Transport facilitation	5
Protein destination	7

*The predicted proteins were classified on the basis of similarities measured by FASTA scores combined with biochemical and genetic data from the literature. In total 3167 ORFs were assigned to at least one category (data from Mewes *et al.*, 1997).

In general, there are two categories of experimental approach in elucidating the biological functions of newly discovered genes: the large scale genome-wide analysis, which is aimed at obtaining a little information about a lot of genes, or detailed analysis focusing on a single gene or small group of interacting genes.

GENOME-WIDE FUNCTIONAL ANALYSIS

These approaches are aimed at providing a basic functional classification of the genes rather than detailed information concerning individual genes.

The generation and analysis of *yeast::lacZ* fusion genes

A very ingenious method of large scale analysis combining gene expression, gene inactivation and protein localisation was developed by Michael Synder and colleagues (Burns *et al.*, 1994). The yeast genomic library was constructed in *E. coli* and the library was mutagenised using a mini *Tn3::LEU2* transposon containing the *LacZ*-coding sequence without a promoter or initiator ATG codon. The mutagenised yeast DNA was isolated and used to transform a diploid yeast strain. In this way random *LacZ* insertions were made throughout the yeast genome. If the *LacZ* insertion is in an ORF and is in frame with the ORF, then a β-galactosidase fusion protein will be produced when the gene is expressed. The identification of genes expressed under a given growth condition is determined by measuring the β-galactosidase activity of independent transformants using X-gal. The subcellular localisation of the different fusion proteins can sometimes be determined by indirect fluorescence using rabbit anti-β-gal antibodies. To identify fusion genes of interest, the DNA adjacent to the *LacZ* insertion is cloned and sequenced.

As there are about 6000 ORFs in yeast genome it is necessary to analyse 30000 in frame fusions to have a 98% chance of having a fusion in each gene. The authors screened 25% of the yeast genome and the results presented suggest that more than 80% of yeast genes are expressed during vegetative growth, this is in good agreement with other estimates (Yoshikawa & Isono, 1990; 1991; Richard *et al.* 1997). The study also found 93–135 genes whose expression is induced during meiosis; however, about 60% of these genes are dispensable for meiosis. From the data on localisation of fusion proteins it appeared that at least 3% of gene products are localised in the nucleus and about 1% are localised in the mitochondria. This is a very flexible system, and once the grids of transformants have been set up they can easily be tested in a large variety of conditions.

Genomic foot-printing

Another system based on transposon mutagenesis, called Genomic Foot-printing, has been developed in the Davis laboratory in Stanford (U.S.A.). These are essentially population competition experiments, where the presence of individual mutated genes in a population is followed by PCR. This can rapidly gather a lot of basic phenotypic information and many parts of the experiments are suitable to automation. The disadvantage of this system is that although it generates a lot of information, it does not create any biological material that can be exploited in subsequent experiments. This method has already been applied to all the genes of chromosome V (Smith *et al.*, 1996).

Large scale phenotypic analysis

In the search for gene function Slonimski and his co-workers have developed a large-scale screening for the identification of gene function *via* the systematic phenotypic analysis of yeast mutants carrying a targeted deletion of a single gene (Rieger *et al.*, 1997). In these studies a large number of unusual growth conditions and the sensitivity to many antibiotics and other inhibitors are tested. Although in most cases the observed phenotypes do not define the function of the gene, they do allow a classification. Also when very clear phenotypes are observed, they can be used as the starting point for more sophisticated genetic analyses.

Characterisation of transcriptome

This new term denotes the evaluation of gene expression by the analysis of the global pattern of transcription. To characterise the yeast transcriptome Velculescu *et al.* (1997) used the SAGE technology (Serial Analysis of

Gene Expression) (Velculescu *et al.*, 1995). The authors analysed more than 60000 of transcripts and detected 4665 different genes, encoding proteins larger than 100 amino acids which were expressed at between 0.3–200 copies per cell. The results indicate an even distribution of transcripts among chromosomes, a low expression of regions within 10 kb of telomeres and allow the detection of highly expressed genes which mostly corresponded to enzymes involved in energy metabolism and protein synthesis. Interestingly, 2684 of the genes transcribed at detectable levels represented uncharacterised ORFs. The results presented by Velculescu *et al.* (1997) are in good agreement with the known transcript levels of well characterised genes.

The comparison of transcriptomes from cells grown in a variety of conditions will provide insight into genes that are specifically required in a given condition. This technique is complementary to the various forms of large scale mRNA hybridisation that have been developed in many laboratories (DeRisi *et al.*, 1997).

Proteome analysis

The term proteome denotes the complete set of proteins that a given organism is capable of synthesizing (Wilkins *et al.*, 1996). The completion of the yeast genome sequence allows the prediction of the proteome of the *S. cerevisiae* cell. With the development of new techniques, the analysis of gene expression at the level of their final product, i.e. proteins, became realistic. It appeared already from preliminary work on chromosome III and from the data obtained by Burns *et al.* (1994) that a clear-cut simple phenotype is only associated with the inactivation of a minority of genes. The large-scale determination of individual protein levels may solve that problem. Two dimensional (2-D) gel electrophoresis allows the resolution of 1000–3000 soluble yeast proteins; in the first dimension they are separated by charge and in the second by size. It is hoped that changes in the characteristic pattern of spots in the 2-D gel analysis of a disruptant or transformant over-expressing the gene under study, will allow the analysis of qualitative and quantitative changes in the levels of the products of related genes. At present this approach is limited by several factors: only one third to a half of yeast proteins are visualised on the gel, high molecular mass and membrane proteins cannot be resolved on the gel. The use of the matrix-assisted laser desorption/ionization (MALDI) mass spectrometry, should rapidly overcome the problem of identification (Khan, 1995; Boucherie *et al.*, 1995; Shevchenko *et al.*, 1996).

THE ANALYSIS OF AN INDIVIDUAL GENE

The standard procedures used in analysis of an individual gene are presented in Fig. 2.

In silico analysis of an individual gene

The sequence of the ORF is a major source of information which sometimes allows the prediction of the biochemical and/or physiological function of its product. The analysis *in silico* although very helpful in planning the experiments, by itself is not sufficient to reveal the function of the gene and sometimes may be misleading.

Below is presented a brief description of some of the commonly used computer programmes. The first group, *BLAST, FASTA, BestFit* and *PileUp* are concerned with sequence homology. *MOTIFS* searches for functional motifs with a proteins sequence, while Codon bias and *PSORT* are concerned with the level of expression of a gene and the localisation of its product.

BLAST – searches for sequences similar to a query sequence. The query and the database searched can be either protein or nucleic acid and the programme can translate a nucleic acid sequence in all phases and compare it to

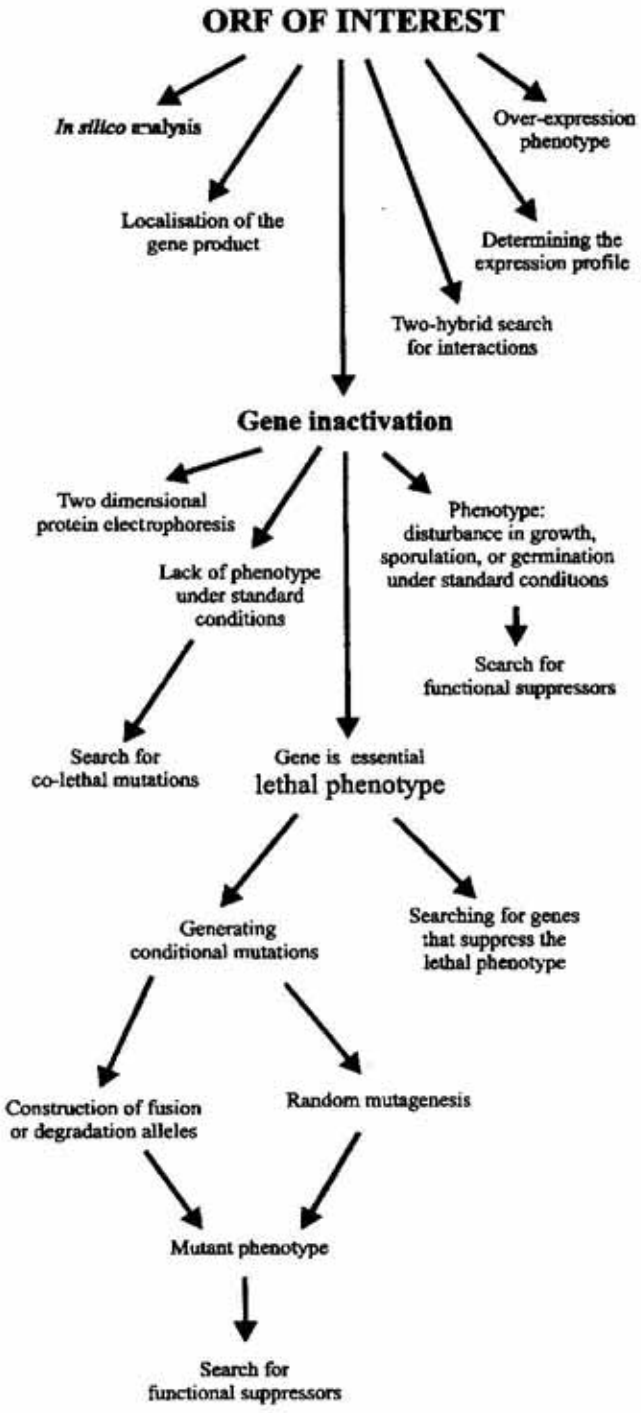

Figure 2 . Schema of the standard procedures used in analysis of an individual gene.

the protein database. *BLAST* finds the highest scoring locally optimal alignments between a query sequence and a database. The *BLAST* algorithm and family of programmes rely on work on the statistics of ungapped sequence alignments by Karlin & Altschul (1990).

FASTA – performs a Pearson & Lipman (1988) search for similarity between a query sequence and a group of sequences of the same type (nucleic acid or protein). For nucleotide searches, *FASTA* may be more sensitive than *BLAST*.

BestFit – uses the local homology algorithm of Smith & Waterman (1981) to find the optimal alignment of the best segment of similarity between two sequences. Optimal alignments are found by inserting gaps to maxi-

mise the number of matches. The sequences can be of very different lengths and have only a small segment of similarity between them.

PileUp – creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments in a simplification of the progressive alignment method of Feng & Doolittle (1987). The results can be presented on the form of a dendrogram showing the clustering relationships used to create the alignment.

Motifs and FindPatterns – the programmes like *Motifs* and *FindPatterns* from the UWGCG Package look for protein sequence motifs associated with a particular function or structure that has been identified and characterised in a group of proteins. The protein sequence is checked for all sequence patterns in the *PROSITE* Dictionary of Protein Sites and Patterns written by Dr. Bairoch of the University of Geneva (Switzerland). This database contains more than 500 protein motifs and links to the current literature concerning motifs.

Codon bias – the frequencies with which individual synonymous codons are used to code their cognate amino acids vary from genome to genome and within a given genome, from gene to gene. Genes which are highly expressed in yeast show a biased use of synonymous codons, with a preference for codons that correspond to the major cellular isoacceptor tRNA species (Bennetzen & Hall, 1982). Therefore, the bias in codon usage gives an indication of the level of gene expression.

PSORT – is a system for the prediction of protein localisation within cells. It analyses the input sequence by applying the stored rules for various sequence features of known protein sorting signals in different organisms. Then, it reports the probability that the input protein is targeted to one of several subcellular localisations. The original programme was developed by Dr. Nakai from Institute for Molecular and Cellular Biology, Osaka University (Japan); and can be used at the following website: http://psort.nibb.ac.jp

Gene expression profiles

In yeast, most genes are transcribed constitutively with similar low levels of mRNA, approximately one to two molecules per cell (Struhl, 1986). Nevertheless, certain genes do show significant changes in the level of expression in response to environmental conditions, in which case this pattern of regulation may or give some idea of the genes function. The simplest and most direct method for the detection of the steady state transcript level of the gene of interest is to reveal the mRNA of interest by hybridisation with a probe specific for the gene. To compare the levels of a given messenger under different conditions it is necessary to have an internal control that is expressed at a constant level. In most cases the actin mRNA (product of the *ACT1* gene) is used.

In transcription studies often are used the gene fusions. The most popular reporter gene is the *E. coli LacZ* gene. This can be used to produce functional hybrid β-galactosidase proteins in yeast (Guarente & Ptashne, 1981; Rose *et al.*, 1981). Typically, the *lacZ* gene is fused to a promoter, and the level of β-galactosidase in quantitative assays is taken as a measure of the level of transcription. *LacZ* gene fusions are one of most popular methods used in studies on the regulation of gene expression, promoter "dissections" and studies on the role of *cis* and *trans* acting elements in transcriptional control (Guarente, 1983). However, it is important to remember that there are several limitations to these studies. Most of the time the constructions are carried on plasmids, but under different conditions plasmid copy number is not invariant, also the role of chromatin in transcriptional regulation is automatically eliminated. Finally, in some cases elements that control the stability of an mRNA are part of the coding sequence

and thus will be eliminated in these constructions.

Localisation of the gene product

The knowledge of the localisation of a protein within the cell can give significant information about its function. Two types of experiment are possible, either the cells can be fractionated into different compartments and the protein associated with one or more compartments is identified by immunoprecipitation, or the localisation can be followed microscopically in intact cells. For both these methods it is necessary to be able to detect the protein. This can be done using antibodies to the protein itself, or an epitope that has been added to it, or by fusing the protein to another which can be easily detected. The most powerful techniques are those that allow the recognition of the active protein *in vivo*.

Protein tagging and immuno-detection

A popular method for determining the subcellular localisation of different gene products is immunological detection using fluorescent tagged (fluorescence microscopy) or metal tagged (electron microscopy) secondary antibodies that label a primary antibody that is able recognise the protein expressed at its physiological level within the cell. To avoid the procedure of protein purification and subsequent preparation of antibodies, small peptide epitopes, or tags, for which antibodies are commercially available, are often fused to the protein of interest. Several yeast–*E. coli* shuttle vectors are available which allow the fusion of any protein of interest to one of several epitopes e.g. the influenza haemagglutinin (HA) epitope or an epitope of the human c-*myc* oncogene protein (Reisdorf *et al.*, 1993). The epitope may be placed at either the N or the C terminus. Preferably, the tag will not affect the activity of the protein, the gene will be under the control of its own promoter and integrated at its normal locus in the genome. The conjunction of these conditions constitutes a strong argument for the correct localisation of hybrid protein.

GFP fusion proteins

The versatile reporter gene, *GFP*, derived from the jellyfish *Aequorea victoria*, encodes the Green Fluorescent Protein (GFP) which is an extremely stable 27 kDa monomer of 238 amino acids. GFP emits bright green light when exposed to UV or blue light. Excitation at 395 nm yields an emission maximum at 508 nm. Both the amino and carboxyl GFP termini could be successfully fused to a wide range of cytosolic and membrane proteins. The resulting fusion can be studied directly *in vivo* by fluorescence microscopy. The powerful yeast recombination system allows the incorporation of fusion constructions at the normal chromosomal locus (Wach *et al.*, 1997). Again, it is important that the fusion protein is active if the localisation is to be considered reliable.

Gene inactivation

The starting point for the genetic and phenotypic analysis of a new gene is the creation of a "null" allele. This is a completely non-functional version of the gene, normally the ORF is deleted. When working with *S. cerevisiae* several different methods can be used, but they are all based on the efficient recombination and/or gene conversion systems. A disruption cassette with a selectable marker (e.g. *URA3*, *TRP1*, *LEU2*...) flanked by 30–500 bp fragments upstream and downstream of the gene of interest is constructed and used to transform yeast. The transformants are screened by PCR or preferably by Southern blot to verify that the integration has occurred at the correct locus. The shorter the flanking regions the greater the chance of ectopic insertion. In general, if the flanking sequences are > 100 bp, > 90% of the integrants are correct; with > 50 bp flanking sequences only about 1% of the integrants are correct.

In the EUROFAN project the *kanMX* module was chosen as the selectable marker. The cassette consists of the coding sequence of the *kan*^{r} gene of the transposon Tn*903* encoding an aminoglycoside 3′-phosphotransferase (Oka *et al.*, 1981) and the promoter and terminator of the *TEF* gene from the fungus *Ashyba gossypii* (Steiner & Philippsen, 1994). The aminoglycoside transferase activity renders *S. cerevisiae* resistant to the drug geneticin (G418) (Jimenez & Davies, 1980). This cassette is well suited to the PCR amplification strategy (Wach *et al.*, 1994; 1996).

Routinely, the inactivation of a new gene will be performed in a diploid strain, only one allele will be disrupted and a tetrad analysis will reveal whether the haploid segregants carrying the deleted gene are viable. When the deleted haploid strain is viable, this will be the starting point for a genetic and phenotypic analysis.

SEARCHING FOR INTERACTIONS

The early results of global gene function analysis indicate that most of the deletion mutants (about 75%) do not display a clear-cut, easy to score phenotype or they exhibit a phenotype which is not directly informative from a functional point of view. Many of the subsequent experimental approaches are designed to elucidate interactions of various kinds.

THE TWO-HYBRID SYSTEM

The two-hybrid system was developed by Stanley Fields and co-workers (Fields & Sango, 1989; Fields & Sternglanz, 1994) and allows the detection protein–protein interactions *in vivo* through the reconstitution of the activity of a transcriptional activator. The method is based on properties of the many transcriptional activators which are composed of two distinct domains, a DNA-binding domain and an activation domain. Typically, the yeast activator product of *GAL4* gene (Gal4p) is used. The principle is that when these two domains are separated there is no transcriptional activation. However, if the two domains are separately fused to two proteins that interact physically, this interaction will bring the DNA-binding and activation domains into close proximity and restore transcriptional activation. In practice two plasmids encoding hybrid proteins are used:

- The "bait" plasmid encodes a hybrid of the Gal4p DNA-binding domain and the protein of interest X. The hybrid protein binds to DNA but will not activate transcription if X does not have an activation domain;
- The "prey" plasmid encodes a hybrid of the Gal4p activation domain and a second protein Y. This hybrid protein will not activate transcription because it does not bind to the upstream activation sequence, therefore, it is not correctly positioned.

When both hybrid proteins are present and there is a physical interaction between proteins X and Y, the Gal4p activation domain will be correctly positioned and able to activate the transcription of a reporter gene (normally *LacZ* and/or *HIS3*).

In a standard form of the two-hybrid system, a strain with the appropriate reporter genes and the bait plasmid is transformed by a bank of "pray" plasmids, and positive colonies are selected for further study. As well as this, the two-hybrid system is very well suited to probing the interactions between two specific proteins and defining the domains that interact.

The two-hybrid system, like all methods has its limitations, not all protein–protein interactions will be detected and in library searches false positives often occur. The false positives may often be due to the fact that the two components in the system are expressed at high levels and in practice it is often difficult to validate the biological significance of a new interaction in the absence of other corroborating data.

EXTRAGENIC SUPPRESSORS

Extragenic suppression of a mutant phenotype has long been used as a way of identifying genetic interactions. It is a prerequisite for all suppressor studies that a phenotype is associate with a mutation or deletion. The method is simple and is based on selecting pseudo-revertants that correct, or partially correct the mutant phenotype. A subsequent genetic analysis of the suppressors will determine if they are recessive or dominant and place them into different complementation groups, which define the number of genes involved. The next step in the analysis is the identification of these genes. In the case of dominant suppressors, a bank of the genomic DNA of the suppressor strain is constructed and the gene is cloned by its ability to suppress the original mutant or deletion. In the case of recessive suppressors, the situation is more complex, in theory, the suppressor strain transformed by a bank of wild type genomic DNA will lose the suppressor phenotype when the corresponding wild type gene is cloned. In practice, this approach rarely meets with success.

MULTICOPY SUPPRESSORS

Over-expression of one gene can suppress a mutation in a different gene (Hinnebusch & Fink, 1983; Vallen *et al.*, 1994). Experimentally, this is a much simpler system than the analysis of extragenic suppressors. The mutant strain is transformed with a bank of wild type genomic DNA cloned on a multicopy vector and transformants in which the mutant phenotype is lost are selected. It is obvious that the wild type gene corresponding to the mutant allele should also be selected in this screen, this acts as a useful internal control. For example, if many copies of the wild type gene (i.e. > 10) are isolated but no suppressors, it is probable that the bank does not contain any multicopy suppressors of the mutation. This system can also be adapted to look for multicopy suppressors of a lethal mutation. In this case a plasmid shuffling protocol is used. The chromosomal copy of the essential gene is deleted or otherwise mutated. The viability of the cell is maintained by the presence of a plasmid carrying a wild type copy of this essential gene and a *URA3* marker. In this situation, the *URA3* plasmid cannot be lost so the cells are unable to grow on medium containing 5-fluoroorotic acid (5FOA). After transformation with a yeast genomic bank in a multicopy vector the cells are replica plated onto the 5FOA containing medium. The *URA3* gene encodes orotidine-5′-phosphate decarboxylase which is able to metabolise 5FOA, this will eventually be transformed into fluorouracil which is toxic for the cell. Thus the presence of 5FOA provides a counter-selection for *URA3* cells (Boeke *et al.*, 1984). Colonies that are able to grow on the 5FOA medium have lost the *URA3* containing plasmid with the wild type gene, thus they must have acquired another gene which is able to compensate for the lose of this gene. Here again, the wild type gene should be isolated and can act as an internal control to see if the bank has been exhaustively probed.

The advantage of multicopy suppressors compared to extragenic suppressors is that they are easier to analyse and the genes involved are more readily identified. However, in general extragenic suppressors are more informative.

CO-LETHAL MUTATIONS

A synthetic phenotype is one that is due to the combination of mutant alleles of two different genes, which individually do not exhibit the phenotype. The extreme synthetic phenotype is lethality, or more correctly co-lethality. The co-lethal phenotype can provide evidence for an interaction between gene products (Huffaker, 1987) and is particularly useful as it enables interactions to be uncovered when

the first mutation has no phenotype. In the light of all the preliminary studies which show that many gene inactivations have no clear phenotype this would seem to be a very informative approach.

To isolate co-lethal mutations in yeast a plasmid shuffling strategy is employed, again this system is based on the ability to counter-select the *URA3* gene with 5FOA. Other systems using the white/red colour of *ade2, ade3 vs. ade2, ADE3* strains or a combination of both 5FOA and white/red have been developed. A strain carrying a mutation or deletion that has no phenotype and a *URA3* plasmid encoding the wild type gene is mutagenised. After regeneration, the cells are plated on uracil containing medium and replicated onto 5FOA medium. Colonies carrying a co-lethal mutation will not be able to lose the *URA3* plasmid and hence will not be able to grow on the 5FOA medium.

The gene corresponding to the co-lethal mutation can be isolated by transforming the double mutant strain and looking for transformants that are now able to grow on 5FOA.

ANALYSIS OF ESSENTIAL GENES

The preliminary studies that have been performed in many laboratories show that about 10% of yeast genes are essential for viability. The discovery that a gene under study is essential is both exciting and frustrating. It is exciting, because by definition the gene plays a vital role in the cell, but it is frustrating because many of techniques described above are not applicable to essential genes. All of the experimental approaches that do not require a mutant, such as the localisation of the gene product, the analysis of expression profiles and two-hybrid studies are applicable. As well as these techniques, modulation of the expression of an essential gene can be informative, especially of other data have already given some clue to the function of the gene. The other principal approach to the study of an essential gene is to isolate a conditional allele, when this has been done, the other standard genetic analyses can be applied to the study of the gene.

MODULATION GENE EXPRESSION

The expression of a gene can be modulated by the use of conditional promoters to induce, or repress transcription, or by constructing fusion proteins that will be very quickly degraded. The latter are normally combined with a conditional promoter. The over-expression of a gene normally leads to an increased concentration of the wild-type protein in the cell. If the interacting proteins must be produced in appropriate stoichiometry then over-expression may create novel phenotypes that give information concerning the function of the protein (reviewed in Rine, 1991). The simplest way to achieve over-expression is placing the gene on a high copy number plasmid. This method has several disadvantages: the level of expression may not be high enough to provoke a phenotype, or for genes whose products are very toxic, the plasmid may simply be lost, or rearranged to stop expression of the gene. More commonly, an inducible strong promoter is used, fused to the gene of interest. The most frequently used promoter construction is a combination of the upstream activating sequence (UAS) of *GAL1-10* fused to the *CYC1* promoter (UAS*GAL10-CYC1* promoter), this can be induced over 1000-fold when cells are grown on galactose medium (Schena *et al.* 1991). The same type of construction can be used to down-regulate gene expression. When glucose is added to cells carrying a gene fused to the UAS*GAL10-CYC1* promoter, expression will be turned off. Two problems are associated with the shut-off experiment: first, for poorly expressed genes the level of expression from the repressed UAS*GAL10-CYC1* promoter may be sufficient, second, if the protein has a long half-life and it was originally highly over-expressed, it

may take many generations to deplete the protein.

The first problem can be solved by using other conditional promoters, such as promoters from genes involved in amino-acid biosynthesis, these have a lower level of expression and are much more tightly repressed than the UAS*GAL10*. The second problem can be solved by arranging for the protein to have a rapid turnover (Park *et al.*, 1992).

Promoter fusions such as those described above are conditional alleles, however, they are not suitable to a genetic analysis as most of the mutations identified would concern the regulation of the galactose genes.

GENERATING CONDITIONAL ALLELES

The principal for isolating conditional alleles for an individual essential gene is essentially the same for all methods. A strain carrying the deleted gene in the genome and a wild type copy on a *URA3* plasmid is transformed by a bank of the mutagenised wild type gene on a plasmid with a different marker. Plasmid shuffling and counter-selection of the *URA3* marker with 5FOA are then used to try to identify mutated plasmids that are able to replace the wild type gene under one condition (i.e. at 24°C), but not at another (i.e. at 36°C). The differences are essentially in the methods that are used to mutagenise the second plasmid carrying the wild type gene. The most common methods of *in vitro* mutagenesis of a cloned gene are by hydroxylamine (Rose & Fink, 1987), the passage of the plasmid through a mutator strain of *E. coli* (Greener *et al.*, 1997), PCR mutagenesis in the presence of manganese or using dITP or nucleotide starvation to increase the error rate (Spee *et al.*, 1993). Once generated, the pool of the mutagenised plasmids are treated as described above.

SUMMARY

Thanks to close cooperation with the group of Professor P.P. Słonimski from Centre de Genetique Moleculaire CNRS our Institute participated in the project to sequence chromosomes II and X (Becam *et al.*, 1994; Zagulski *et al.*, 1995) and is taking part in the functional analysis of the yeast genome (Gromadka *et al.*, 1996a; 1996b; Kucharczyk *et al.*, 1998), this prompted us to write this brief survey. It shows that the yeast *Saccharomyces cerevisiae* occupies a unique position in modern biology, the genome sequence has been completed and for the first time we know all the elements that make up a simple eukaryotic cell. The challenge that lies before us is to interpret that information. The techniques described above and others that are being developed make *S. cerevisiae* the ideal choice for the functional analysis of the large number ORFs of unknown function. That fact that many of these ORFs are conserved in higher eukaryotes and even man, means that this analysis is of great significance, not only to yeast geneticists, but also to the wider biological community.

REFERENCES

Barton, A.B. & Kaback, D.B. (1994) Molecular cloning of chromosome I DNA from *Saccharomyces cerevisiae*: Analysis of the genes in the *FUN38-MAK16-SPO7* region. *J. Bacteriol.* **176**, 1872–1880.

Becam, A.-M., Cullin, C., Grzybowska, E., Lacroute, F., Nasr, F., Ozier-Kalogeropoulos, O., Pałucha, A., Słonimski, P.P., Zagulski, M. & Herbert, C.J. (1994) The sequence of 29.7 kb from right arm of chromosome II reveals 13 complete open reading frames, of which ten correspond to new genes. *Yeast* **10**, S1–S11.

Bennetzen, J.L. & Hall, B.D. (1982) Codon selection in yeast. *J. Biol. Chem.* **257**, 3026–3031.

Boeke, J.D., LaCroute, F. & Fink, G.R. (1984) A positive selection for mutants lacking orotidine-5'-phosphate decarboxylase activity in yeast: 5-Fluoro-orotic acid resistance. *Mol. Gen. Genet.* **197**, 345-346.

Botstein, D. & Cherry, J.M. (1997) Molecular linguistics: Extracting information from gene and protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 5506-5507.

Boucherie, H., Dujardin, G., Kermorgant, M., Monribot, C., Slonimski, P. & Perrot, M. (1995) Two-dimensional protein map of *Saccharomyces cerevisiae:* Construction of a gene-protein index. *Yeast* **11**, 601-613.

Burns, N., Grimwade, B., Ross-Macdonald, P.B., Choi, E.Y., Finberg, K., Roeder, G.S. & Snyder, M. (1994) Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae. Genes Dev.* **8**, 1087-1105.

DeRisi, J.L., Iyer, V.R. & Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686.

Feng, D.F. & Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351-360.

Fields, S. & Song, O-k. (1989) A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-246.

Fields, S. & Sternglanz, R. (1994) The two-hybrid system: An assay for protein-protein interactions. *Trends Genet.* **10**, 286-292.

Greener, A., Callahan, M. & Jerpseth, B.A. (1997) An efficient random mutagenesis technique using an *E. coli* mutator strain. *Mol. Biotechnol.* **7**, 189-195.

Gromadka, R., Góra, M., Zielenkiewicz, U., Słonimski, P.P. & Rytka, J. (1996a) Subtelomeric duplications in *Saccharomyces cerevisiae* chromosomes III and XI: Topology, arrangements, correction of sequence and strain-specific polymorphism. *Yeast* **12**, 583-591.

Gromadka, R., Kaniak, A., Słonimski, P.P. & Rytka, J. (1996b) A novel cross phylum of proteins comprises a *KRR1* (YCL 059c) gene which is essential for viability of *Saccharomyces cerevisiae. Gene* **171**, 27-32.

Guarente, L. & Ptashne, M. (1981) Fusion of *Escherichia coli lacZ* to the cytochrome *c* gene of *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U.S.A.* **78**, 2199-2203.

Guarente, L. (1983) Yeast promoters and *lacZ* fusions designed to study expression of cloned genes in yeast. *Methods Enzymol.* **101**, 181-191.

Hinnebusch, A.G. & Fink, G.R. (1983) Positive regulation in the general amino acid control of *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U.S.A.* **80**, 5374-5378.

Huffaker, T.C., Hoyt, M.A. & Botstein, D. (1987) Genetic analysis of the yeast cytoskeleton. *Annu. Rev. Genet.* **21**, 259-284.

Jimenez, A. & Davies, J. (1980) Expression of a transposable antibiotic resistance element in *Saccharomyces. Nature* **287**, 869-871.

Kaback, D.B., Angerer, L.M. & Davidson, N. (1979) Improved methods for the formation and stabilization of R-loops. *Nucleic Acids Res.* **6**, 2499-2317.

Kahn, P. (1995) From genome to proteome: Looking at a cell's proteins. *Science* **270**, 369-370.

Karlin, S. & Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2264-2268.

Kucharczyk, R., Zagulski, M., Rytka, J. & Herbert, C.J. (1998) The yeast gene *YJR025c* encodes a 3-hydroxyanthranilic acid dioxygenase and is involved in nicotinic acid biosynthesis. *FEBS Lett.* **424**, 127-130.

Mewes, H.W., Albermann, K., Bahr, M., Frishman, D. *et al.* (1997) Overview of the yeast genome. *Nature* (Suppl.) **387**, 7-65.

Oka, A., Sugisaki, H. & Takanami, M. (1981) Nucleotide sequence of the kanamycin resistance transposon Tn*903*. *J. Mol. Biol.* **147**, 217-226.

Park, E.C., Finley, D. & Szostak, J.W. (1992) A strategy for the generation of conditional mutations by protein destabilization. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1249-1252.

Pearson, W.R. & Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444–2448.

Reisdorf, P., Maarse, A.C. & Daignan-Fornier, B. (1993) Epitope-tagging vectors designed for yeast. *Curr. Genet.* **23**, 181–183.

Richard, G.F., Fairhead, C. & Dujon, B. (1997) Complete transcriptional map of yeast chromosome XI in different life conditions. *J. Mol. Biol.* **268**, 303–321.

Rieger, K.-J., Kaniak, A., Coppee, J.-Y., Aljinovic, G., Baudin-Baillieu, A., Orłowska, G., Gromadka, R., Groudinsky, O., Di Rago, J.-P. & Słonimski, P.P. (1997) Large scale phenotypic analysis. The pilot project on yeast chromosome III. *Yeast* **13**, 1547–1562.

Rine, J. (1991) Gene overexpression in studies of *Saccharomyces cerevisiae*. *Methods Enzymol.* **194**, 239–251.

Rose, M.D., Casadaban, M.J. & Botstein, D. (1981) Yeast genes fused to beta-galactosidase in *Escherichia coli* can be expressed normally in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **78**, 2460–2464.

Rose, M.D. & Fink, G.R. (1987) *KAR 1*, a gene required for function of both intracellular and extracellular microtubules in yeast. *Cell* **48**, 1047–1060.

Schena, M., Picard, D. & Yamamoto, K.R. (1991) Vectors for constitutive and inducible gene expression in yeast. *Methods Enzymol.* **194**, 389–398.

Shevchenko, A., Jensen, O.N., Podtelejnikov, A.V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H. & Mann, M. (1996) Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 14440–14445.

Shoemaker, D.D., Lashkari, D.A., Morris, D., Mittmann, M. & Davis, R.W. (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy *Nature Genet.* **14**, 450–456.

Slonimski, P.P. & Brouillet, S. (1993) A data-base of chromosome III of *Saccharomyces cerevisiae*. *Yeast* **9**, 941–1029.

Smith, R.C. & Waterman, M.S. (1981) Comparison of bio-sequences. *Adv. Appl. Math.* **2**, 482–489.

Smith, V., Chou, K.N., Lashkari, D., Botstein, D. & Brown, P.O. (1996) Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069–2074.

Spee, J.H., de Vos, W.M. & Kuipers, O.P. (1993) Efficient random mutagenesis method with adjustable mutation frequency by use of PCR and dITP. *Nucleic Acids Res.* **21**, 777–778.

Steiner, S. & Philippsen, P. (1994) Sequence and promoter analysis of the highly expressed *TEF* gene of the filamentous fungus *Ashbya gossypii*. *Mol. Gen. Genet.* **242**, 263–271.

Struhl, K. (1986) Constitutive and inducible *Saccharomyces cerevisiae* promoters: Evidence for two distinct molecular mechanisms. *Mol. Cell Biol.* **6**, 3847–3853.

Wach, A. (1996) PCR-synthesis of marker cassettes with long flanking homology regions for gene disruptions in *S. cerevisiae*. *Yeast* **12**, 259–265.

Wach, A., Brachat, A., Pohlmann, R. & Philippsen, P. (1994) New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast* **10**, 1793–1808.

Wach, A., Brachat, A., Alberti-Segui, C., Rebischung, C. & Philippsen, P. (1997) Heterologous *HIS3* marker and GFP reporter modules for PCR-targeting in *Saccharomyces cerevisiae*. *Yeast* **13**, 1065–1075.

Wilkins, M.R., Sanchez, J.C., Gooley, A.A., Appel, R.D., Humphery-Smith, I., Hochstrasser, D.F. & Williams, K.L. (1996) Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it *Biotechnol. Genet. Eng. Rev.* **13**, 19–50.

Vallen, E.A., Ho, W., Winey, M. & Rose, M.D. (1994) Genetic interactions between CDC31 and KAR1, two genes required for duplication of the microtubule organizing center in *Saccharomyces cerevisiae*. *Genetics* **137**, 407–422.

Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. (1995) Serial analysis of gene expression. *Science* **270**, 484–487.

Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Jr., Hieter, P., Vogelstein, B. & Kinzler, K.W. (1997) Characterization of the yeast transcriptome. *Cell* **88**, 243–251.

Yoshikawa, A. & Isono, K. (1990) Chromosome III of *Saccharomyces cerevisiae:* An ordered clone bank, a detailed restriction map and analysis of transcripts suggest the presence of 160 genes. *Yeast* **6**, 383–401.

Yoshikawa, A. & Isono, K. (1991) Construction of an ordered clone bank and systematic analysis of the whole transcripts of chromosome VI of *Saccharomyces cerevisiae. Nucleic Acids Res.* **19**, 1189–1195.

Zagulski, M., Babinska, B., Gromadka, R., Migdalski, A., Rytka, J., Sulicka, J. & Herbert, C.J. (1995) The sequence of 24.3 kb from chromosome X reveals 5 complete open reading frames all of which correspond to new genes, and a tandem insertion of a Ty1 transposon. *Yeast* **11**, 1179–1186.