The EMBO Journal vol.13 no.24 pp.5795–5809, 1994

Complete DNA sequence of yeast chromosome II

H.Feldmann[1,2], M.Aigle[3], G.Aljinovic[4], B.André[5], M.C.Baclet[3], C.Barthe[3], A.Baur[6], A.-M.Bécam[7], N.Biteau[3], E.Boles[6], T.Brandt[8], M.Brendel[9], M.Brückner[10], F.Bussereau[11], C.Christiansen[8], R.Contreras[12], M.Crouzet[3], C.Czepluch[13], N.Démolis[11], Th.Delaveau[14], F.Doignon[3], H.Domdey[15], S.Düsterhus[16], E.Dubois[17], B.Dujon[18], M.El Bakkoury[17], K.-D.Entian[16], M.Feuermann[19], W.Fiers[12], G.M.Fobo[20], C.Fritz[21], H.Gassenhuber[15], N.Glansdorff[17], A.Goffeau[22,23], L.A.Grivell[24], M.de Haan[24], C.Hein[5], C.J.Herbert[7], C.P.Hollenberg[21], K.Holmstrøm[8], C.Jacq[14], M.Jacquet[11], J.C.Jauniaux[5,13], J.-L.Jonniaux[22], T.Kallesøe[8], P.Kiesau[16], L.Kirchrath[21], P.Kötter[16], S.Korol[16], S.Liebl[20], M.Logghe[12], A.J.E.Lohan[25], E.J.Louis[26], Z.Y.Li[9], M.J.Maat[24], L.Mallet[11], G.Mannhaupt[1], F.Messenguy[17], T.Miosga[6], F.Molemans[12], S.Müller[10], F.Nasr[7], B.Obermaier[15], J.Perea[14], A.Piérard[17], E.Piravandi[15], F.M.Pohl[27], T.M.Pohl[4], S.Potier[19], M.Proft[16], B.Purnelle[22], M.Ramezani Rad[21], M.Rieger[10], M.Rose[16], I.Schaaff-Gerstenschläger[6], B.Scherens[17], C.Schwarzlose[1], J.Skala[22], P.P.Slonimski[7], P.H.M.Smits[24], J.L.Souciet[19], H.Y.Steensma[28], R.Stucka[1], A.Urrestarazu[5], Q.J.M.van der Aart[28], L.van Dyck[22], A.Vassarotti[23], I.Vetter[1], F.Vierendeels[17], S.Vissers[5], G.Wagner[10], P.de Wergifosse[22], K.H.Wolfe[25], M.Zagulski[7], F.K.Zimmermann[6], H.W.Mewes[20] and K.Kleine[20]

[1]Institut für Physiologische Chemie, Physikalische Biochemie und Zellbiologie der Universität München, Schillerstraße 44, D-80336 München, Germany, [3]Université de Bordeaux II, Laboratoire de Biologie Moléculaire et de Séquencage, Rue Léo Saignat, F-33076 Bordeaux Cedex, France, [4]GATC-Gesellschaft für Analyse Technik und Consulting, Fritz-Arnold-Strasse 23, D-78467 Konstanz, Germany, [5]Laboratoire de Physiologie Cellulaire et de Génétique des Levures, Université Libre de Bruxelles, Campus de la Plaine, CP244, Boulevard du Triomphe, B-1050, Bruxelles, Belgium, [6]Institut für Mikrobiologie, TH Darmstadt, Schnittspahnstraße 10, D-64287 Darmstadt, Germany, [7]Centre National de la Recherche Scientifique (CNRS), Centre de Génétique Moléculaire, F-91198 Gif-sur-Yvette Cedex, France, [8]Research Institute for Food Technology, Agro-Industrial Technology and Molecular Biotechnology, Biotechnological Institute, Lundtoftevej 100, Building 227, PO Box 199, DK-2800 Lyngby, Denmark, [9]Johann Wolfgang Goethe-Universität Frankfurt, Institut für Mikrobiologie, Theodor-Stern-Kai 7, Haus 75A, D-60596 Frankfurt/M, Germany, [10]Genotype GmbH, Biotechnologische und Molekularbiologische Forschung, Angelhofweg 39, D-69259 Wilhelmsfeld, Germany, [11]Université de Paris-Sud, Institut de Génétique et Microbiologie, URA1354 du CNRS, Laboratoire Information Génétique et Développement, Bat. 400, F-91405 Orsay Cedex, France, [12]Rijksuniversiteit Gent, Laboratorium voor Moleculaire Biologie, K.L.Ledeganckstraat 35, B-9000 Gent, Belgium, [13]Angewandte Tumorvirologie and Virologie appliquée à l'oncologie (Unité INSERM 375), Deutsches Krebsforschungszentrum, Abt. 0610, P.101949, D-69009 Heidelberg, Germany, [14]Ecole Normale Superieur, Laboratoire de Génétique Moléculaire, CNRS URA 1302, Rue d'Ulm 46, F-75230 Paris Cedex 05, France, [15]Genzentrum der Ludwig-Maximilians-Universität München, Laboratorium für Molekulare Biologie, Am Klopferspitz 18a, D-82152 Martinsried/München, Germany, [16]Johann Wolfgang Goethe-Universität Frankfurt, Institut für Mikrobiologie, Biozentrum, Marie-Curie-Straße 9, D-60439 Frankfurt/M, Germany, [17]Institut de Recherches du CERIA, COOVI, Laboratoire de Microbiologie, Université Libre de Bruxelles and Laboratorium voor Erfelikheidesleer en Microbiologie, Vrije Universiteit Brussel, Avenue E.Gryson 1, B-1070 Brussels, Belgium, [18]Unité de Génétique Moléculaire des Levures (URA 1149 du CNRS), Département de Biologie Moléculaire, Institut Pasteur, F-75724 Paris Cedex 15, France, [19]CNRS, Institut de Botanique, 28 Rue Goethe, F-67083 Strasbourg Cedex, France, [20]MIPS, Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, D-82152 Martinsried, Germany, [21]Institut für Mikrobiologie der Heinrich-Heine-Universität Düsseldorf, Geb. 26.12, Universitätstrasse 1, D-40225 Düsseldorf, Germany, [22]Unité de Biochimie Physiologique, Université Catholique de Louvain, Place Croix du Sud 2-20, B-1348 Louvain-la-Neuve, Belgium, [23]Commission of the European Communities, B-1049, Brussels, Belgium, [24]Universiteit van Amsterdam, Sectie Moleculaire Biologie, Vakgroep Moleculaire Celbiologie, Kruislaan 318, NL-1098 SM Amsterdam, The Netherlands, [25]University of Dublin, Department of Genetics, Lincoln Place Gate, Trinity College, Dublin 2, Ireland, [26]Yeast Genetics, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK, [27]Fakultät für Biologie der Universität Konstanz, Postfach 55 60, D-78434 Konstanz, Germany, [28]Leiden University, Clusius Laboratory, Department of Cell Biology and Genetics, Wassenaarseweg 64, NL-2333 AL Leiden, The Netherlands

[2]Corresponding author

Communicated by H.Feldmann

In the framework of the EU genome-sequencing programmes, the complete DNA sequence of the yeast *Saccharomyces cerevisiae* chromosome II (807 188 bp) has been determined. At present, this is the largest eukaryotic chromosome entirely sequenced. A total of 410 open reading frames (ORFs) were identified, covering 72% of the sequence. Similarity searches revealed that 124 ORFs (30%) correspond to genes of known function, 51 ORFs (12.5%) appear to be homologues of genes whose functions are known, 52 others (12.5%) have homologues the functions of which are not well defined and another 33 of the novel putative genes (8%) exhibit a degree of similarity which is insufficient to confidently assign function. Of the genes on chromosome II, 37–45% are thus of unpredicted function. Among the novel putative genes, we found several that are related to genes that perform differentiated functions in multicellular organisms or are involved in malignancy. In addition to a compact arrangement of potential protein coding sequences, the analysis of this chromosome confirmed general

chromosome patterns but also revealed particular novel features of chromosomal organization. Alternating regional variations in average base composition correlate with variations in local gene density along chromosome II, as observed in chromosomes XI and III. We propose that functional ARS elements are preferably located in the AT-rich regions that have a spacing of ~110 kb. Similarly, the 13 tRNA genes and the three Ty elements of chromosome II are found in AT-rich regions. In chromosome II, the distribution of coding sequences between the two strands is biased, with a ratio of 1.3:1. An interesting aspect regarding the evolution of the eukaryotic genome is the finding that chromosome II has a high degree of internal genetic redundancy, amounting to 16% of the coding capacity.

Key words: compositional bias/gene function/gene redundancy/genome organization/putative replication origins

Introduction

The current genome projects endeavour to decipher the genetic information of a number of organisms by establishing detailed maps and finally complete sequences of their genomes. With the present level of sequencing methodology, early efforts at genome sequencing have been concentrated on organisms with less complex genomes. In this context, model organisms like bacteria (Kunst and Devine, 1991; Daniels *et al.*, 1992; Honore *et al.*, 1993) or organisms with genomes of intermediate sizes such as *Caenorhabditis elegans* (Wilson *et al.*, 1994) or *Arabidopsis thaliana* (Meyerowitz and Pruitt, 1985) assume great importance as experimental systems. Among all eukaryotic model organisms, *Saccharomyces cerevisiae* combines several advantages: (i) this yeast has a genome size of only 13.5 Mb, i.e. 220 times smaller than that of the human genome; (ii) the yeast system is tractable to powerful genetic techniques; and (iii) functions in yeast have been studied in great detail biochemically. Based on present data, one can calculate that a repertoire of 6500–7000 genes is sufficient to build this simple eukaryotic cell. Considering recent progress and worldwide studies of yeast genome sequencing (Vassarotti and Goffeau, 1992; Goffeau, 1994), we can be confident of deciphering its genetic potential within a reasonable time period and with relatively limited effort.

Since a large variety of examples provide evidence that substantial cellular functions are highly conserved from yeast to mammals, and that corresponding genes can often complement each other, the wealth of sequence information obtained in yeast will be extremely useful as a reference against which sequences of human, animal or plant genes may be compared. Moreover, the ease of genetic manipulation in yeast opens up the possibility of functionally dissecting gene products from other eukaryotes in the yeast system.

Two years ago a consortium of 35 European laboratories published the first complete sequence of a eukaryotic chromosome: chromosome III of *S.cerevisiae* (Oliver *et al.*, 1992). For the past 3 years our consortium has turned its efforts to the sequencing of yeast chromosomes XI and II and will continue to contribute to the sequencing of the yeast genome. The sequence of chromosome XI, the second eukaryotic chromosome entirely sequenced, has been published recently (Dujon *et al.*, 1994). We report here the complete sequence of chromosome II (807 188 bp), the largest eukaryotic chromosome sequence ever entirely determined. The sequence of chromosome II, which constitutes ~6% of the yeast genome, adds considerably to the body of information we have gained so far from chromosomes III and XI, which together make up ~7.3% of the genome. Apart from the many novel genes detected in chromosome II, we have also arrived at a more precise description of the organization of the yeast genome. The size of chromosome II is sufficient to reveal specific novel chromosomal organization patterns; combined with the previous data from chromosomes III and XI, its analysis permits us to substantiate general principles of chromosomal organization in yeast.

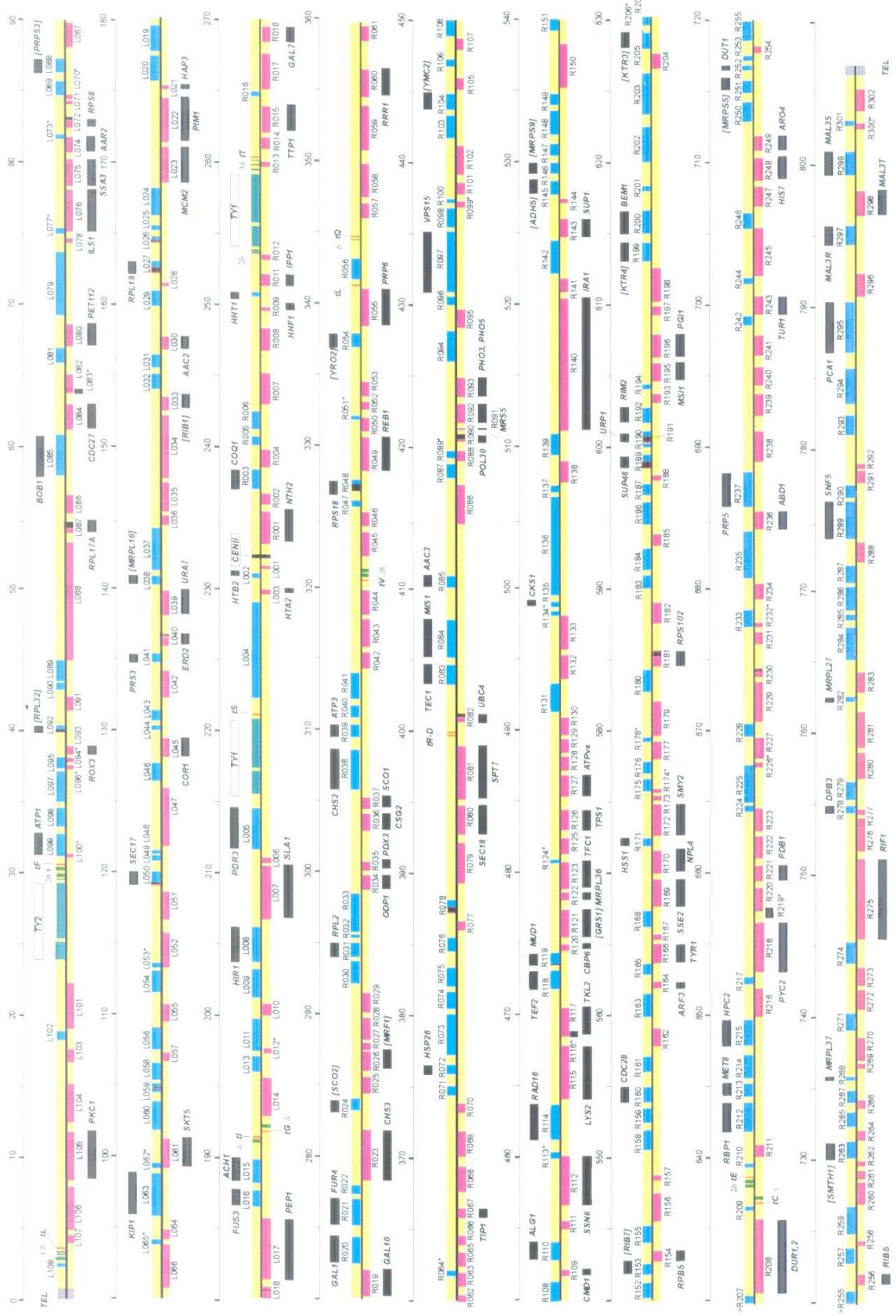
Results

Assembly and verification of sequence

The sequence was determined from a set of 43 selected partially overlapping cosmid clones of a purpose-built genomic library from *S.cerevisiae* strain αS288C, supplemented by an overlapping plasmid clone containing the right telomere. By cross-reference with an ordered library from strain C836, established prior to this work (Stucka, 1992), and by chromosomal walking, a set of overlapping cosmid clones for chromosome II from strain αS288C was generated. These cosmids then served to construct the physical map using the restriction enzymes *Bam*HI, *Sal*I, *Xho*I and *Xba*I (average resolution ~2 kb).

Clones were distributed between the collaborating laboratories according to a scheme to be presented elsewhere (H.Feldmann *et al.*, manuscript in preparation). Assembly and interpretation of the sequence followed the same principles as those applied for chromosome XI (Dujon *et al.*, 1994). Telomeres were physically mapped relative to the terminal-most cosmid inserts using the I-*Sce*I chromosome fragmentation procedure described by Thierry and Dujon (1992). From this analysis it follows that the right telomere is completely contained in the sequence presented here. This sequence was determined from a specific plasmid clone (pEL19B2) obtained by

Fig. 1. *Saccharomyces cerevisiae* chromosome II map as deduced from the complete sequence. The map is drawn to scale from the sequence and coordinates (top line) are in kb. The genetic elements on the two strands are shown as coloured bars. The top strand (designated 'Watson' strand) is oriented 5′ to 3′ from left to right. The sequence has been interpreted using the principles detailed in Materials and methods. This procedure identified 410 ORFs (blue and purple boxes), which have been numbered in increasing order from the centromere and designated L for the left arm and R for the right arm (note that the database entries will use a more complex nomenclature, namely YBL for ORFs on the left arm and YBR for ORFs on the right arm, followed by a w/c suffix indicating their location on the Watson–Crick coding strand; see also Table I). ORFs corresponding to known genes are indicated by black bars. Tentative gene names are in brackets. Ty elements (or remnants thereof) are shown as green bars. δ, σ and τ refer to the LTRs of the Ty1/2, Ty3 and Ty4 elements, respectively, or remnants thereof. tRNA genes (red bars) are symbolized by a t and the one-letter code for the amino acid accepted.



'telomere trap cloning' (Louis *et al.*, 1994) and was shown to overlap the right-most cosmid by some 12 kb. At its left end, the sequence presented here starts with a telomere consensus sequence contained in the left-most cosmid. According to our physical mapping data, chromosome II extends some 7 kb beyond this telomere consensus. A detailed analysis of yeast telomeres indicated that chromosome II carries an additional 5.2 kb Y′ element beyond the left telomere consensus sequence (Louis *et al.*, 1994), which is in full agreement with the mapping data.

During sequence assembly, quality controls were performed following the same principles as applied in sequencing yeast chromosome XI (Dujon *et al.*, 1994; see Materials and methods). We are confident, therefore, that a level of 99.97% accuracy has been achieved. Partial sequences of chromosome II have been published independently by the authors of this work (Delaveau *et al.*, 1992, 1994; Skala *et al.*, 1992, 1994; Van Dyck *et al.*, 1992, 1993, 1994; Baur *et al.*, 1993; Bussereau *et al.*, 1993; Démolis *et al.*, 1993, 1994; Doignon *et al.*, 1993a,b; Miosga and Zimmermann, 1993; Schaaff-Gerstenschläger *et al.*, 1993a,b, 1994; Scherens *et al.*, 1993; Bécam *et al.*, 1994; De Wergifosse *et al.*, 1994; Holmstrøm *et al.*, 1994; Logghe *et al.*, 1994; Mallet *et al.*, 1994; Mannhaupt *et al.*, 1994; Nasr *et al.*, 1994a,b; Ramezani Rad *et al.*, 1994; Smits *et al.*, 1994; van der Aart *et al.*, 1994; Wolfe and Lohan, 1994; Zagulski *et al.*, 1994).

Definition of open reading frames (ORFs) and other genetic elements

The map of chromosome II, as deduced from the complete sequence, is shown in Figure 1. A total of 410 ORFs were identified in the entire chromosome using the principles explained in Materials and methods, disregarding the six ORFs contributed by three complete yeast retrotransposons (Ty elements); 17 ORFs, mainly encoding ribosomal proteins, are interrupted by introns. The list includes 30 pairs of partially overlapping ORFs, five representing parallel and 25 representing anti-parallel overlaps; 11 pairs each include a gene whose function is known, whereas 10 other pairs include an ORF whose predicted product has a homologue in the databases, suggesting that it corresponds to a real gene (for details see H.Feldmann *et al.*, manuscript in preparation). Moreover, in all such cases the partially overlapping partner ORF is shorter, suggesting that it may not correspond to a real gene. This leaves uncertainty for nine of the pairs of overlapping ORFs.

In the analysis of chromosome XI (Dujon *et al.*, 1994), each ORF was evaluated using the codon adaptation index (CAI; Sharp and Li, 1987) and ORF sizes as criteria: ORFs that were both <150 codons in length and had a CAI <0.110 were considered as 'questionable'. If the same criteria are applied to chromosome II, 38 ORFs fall into this category; of these, 21 also belong to the set of partially overlapping ORFs. These criteria, however, can be used only as a guideline rather than a strict rule; in chromosome II, three genes of known function are <150 codons in length and have CAI values between 0.091 and 0.067. If we exclude from the list of 410 predicted chromosome II ORFs 21 of the partially overlapping ORFs and the residual 17 out of the 'questionable' ORFs, we arrive at a total number of 372 ORFs that might correspond to real genes. Necessarily, this estimate remains uncertain until the number of expressed genes in chromosome II has been determined by experimental methods.

A total of 13 tRNA genes, one of them containing an intron, have been identified on chromosome II. Among these is one copy of the tandemly arranged pairs of tRNA(Arg)−tRNA(Asp) genes. As anticipated (e.g. Hauber *et al.*, 1988; Ji *et al.*, 1993), most of the tRNA genes are associated with complete Ty elements, with their long terminal repeats (δ, σ and τ) and/or remnants thereof. Two of the three complete Ty elements encountered on chromosome II belong to class 1 and one to class 2 elements.

Analysis of the predicted protein products

Comparison of the present sequence with public databases revealed that 124 of the 410 ORFs (30%) correspond either to previously known protein-encoding genes or to genes whose functions have been determined during this work. In all, 70% of the total ORFs represent novel putative yeast genes; 51 of them (12.5% of the total) have homologues among gene products from yeast or other organisms whose functions are known, whereas 52 others (12.5% of the total) have homologues whose functions are not well defined. A further 33 ORFs of the novel putative genes (8% of the total) show a degree of similarity which is insufficient to confidently assign function (see Materials and methods). The remaining 150 ORFs (37% of the total) have either homologues to ORFs of unknown function on other chromosomes or no homologues in data libraries at all (note that this last set includes 17 'questionable' ORFs). Overall, between 37 and 45% of the genes of chromosome II are thus of unpredicted function. All of the above figures are similar to those obtained for chromosomes III (Oliver *et al.*, 1992) and XI (Dujon *et al.*, 1994). Table I lists the known genes plus all those ORFs which are considered to be homologues of gene products from yeast or other organisms whose functions are known or whose functions can be predicted from similarity scores and protein signatures.

We have analysed the chromosome II ORFs by using the ALOM algorithm (Klein *et al.*, 1985) to predict putative membrane spans. A total of 142 ORFs (disregarding 21 'questionable' ORFs from the partially overlapping pairs) were found to contain from one to 14 potential membrane transversions (one ORF even showed 21 such spans). These results were confirmed by visual inspection of hydropathy plots of the ORFs in question. Thus, we arrive at an estimate that some 38% of the 'real' genes in chromosome II may code for transmembrane proteins. A similarly high figure has already been found with chromosome III (Goffeau *et al.*, 1993a,b). Preliminary data obtained from other systematically sequenced yeast chromosomes suggest that this may apply as a general rule in yeast (A.Goffeau, personal communication). Even though the algorithm may give a somewhat high estimate, possibly a third of the yeast proteins have to be considered to be associated with membrane structures. We also examined chromosome II ORFs for the occurrence of putative mitochondrial target signal sequences. A rough estimate is that 8–10% of the proteins may be designed for mitochondrial import. Details of these analyses will

Table 1. Genes of known or predicted function on chromosome II

ORF	Size (aa)[a]	Gene[b]	Function	CAI
YBL106c	1010		probable G-protein, β-transducin type	0.123
YBL105c	1151	*PKC1*	protein kinase C-like protein (Ser/Thr-specific)	0.171
YBL103c	316		probable cytochrome *c* subunit, copper binding	0.125
YBL000	438	*TY2A*		0.180
YBL000	1770t	*TY2B*		0.150
YBL099w	545	*ATP1*	mitochondrial ATPase, α chain precursor	0.470
YBL093c	220	*ROX3*	nuclear protein involved in CYC7 expression	0.104
YBL092w	130i	*[RPL32e]*	probable ribosomal protein L32.e	0.817
YBL091c	349		probable IF2-associated glycoprotein	0.211
YBL088c	2787		probable PI3 kinase, DRR1 homologue	0.122
YBL087c	137i	*RPL17A*	probable ribosomal protein L23.e or YL17a	0.623
YBL085w	980	*BOB1*	BEM1 binding protein	0.132
YBL084c	758	*CDC27*	cell division control protein CDC27	0.131
YBL080c	541	*PET112*	maintenance of rho^+ mitochondrial DNA	0.132
YBL076c	1072	*ILS1*	isoleucyl-tRNA synthetase	0.342
YBL075c	649	*SSA3*	heat-shock protein, 70 kDa	0.177
YBL074c	355	*AAR2*	MATa1-mRNA splicing factor	0.143
YBL072c	200i	*RPS8*	ribosomal protein S8.e	0.746
YBL068w	355	*[PRPS3]*	probable ribose-phosphate pyrophosphokinase	0.191
YBL066c	1057		probable regulatory Zn-finger protein	0.137
YBL064c	261		homologue to thiol-specific antioxidant	0.206
YBL063w	1111	*KIP1*	kinesin-related protein	0.143
YBL061c	696	*SKT5*	probable Ca^{2+} binding protein (prenylated)	0.113
YBL056w	468		probable phosphoprotein (Ser/Thr) phosphatase	0.177
YBL054w	525		homologue to myb transforming proteins	0.142
YBL050w	292i	*SEC17*	transport vesicle fusion protein	0.158
YBL047c	1381		cytoskeletal-related transport protein, Ca^{2+} binding	0.196
YBL045c	457	*COR1*	ubiquinol−cytochrome *c* reductase	0.293
YBL042c	639		FUR4 homologue, uracil transport protein	0.168
YBL041w	241	*PRS3*	proteasome subunit 3	
YBL040c	219i	*ERD2*	ER lumen protein retaining receptor	0.181
YBL039c	579	*URA7*	cytidine triphosphate synthase	0.308
YBL038w	232	*[MRPL16]*	probable mitochondrial ribosomal protein L16	0.154
YBL036c	257		homologue to twitching motility protein	0.237
YBL033c	345	*[RIB1]*	probable GTP cyclohydrolase II	0.111
YBL030c	318	*AAC2*	mitochondrial ATP/ADP carrier	0.537
YBL027w	189i	*RPL19*	ribosomal protein L19.e	0.707
YBL026w	95i		probable snRNP-related protein	0.168
YBL024w	684		probable proliferating-cell nucleolar antigen (human p120)	0.270
YBL023c	868	*MCM2*	transcription factor	0.175
YBL022c	1133	*PIM1*	mitochondrial ATP-dependent lon-like serine proteinase	0.186
YBL021c	144	*HAP3*	transcription factor	0.091
YBL017c	1579	*PEP1*	carboxypeptidase Y sorting precursor	0.163
YBL016w	353	*FUS3*	protein kinase (cell cycle and cell fusion)	0.110
YBL015w	526	*ACH1*	acetyl-CoA hydrolase	0.217
YBL013w	393		probable met-tRNA formyltransferase, mitochondrial	0.073
YBL008w	840	*HIR1*	regulator of histone gene transcription	0.128
YBL007c	1244	*SLA1*	cytoskeleton assembly control protein	0.189
YBL005w	976	*PDR3*	pleiotropic drug resistance protein 3	0.147
YBL000	440	*TY1A*		0.140
YBL000	1755t	*TY1B*		0.140
YBL003c	132	*HTA2*	histone H2A.2	0.542
YBL002w	131	*HTB2*	histone H2B.2	0.562
YBR001c	780	*NTH2*	α,α-trehalase	0.123
YBR003w	473	*COQ1*	hexaprenyl-pyrophosphate synthase precursor	0.140
YBR006w	435		probable aldehyde dehydrogenase	0.187
YBR008c	548		probable benomyl/methotrexate resistance protein	0.155
YBR009c	103	*HHF1*	histone H4	0.733
YBR010w	136	*HHT1*	histone H3	0.621
YBR011c	287	*IPP1*	inorganic pyrophosphatase	0.620
YBR000	440	*TY1A*		0.150
YBR000	1756t	*TY1B*		0.140
YBR014c	203		glutaredoxin homologue	0.154
YBR015c	597	*TTP1*	type II transmembrane protein	0.169
YBR018c	366	*GAL7*	galactose-1-phosphate uridylyltransferase	0.221
YBR019c	699	*GAL10*	UDP-glucose-4-epimerase	0.185
YBR020w	528	*GAL1*	galactokinase	0.194
YBR021w	633	*FUR4*	uracil transport protein	0.186
YBR023c	1165	*CHS3*	chitin synthase 3	0.166
YBR024w	301	*[SCO2]*	SCO1 protein homologue	0.155
YBR025c	394		probable purine nucleotide binding protein	0.566
YBR026c	380	*[MRF1]*	probable (mitochondrial) ssDNA binding protein	0.165

Table I. *Continued*

ORF	Size (aa)[a]	Gene[b]	Function	CAI
YBR028c	525		probable Ser/Thr-specific protein kinase	0.149
YBR031w	362	*RPL2*	ribosomal protein L2A	0.802
YBR033w	919		probable regulatory Zn-finger protein	0.115
YBR034c	348	*[ODP1]*	ORF adjacent to PDX3	0.267
YBR035c	228	*PDX3*	pyridoxamine-phosphate oxidase	0.242
YBR036c	410	*CSG2*	Ca^{2+}-dependent regulatory protein	0.142
YBR037c	295	*SCO1*	cytochrome oxidase assembly protein precursor	0.110
YBR038w	963	*CHS2*	chitin synthase 2	0.172
YBR039w	311		probable H^+-transporting ATPase [F(1)-ATPase γ]	0.337
YBR041w	623		probable AMP binding protein	0.186
YBR042c	397		probable membrane-bound small GTPase	0.132
YBR043c	689		probable pleiotropic resistance protein	0.124
YBR044c	573		homologue to mitochondrial chaperonin hsp60	0.126
YBR046c	334		homologue to quinone oxidoreductase (*E.coli*)	0.156
YBR048w	156i	*RPS18B*	ribosomal protein S11.e.B	0.733
YBR049c	810	*REB1*	DNA binding regulatory protein	0.199
YBR052c	210		homologue to Trp repressor binding protein (*E.coli*)	0.150
YBR054w	344	*[YRO2]*	homologue to HSP30 heat-shock protein	0.456
YBR055c	899	*PRP6*	pre-mRNA splicing factor	0.126
YBR056w	501		homologue to glucan-1,3-β-glucosidase	0.202
YBR059c	1108		probable protein kinase	0.145
YBR060c	620	*RRR1*	origin recognition complex, 72 kDa subunit	0.140
YBR061c	310		homologue to ftsJ protein (*E.coli*)	0.143
YBR063c	404		probable phosphopantethein binding protein	0.131
YBR066c	220		probable Zn-finger protein	0.124
YBR067c	210	*TIP1*	temperature shock-inducible protein precursor SRP1/TIP1	0.449
YBR068c	609		probable amino acid transport protein	0.157
YBR069c	619		probable amino acid transport protein	0.151
YBR072w	214	*HSP26*	heat-shock protein, 30 kDa	0.337
YBR073w	958		probable RAD protein, DNA repair helicase	0.131
YBR074w	413		homologue to aminopeptidase Y	0.123
YBR078w	226i		homologue to sporulation-specific protein SPS2	0.614
YBR080c	758	*SEC18*	vesicular fusion protein	0.192
YBR081c	1332	*SPT7*	probable transcription factor, suppressor of Ty transcription	0.154
YBR082c	148i	*UBC4*	ubiquitin conjugating enzyme E2, 16 kDa subunit	0.313
YBR083w	486	*TEC1*	Ty transcription activator	0.120
YBR084w	975	*MIS1*	C1-tetrahydrofolate synthase precursor, mitochondrial	0.207
YBR085w	307	*AAC3*	mitochondrial ATP, ADP carrier	0.198
YBR086c	946		probable transmembrane protein	0.175
YBR087w	354		replication factor RFC3 homologue	0.151
YBR088c	258	*POL30*	proliferating cell nuclear antigen	0.256
YBR091c	109	*MRS5*	nuclear protein involved in mitochondrial intron splicing	0.067
YBR092c	467	*PHO3*	acidic phosphatase, constitutive	0.353
YBR093c	467	*PHO5*	acidic phosphatase, repressible	0.460
YBR097w	1454	*VPS15*	protein kinase, vacuolar transport	0.134
YBR104w	329	*[YMC2]*	mitochondrial carrier protein	0.119
YBR108w	848		probable transcription factor	0.106
YBR109c	147	*CMD1*	calmodulin	0.219
YBR110w	449	*ALG1*	α-mannosyltransferase	0.140
YBR111c	231	*[YSA1]*	homologue to *Drosophila* serendipity protein	0.246
YBR112c	966	*SSN6*	transcription regulatory protein	0.161
YBR114w	790	*RAD16*	radiation repair protein, putative DNA helicase	0.162
YBR115c	1392	*LYS2*	α-aminoadipate reductase	0.212
YBR117c	681	*TKL2*	transketolase 2 (EC 2.2.1.1)	0.168
YBR118w	458	*TEF2*	translational elongation factor α-1	0.875
YBR119w	298i	*MUD1*	U1snRNP-specific A protein	0.112
YBR120c	162	*CBP6*	cytochrome *b* pre-mRNA processing protein 6	0.126
YBR121c	667	*[GRS1]*	probable glycyl-tRNA synthase	0.413
YBR122c	196	*MRPL36*	mitochondrial ribosomal protein YmL36	0.173
YBR123c	649	*TFC1*	transcription factor TFIIIC, 95 kDa subunit	0.135
YBR125c	393		probable phosphoprotein phosphatase	0.128
YBR126c	495	*TPS1*	α,α-trehalose-phosphate synthase (*CIF1*)	0.187
YBR127c	517	*ATPvs*	H^+-transporting ATPase, vacuolar	0.390
YBR132c	596		probable amino acid transport protein	0.142
YBR135w	150	*CKS1*	CDC28 kinase complex, regulatory subunit	0.143
YBR136w	2368		probable phosphatidyl inositol kinase	0.136
YBR139w	508		probable serine-type carboxypeptidase	0.150
YBR140c	3092	*IRA1*	GTPase-activating protein of the RAS-cAMP pathway	0.139
YBR142w	773		probable DEAD box RNA helicase	0.182
YBR143c	437	*SUP1*	ominipotent suppressor protein of nonsense codons	0.333
YBR145w	351	*[ADH5]*	alcohol dehydrogenase	0.253
YBR146w	278	*[MRPS9]*	probable mitochondrial ribosomal protein S9	0.137

Table I. *Continued*

ORF	Size (aa)[a]	Gene[b]	Function	CAI
YBR149w			probable aldehyde reductase	
YBR150c			probable regulatory Zn-finger protein	
YBR153w		*[RIB7]*	riboflavin biosynthetic protein	
YBR154c		*RPB5*	RNA polymerases I, II and III, 27 kDa subunit	
YBR160w		*CDC28*	cell division control protein	
YBR161w			SUR1 homologue	
YBR164c		*ARF3*	GTP binding ADP ribosylation factor 3	
YBR166c		*TYR1*	prephenate dehydrogenase ($NADP^+$)	
YBR169c		*SSE2*	heat-shock protein, 70 kDa	
YBR170c		*NPL4*	suppressor of SEC63, ER translocation component	
YBR171w		*HSS1*	ER translocation complex subunit SEC66	
YBR172c		*SMY2*	kinesin-related protein suppressing myosin defects	
YBR175w			probable GTP binding protein	
YBR176w			probable 3-methyl-2-oxobutanoate hydroxymethyltransferase	
YBR177c			probable membrane receptor	
YBR179c			probable purine nucleotide binding protein	
YBR180w			probable drug resistance protein	
YBR181c		*RPS10I*	ribosomal protein S6.e	
YBR182c			probable DNA binding transcription factor	
YBR186w			probable ATP binding protein	
YBR187w			probable membrane protein	
YBR189w		*SUP46*	suppressor, ribosomal protein S13	
YBR191w		*URP1*	ribosomal protein L21.e	
YBR192w		*RIM2*	probable carrier protein, mitochondrial	
YBR195c		*MSI1*	multicopy suppressor of IRA1, G-protein	
YBR196c		*PGI1*	phosphoglucose isomerase	
YBR198c			probable transcription-associated factor protein	
YBR199w		*KTR4*	α-1,2-mannosyltransferase homologue	
YBR200w		*BEM1*	bud emergence mediator	
YBR202w			MCM3 protein homologue	
YBR204c			probable serine-active lipase, peroxisomal	
YBR205w		*[KTR3]*	KTR3 protein	
YBR207w			probable membrane protein	
YBR208c		*DUR1,2*	urea carboxylase	
YBR212w		*RBP1*	RNA binding protein, NGR1	
YBR213w		*MET8*	effector of PAPS reductase and sulfite reductase	
YBR215w		*HPC2*	cell cycle regulatory protein	
YBR218c		*PYC2*	pyruvate carboxylase 2	
YBR221c		*PDB1*	pyruvate dehydrogenase (lipoamide), β-chain	
YBR222c			probable AMP binding protein	
YBR227c			homologue to ATP binding protein clpX (*E.coli*)	
YBR229c			homologue to α-1,4-glucosidase	
YBR233w			homologue to human hnRNP complex K protein	
YBR236c		*ABD1*	protein with mutational synergism related to BEM1	
YBR237w		*PRP5*	pre-mRNA processing protein, RNA helicase	
YBR239c			probable Zn-finger protein	
YBR240c			probable Zn-finger protein	
YBR241c			probable sugar transport protein	
YBR242w			probable ATP/GTP binding protein	
YBR243c		*TUR1*	UDP-*N*-acetylglucosamin-1-phosphate transferase	
YBR244w			probable glutathione peroxidase	
YBR245c			homologue to SNF2/SWI2 DNA binding regulatory protein	
YBR248c		*HIS7*	glutamine amido transferase	
YBR249c		*ARO4*	2-deoxy-3-deoxyphosphoheptanoate aldolase	
YBR251w		*[MRPS5]*	probable mitochondrial ribosomal protein S5	
YBR252w		*DUT1*	mitochondrial dUTP pyrophosphatase	
YBR254c			probable membrane protein	
YBR256c		*RIB5*	riboflavin synthase α-chain	
YBR263w		*[SHMT1]*	serine hydroxymethyltransferase	
YBR264c			probable small GTP binding protein	
YBR265w			probable membrane protein	
YBR266c			probable membrane protein	
YBR267w			probable Zn-finger protein (C_2H_2 type)	
YBR268w		*MRPL37*	probable mitochondrial ribosomal protein L37	
YBR270c			probable ATP/GTP binding protein	
YBR274w			probable protein kinase (cytokine receptor family)	
YBR275c		*RIF*	RAP1-interacting regulatory protein	
YBR276c			probable tyrosine-specific protein phosphatase	
YBR278w		*DPB3*	DNA-directed DNA polymerase, chain C	
YBR281c			probable G-protein, β-transducin type	
YBR282w		*MRPL27*	mitochondrial ribosomal protein YmL27	
YBR283c			probable SEC61 homologue	

Table I. *Continued*

ORF	Size (aa)[a]	Gene[b]	Function	CAI
YBR286w	564		aminopeptidase Y	0.331
YBR289w	905	*SNF5*	general transcriptional activator	0.119
YBR291c	299		probable mitochondrial carrier protein	0.148
YBR293w	474		probable multidrug resistance protein	0.087
YBR294w	859		probable sulfate transport protein	0.130
YBR295w	1216	*PCA1*	P-type copper-transporting ATPase	0.146
YBR296c	574		homologue to phosphate-repressible phosphate permease	0.254
YBR297w	468	*MAL3R*	maltose fermentation regulatory protein	0.123
YBR298c	614	*MAL3T*	maltose permease	0.164
YBR299w	584	*MAL3S*	maltase	0.227

Detailed lists of all chromosome II ORFs (including GC content and CAI values), intron-containing genes, tRNA genes and proteins with putative membrane spans can be found in tables deposited together with the sequence data (see Acknowledgements).
[a]'i' indicates an intron-containing ORF; t indicates TYB protein produced with an internal +1 frameshift.
[b]Suggested gene names are in parentheses.

Table II. Related genes from chromosome II

Gene/ORF on chromosome II	Related gene/ ORF on other chromosome[a]	Functional description
HTA2	*HTA1* (4R)	histones H2A
HTB2	*HTB1* (4R)	histones H2B
HHT1	*HHT2* (4)	histones H3
HHF1	*HHF2* (4)	histones H4
PYC2	*PYC1* (7)	pyruvate carboxylases
TKL2	*TKL1*	transketolases
TEF2	*TEF1* (16R)	translational elongation fators α
YMC2	*YMC1* (16)	mitochondrial carrier proteins
MCM2	*MCM3* (5L)	transcription factors
IRA1	*IRA2* (15L)	regulators in the cAMP–RAS pathway
KIP1	*KIP2* (16L)	kinesin-related proteins
NTH2	*NTH1* (4)	trehalases
YBR078w	*SPS2*	sporulation-specific proteins
YBR028c	*YKR2* (13R)	protein kinases
YRO2	YCR20c (3)	seven transmembrane proteins
RPS8B	*RPS8A* (5)	ribosomal proteins
RPS18B	*RPS18A*	ribosomal proteins

Gene/ORF on chromosome II	Related gene/ ORF on chromosome II	Functional description
AAC2	*AAC3*	mitochondrial ADP/ATP translocators
CHS2	*CHS3*	chitin synthases
SCO1	*SCO2*	cytochrome oxidase assembly factors
MCM2	YBR202w	probable transcription factors
KTR3	YBR199w	probable mannosyltransferases
RAD16	YBR073w	probable radiation repair proteins
YMC2	YBR291c	probable mitochondrial carrier proteins
YBL088c	YBR136w	probable phosphatidyl inositol kinases
YBR041w	YBR222c	probable AMP binding proteins
YBR068c	YBR069c	probable amino acid transporters
YBR068c	YBR132c	probable amino acid transporters
YBR008c	YBR043c	probable multidrug resistance proteins
YBR008c	YBR293w	probable multidrug resistance proteins
YBL056w	YBR125c	probable phosphoprotein phosphatases

[a]Where known, the chromosomal location is indicated in parentheses.

be presented elsewhere (H.Feldmann *et al.*, manuscript in preparation).

'Redundant' sequences in chromosome II

Several algorithms were used to analyse chromosome II for the occurrence of sequences demonstrating high similarity, both at the nucleotide and the amino acid levels (H.Feldmann *et al.*, manuscript in preparation). The results not only confirm earlier notions (e.g. Dujon *et al.*, 1994) that the degree of internal genetic redundancy in the yeast genome must be high, but also provide a more detailed picture of this phenomenon (Table II). First, in chromosome II we find quite a number of genes that are functionally well characterized and have highly homologous counterparts on other chromosomes. Surprisingly, a second category that we encountered is represented by a number of highly homologous genes on chromosome II itself. Several of these are functionally characterized, while for others only probable functions are predicted. Additionally, 20 of the chromosome II ORFs of unknown function have homologues among ORFs also of unknown function and lying on other systematically sequenced chromosomes or on chromosome II itself.

By applying the program PYTHIA (Milosavljevic and Jurka, 1993) to search for simple repeats, we detected at least 12 sets of regularly repeated trinucleotides along chromosome II (H.Feldmann *et al.*, manuscript in preparation). Concomitant examination of the chromosome II ORFs revealed that these triplets represent repetitious codons for particular amino acids, such as asparagine, glutamine, arginine, aspartic acid, glutamic acid, proline and serine, thus forming homopeptide stretches. Searches in the databases show that there are numerous proteins containing homopeptides built from these amino acids, sometimes of considerable size, in yeast and other organisms. Although the role of such homopeptides is not well defined, it appears that they constitute specific domains enabling the respective proteins to fulfil specific functions.

Organization of the chromosome

The gene density in chromosome II is as high as found previously with chromosomes III and XI: ORFs occupy on average 71.9% of the sequence of chromosome II, excluding the ORFs contributed by the Ty elements. The

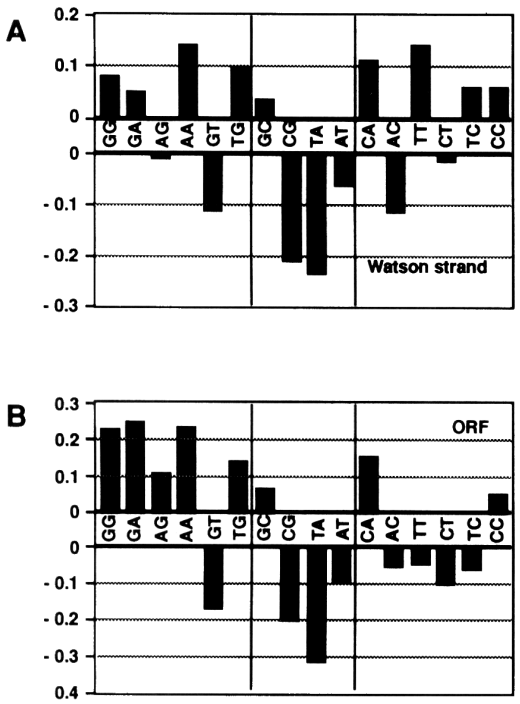


Fig. 2. Compositional symmetry/asymmetry of chromosome II and its constituent elements. Relative deviations of dinucleotide frequencies [(observed − expected)/expected] are shown as vertical bars (expected frequencies are calculated from mononucleotide frequencies). Complementary dinucleotide pairs have been arranged in mirror image to help visualize compositional symmetry or asymmetry. Self-complementary dinucleotides are at the centre. (**A**) Data for the entire chromosome sequence, calculated from the Watson strand. (**B**) Data for ORFs only, calculated in each case from the coding strand.

average ORF size is 475 codons (1425 bp). The mean sizes of inter-ORF regions are 647 bp for 'divergent promoters' and 414 bp for 'convergent terminators', while 'promoter–terminator combinations' are 662 bp in length on average. These values are similar to those reported for chromosome XI. The average base composition of chromosome II is 38.3% GC, a value close to that of chromosomes III (38.5%) and XI (38.1%). As expected, the coding regions have a higher GC content on average (39.6%) than the non-coding regions (35.1%). In sliding windows, coding regions may be discriminated from intergenic regions because 'transitions' in GC content are rather sharp at their borders (data not shown). An almost symmetrical distribution of dinucleotide frequencies over the entire chromosome is apparent (Figure 2A), whereas the base composition of ORFs shows a significant excess of homopurine pairs on the coding strand (Figure 2B). These data are also similar to those obtained for chromosome XI (Dujon *et al.*, 1994).

Contrary to what has been observed in chromosomes III and XI, chromosome II shows a significant bias of coding capacity between the two strands (Table III). Whereas in the two other chromosomes the coding capacity is nearly symmetrical on the two strands, in chromosome II the coding capacity on the 'Crick' strand exceeds that of the 'Watson' strand by 33%. This bias remains virtually unchanged when the 'questionable' ORFs are excluded from the calculations. At present, the significance of this phenomenon is not known; more detailed analyses, e.g. of biased codon usage in the two strands from chromosome II and others, may give further clues. For the putative membrane proteins, the same asymmetrical distribution of ORFs is observed as for the rest of the ORFs. Remarkably, the 'membrane' ORFs appear to occur in clusters on chromosome II and occupy 46.5% of the total coding capacity.

Regional variations of base composition with similar amplitudes were noted along chromosomes III (Sharp and Lloyd, 1993) and XI (Dujon *et al.*, 1994), with major GC-rich peaks in each arm. The analysis of chromosome XI revealed an almost regular periodicity of the GC content, with a succession of GC-rich and GC-poor segments of ~50 kb each; a further interesting observation was that the compositional periodicity correlated with local gene density. Profiles obtained from a similar analysis of chromosome II again show these phenomena (Figure 3). GC-poor peaks coinciding with relatively low gene densities are located at the centromere (around coordinate 230) and at both sides of the centromere with a periodicity of ~110 kb. These minima are more pronounced around coordinates 120, 340 and 560, while they are less so at coordinates 450 and 670. Remarkably, most of the tRNA genes reside in GC-poor 'valleys' and the Ty elements eventually became integrated into these regions. We have also analysed chromosome II for the occurrence of simple repeats, potential ARS elements and putative regulatory signals. Some of the results will be discussed below and a detailed evaluation will be presented elsewhere (H.Feldmann *et al.*, manuscript in preparation).

Comparison of the physical and genetic maps

The genetic map of *S.cerevisiae* (Mortimer *et al.*, 1992) assigned 92 genes or markers to chromosome II; 71 were located on a linear array and 21 remained unmapped. Figure 4 shows a comparison of this map with the physical map deduced from the complete sequence. In all, 42 of the mapped genes and 11 of the unmapped genes could be unambiguously assigned to an ORF or a tRNA gene of the present sequence on the basis of previous partial sequence data, use of probes or gene function; the assignment of four genes remains tentative. Thus, a total of 35 genes or markers remains unassigned on the physical map of chromosome II at present. These include several genes [*pet9* (= *AAC1*); *pdr7* (= *pdr4*); *RNA14*; *rpc19*] whose sequences are known but which do not appear in chromosome II of strain αS288C. This is also true for the *MEL1*, *SUC3* and *MGL2* genes. *CDC25* had been mapped to chromosome II erroneously but has been located to chromosome XII (Johnson *et al.*, 1987). Two suppressors, SUP87 and SUP72, may correspond to the tRNA genes found between coordinates ~320 and ~345 on chromosome II. The order of the genes positioned on chromosome II by genetic and physical mapping is largely the same, with some exceptions. No gross translocations or inversions on the genetic map, as found with chromosome XI (Dujon *et al.*, 1994), were observed here.

Discussion

The network approach to systematic sequencing of the yeast genome started with chromosome III and has been

Table III. Organization of ORFs along yeast chromosomes II, XI and III

Chromosome	W strand				C strand				Ratio of coding capacity (C/W)
	coding		ORFs		coding		ORFs		
	%	aa		average length (aa)	%	aa	*n*	average length (aa)	
II 807 188 bp (overlapping ORFs and Tys excluded)	30.3	81 525	177	475.6	40.5	108 929	204	534.0	1.336
XI 666 448 bp (overlapping ORFs excluded)	36.3	80 742	163	495.3	34.8	77 231		518.3	0.960
III 315 287 bp (overlapping ORFs and Tys excluded)	32.4	34 037	79	430.8	35.5	37 162	104	357.3	1.092

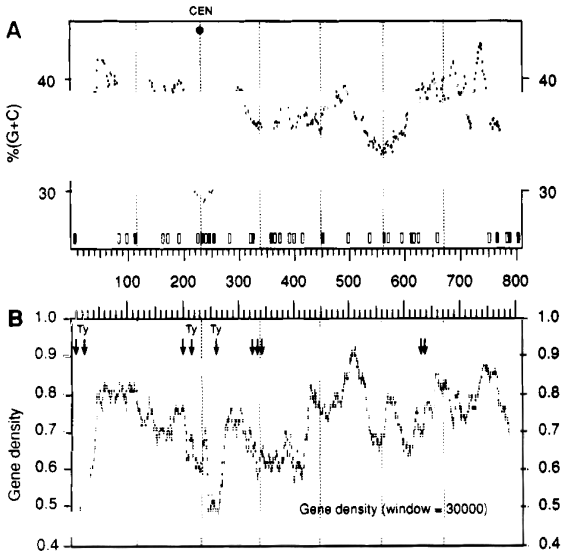


Fig. 3. Compositional variation and gene density along chromosome II. (**A**) Compositional variation along chromosome II calculated as in Dujon *et al.* (1994). Each point represents the average GC composition calculated from the silent positions only of the codons of 15 consecutive ORFs. Similar slopes were obtained when the GC composition was calculated from the entire ORFs or from the inter-ORF regions, or when averages of 13–30 elements were plotted (results not shown). The location of perfect ARS consensus sequences is indicated by the rectangles; filled boxes, ARS patterns fulfilling criteria attributed to functional replication origins (see text). (**B**) Gene density along chromosome II. Gene density is expressed as the fraction of nucleotides within ORFs versus the total number of nucleotides in sliding windows of 30 kb (increments are 1 kb). Similar results were obtained for sliding windows of 20 or 50 kb. The arrows indicate the locations of tRNA genes; tRNA genes associated with complete Ty elements are marked by 'Ty'. The vertical lines have been introduced at a regular spacing of 110 kb, starting from the centromere (coordinate 230) and taking the most prominent troughs at coordinates 120 and 560 as references.

continued successfully with chromosomes XI and II. In the two latter cases, cosmid libraries and fine-resolution physical maps of the respective chromosomes from the same unique strain were first constructed to facilitate sequencing and assembly of the sequences. It should be noted that, by convention, in all laboratories engaged in sequencing the yeast genome, the strain αS288C, or isogenic derivatives thereof, were chosen as the source of DNA because they have been fairly well characterized and employed in many genetic analyses. For cosmid cloning of chromosome II DNA, we employed a vector which carries a yeast marker and therefore can be used in direct complementation experiments (Stucka and Feldmann, 1994). Furthermore, these cosmid clones turned out to be stable for many years under usual storage conditions. Like chromosome XI, the physical map of chromosome II has been constructed without reference to the genetic map and has been confirmed by the final sequence.

The comparison of the physical and genetic maps of chromosome II (Figure 4) shows that most of the linkages have been established to give the correct gene order; however, in many cases the relative distances derived from genetic mapping are rather imprecise. The obvious imprecisions of the genetic map may be due to the fact that different yeast strains have been used to establish the linkages. It is possible that some strains employed in genetic mapping experiments show inversions or translocations which then might contribute to discrepancies between physical and genetic maps, as considered in the case of chromosome XI. However, a more wide-spread phenomenon that may lead to imprecisions in the genetic maps are strain polymorphisms caused by the Ty elements. Detailed information on strain differences resulting from Ty insertions and/or deletions is available for chromosome II, where we can compare the complete Ty patterns from strains αS288C and C836, and local patterns from two other strains, YNN13 and M1417-c (Stucka, 1992). In αS288C, a Ty2 element is associated with the $tRNA^{Phe}$ gene (coordinate ~24), while it is absent in C836 at this position; instead, a Ty2 has been inserted into a 'solo' δ sequence near the $tRNA^{Leu4}$ gene (coordinate ~3.6). The Ty1 element next to *IPP1* (coordinate ~251) is missing in C836, whereas a Ty3 element is found at the equivalent position in YNN13. In C836, the $tRNA^{Cys}$ and $tRNA^{Glu3}$ genes bracket a Ty1 element, which is absent at this location (coordinate ~638) in αS288C; in M1417-c, the Ty1 element and the τ sequence, the LTR of a Ty4 element, are missing. It may be noted that the sequences around the elements are well conserved among all these strains. Many more examples of this kind can be found in the literature. Altogether, this reveals a substantial plasticity of the yeast genome around tRNA gene loci which appear to be the preferred target sites for Ty transpositions (e.g. Hauber *et al.*, 1988; Feldmann, 1988). Experimentally, this latter phenomenon has been proven for yeast chromosome III (Ji *et al.*, 1993). Since these regions do not

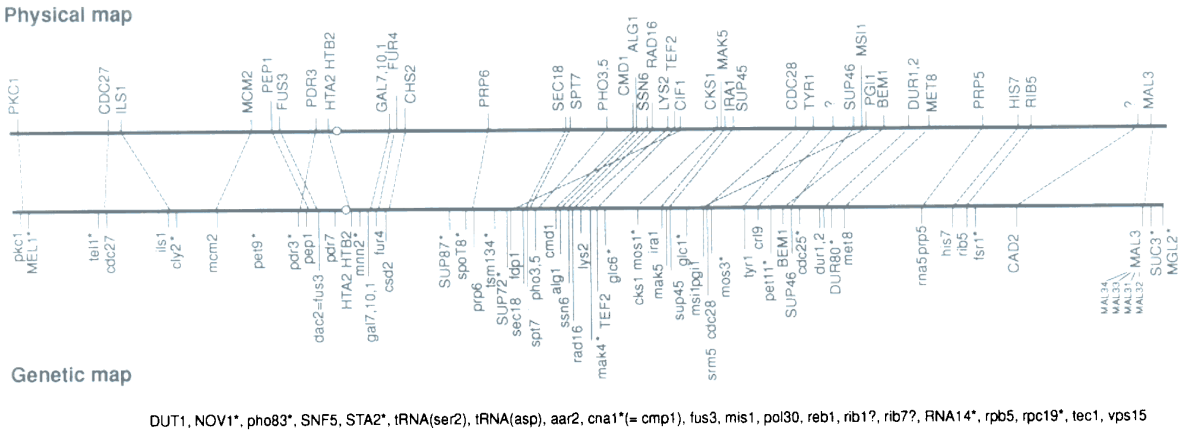


Fig. 4. Comparison of the genetic and physical maps of yeast chromosome II. The genetic map (lower part; 71 mapped genes or markers) is redrawn from *Genetic and Physical Maps of Saccharomyces cerevisae* (edition 11; Mortimer *et al.*, 1992). The unmapped genes are listed beneath. The physical map (upper part) derived from the complete sequence of chromosome II has been drawn to the same scale. The circle indicates the position of the centromere. Genes or markers for which no ORF or RNA gene has been assigned on the physical map as yet are indicated by an asterisk; the assignation of genes marked by '?' is only tentative. Numerous other genes described in this work were not assigned previously to a chromosome (compare Figure 1 and Table I).

contain any special DNA sequences, the region-specific integration of the Ty elements may be due to specific interactions of the Ty integrase(s) with the transcriptional complexes formed over the intragenic promoter elements of the tRNA genes or triggered by positioned nucleosomes in the 5′ flanking regions of the tRNA genes (Feldmann, 1988; Ji *et al.*, 1993). In any case, the Ty integration machinery can detect regions of the genome that may represent 'safe havens' for insertion, thus guaranteeing survival of both the host and the retroelement.

About two thirds of the genes or markers mapped to chromosome II could be assigned to an ORF or an RNA gene on the basis of previous sequence data, the use of probes or gene function. At present, 35 genes or markers remain unassigned. Further assignments must await the correlation of our sequence data and new information that will become available in the literature. Three genes mapped on chromosome II, *MEL1*, *SUC3* and *MGL2*, are absent from the strain αS288C. *MEL* and *SUC* genes, which are involved in carbohydrate metabolism, have been found previously as subtelomeric repeats in several yeast strains. The presence of multiple gene copies could be attributed to selective pressure induced by human domestication, but it appears that they are largely dispensable in laboratory strains (such as αS288C) which are no longer used in fermentation processes. A comparison at the molecular level of αS288C with brewer's yeast strain C836 clearly shows that the *SUC* genes are present on chromosome II of the latter strain (Stucka, 1992). Non-homologous recombination processes may account for the duplication of these and other genes residing in subtelomeric regions (Michels *et al.*, 1992), reflecting the dynamic structure of yeast telomeres in general (Louis *et al.*, 1994). Altogether, the experience gained from the yeast chromosomes sequenced so far shows that genetic maps provide valuable information but that in some cases they may be misleading. Therefore, independent physical mapping and eventual determination of the complete sequences is needed to unambiguously delineate all genes along chromosomes. At the same time, the differences found between various yeast strains demonstrate the need to use one particular strain as a reference system.

As observed in chromosome XI (Dujon *et al.*, 1994), the compositional periodicity in chromosome II correlates with local gene density, as is the case in more complex genomes in which isochores of composition are, however, much larger (Bernardi, 1993). Although the fairly periodic variation of base composition is now evident for the three sequenced yeast chromosomes, its significance remains unclear. Several explanations for the compositional distribution and the location-dependent organization of individual genes have been offered (Bernardi, 1993; Dujon *et al.*, 1994), some of which could be tested experimentally. For example, transcription mapping of a whole chromosome could give a clue as to whether such rules influence the expression of genes. Furthermore, long-range determination of DNase I-sensitive sites may be used to find a possible correlation between compositional periodicity and chromatin structure along a yeast chromosome. Similarly, knowledge of the sequence provides a basis to search for potential ARS elements, thus enabling functional replication origins to be sorted out experimentally. In Figure 3 we have listed the location of 36 ARS elements which completely conform to the 11 bp degenerate consensus sequence (Newlon, 1988; van Houten and Newlon, 1990). Several of these were found associated at their 3′ extensions with imperfect (one to two mismatches) parallel and/or antiparallel ARS sequences or putative ABF1 binding sites, reminiscent of the elements reported to be critical for replication origins (Bell and Stillman, 1992; Marahrens and Stillman, 1992). Remarkably, these patterns are found within the GC valleys, suggesting that functional replication origins might preferably be located in AT-rich regions. A similar correlation is apparent from an analysis of chromosome XI (data not shown) and, more convincingly, when the distribution of functional replication origins mapped in 200 kb of chromosome III (Dershowitz and Newlon, 1993) is compared with the GC profiles of

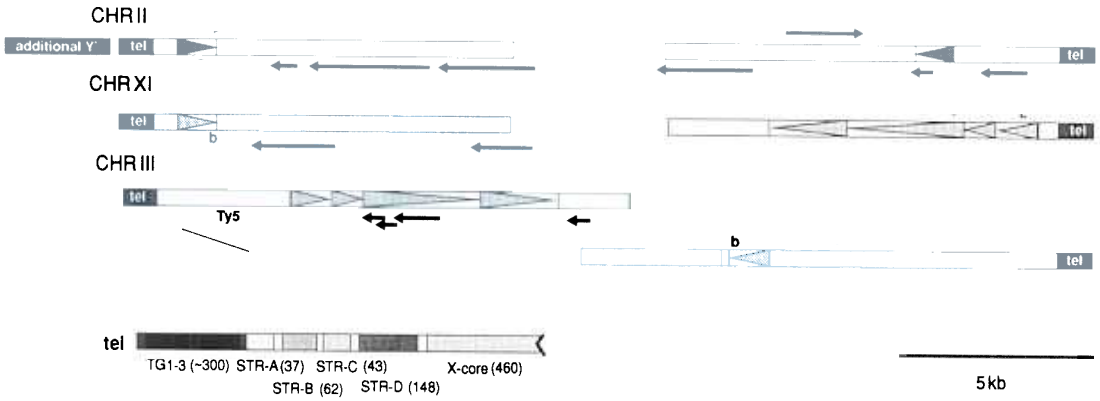


Fig. 5. Organization of telomeric regions. The 10–13 kb from each end of the sequences of chromosomes II, XI and III are represented by the mosaic boxes. Repetitious sequences of different types (a, ~800 bp; b, ~1 kb; c, four consecutive regions of ~1.1, 0.8, 3.0 and 2.0 kb, respectively are indicated by the triangular segments within the boxes. The telomere regions (tel) are shown as black boxes. They conform to the consensus pattern described by Louis *et al.* (1994), consisting of a variable number of $TG_{(1-3)}$ repeats, four types of subtelomeric repeats (STRs) and an X c segment (see insert, not drawn to scale). The locations of ORFs are indicated by arrows above ('Watson' strand) and below ('Crick' strand) each chromosome panel.

this chromosome (Sharp and Lloyd, 1993). The spacing of ~100–110 kb of the AT-rich regions is compelling, because this is also the observed spacing between active origins (for a review see Fangman and Brewer, 1992). Of course, functional ARS elements have yet to be defined for chromosomes II and XI, and also for the remainder of chromosome III. In this context, it would be interesting to see whether the putative origins of replication and the chromosomal centromeres in chromosomes II and XI might maintain specific interactions with the yeast nuclear scaffold (Amati and Gasser, 1988). It is not surprising that ARS elements possibly functioning as replication origins occur next to the histone genes in chromosome II (located at both sides of the centromere), but it is puzzling that the majority of the tRNA genes are flanked by such ARS elements. In all of the yeast chromosomes sequenced thus far, ARS elements located in the subtelomeric regions are closely associated with specific sites for origin binding factors (Eisenberg *et al.*, 1988; Estes *et al.*, 1992).

A comparison of the telomere regions of chromosome II with those of chromosomes III and XI (Figure 5) revealed the characteristic subtelomeric structures ('tel') found in all yeast chromosomes (Louis *et al.*, 1994). As inferred from our mapping data and the detailed analysis of the yeast telomeres (Louis *et al.*, 1994), chromosome II carries an additional 5.2 kb Y′ element at its left end; because of its particular structure, this element from chromosome II could not be cloned as yet. There are two Y′ classes, 5.2 and 6.7 kb in length, both of which include an ORF for a putative RNA helicase of as yet unknown function. Y′s show a high degree of conservation but vary among different strains, as well as within a single strain, with respect to their presence (Louis and Haber, 1992; Louis *et al.*, 1994). Experiments with the *est1* (ever shortening telomeres) mutants, in which telomeric repeats are progressively lost, have shown that the senescence of these mutants can be rescued by a dramatic proliferation of Y′ elements (Lundblad and Blackburn, 1993). Several additional functions have been suggested for these elements (for a review see Palladino and Gasser, 1994), such as extension of telomere-induced heterochromatin, protection of nearby unique sequences from its effects or a role in the positioning of chromosomes in the nucleus. Chromosome II might then offer an experimental system to address the functional significance of a particular Y′ element.

A comparison of the termini of chromosome II with those of chromosomes III and XI revealed that our chromosome II sequence not only extends into genuine telomere regions but that these three chromosomes share extended similarities in their subtelomeric regions by the occurrence of repetitious sequences of different types. While segments b and c (Figure 5) represent interchromosomal subtelomeric duplications (Dujon *et al.*, 1994), an ~800 bp sequence (Figure 5, segment a) is found as an inverted duplication near both termini of chromosome II. These duplicated regions contain ORFs, the putative products of which exhibit high similarity; but their functions remain unclear because no homologues of known function can be found in the databases.

A survey of previous sequence data and sequences obtained in the yeast sequencing programme suggests that there is a considerable degree of internal genetic redundancy in the yeast genome (Dujon *et al.*, 1994). Whereas an estimate of sequence similarity (both at the nucleotide and the amino acid levels) becomes predictive at this stage, it still remains difficult to correlate these values to functional redundancy because only in a limited number of cases have gene functions been defined precisely. Classic examples of redundant genes in yeast are the *MEL*, *SUC* and *MAL* genes that are found in the subtelomeric regions of several chromosomes. There is also a great variety of internal genes that appear to have arisen from duplications, as suggested by the analyses of chromosomes II and XI. In chromosome II, this concerns ~16% of the total ORFs, while this figure is estimated to be only 4% in chromosome XI. However, in these and other cases available from the literature, sequence similarities at the nucleotide level are generally restricted to the coding regions and do not extend into the intergenic regions. Thus, the corresponding gene products share high similarity in terms of amino acid sequence or sometimes are even identical; they may be functionally redundant but their expression will depend on the nature of the regulatory

elements. This has been demonstrated experimentally in numerous examples, prominent cases being the *PHO3* and *PHO5* genes located next to each other on chromosome II. Biochemical studies also revealed that in particular cases 'redundant' proteins can substitute each other, thus accounting for the fact that a large portion of single gene disruptions in yeast do not impair growth or cause abnormal phenotypes. This does not imply, however, that these 'redundant' genes were *a priori* dispensible. Rather, they may be designed to help adapt yeast cells to particular environmental conditions. These notions are of practical importance when carrying out and interpreting gene disruption experiments.

The availability of the complete sequence of chromosome II not only provides further insight into genome organization and evolution in yeast, but extends the catalogue of novel genes detected in this organism. Of general interest may be those that are homologues to genes that perform differentiated functions in multicellular organisms (YBL088c and YBR136w, homologues to phosphatidyl inositol kinases; YBL056w and YBR125c, homologues to phosphoprotein phosphatases; YBR274w, homologue to cytokine family protein kinase; YBR108w, probable homologue to *Drosophila mastermind*) or that might be of relevance to malignancy (YBL024w, homologue to p120, major human antigen associated with malignant tumours; YBR008c, YBR043c and YBR293w, probable multidrug resistance proteins; YBR295w, P-type copper transporting ATPase, homologue to Menkes and Wilson disease gene). Although the role of these genes has still to be clarified, yeast may offer a useful experimental system to identify their function. On the other hand, the wealth of information to be expected when the yeast genome sequencing programme progresses clearly demands that new routes are explored to investigate the functions of novel genes.

Materials and methods

Strains, plasmids, vectors and general methods

The following yeast strains were employed: C836, a diploid brewers yeast; αS288C (YGSC); FY73 (*MATα ura3-52 his3Δ200 GAL2*) derived from the strain αS288C (Thierry and Dujon, 1992). FY73/α224-pAF101 and FY73/α1001.1-pAF101 are transgenic strains derived from FY73 carrying the I-*Sce*I site within the right and left arm telomeric regions of chromosome II, respectively. pYc3030, a cosmid shuttle vector carrying the 2μ plasmid origin of replication and *HIS3* as a genetic marker (Hohn and Hinnen, 1980), was used for cosmid cloning throughout. Cosmids were propagated in *Escherichia coli* strains A490 and HB101. pAF101 is a plasmid carrying the *URA3* marker and the I-*Sce*I site (Thierry *et al.*, 1990). pEL61, a vector derived from pGEM-3Zf(−) by the insertion of a $(G_{1-3}T)_{300}$ repeat sequence and carrying *URA3* as a selective marker, was used for telomere cloning. Standard procedures were used in recombinant DNA techniques (Sambrook *et al.*, 1989). Yeast transformation was carried out by the procedure of Ito *et al.* (1983).

Chromosome II DNA

Construction of cosmid libraries, restriction mapping and cosmid distribution. A set of overlapping cosmid clones containing chromosome II inserts and issued from a genomic library of yeast strain αS288C was used as the DNA material. Similar to procedures described earlier in the construction of a chromosome II-specific cosmid library from strain C836 (Hauber *et al.*, 1988; Nelböck, 1988; Stucka, 1992), total DNA from αS288C was submitted to partial digestion with *Sau*3A, size-fractionated fragments cloned into the vector pYc3030, DNA samples packaged *in vitro* into lambda particles and *E.coli* A490 transfected with these. From a total of 200 000 clones, 3000 (about seven genome equivalents) were individually amplified and kept as an ordered cosmid library. DNA samples prepared from these clones were transferred to gridded filters and used for hybridizations (Stucka, 1992). A set of overlapping cosmid clones containing chromosome II inserts was established by (i) hybridizations of the ordered cosmid clones with chromosome II DNA; (ii) chromosomal walking and (iii) by using a collection of ~100 unique restriction fragments precisely mapped on C836 chromosome II as a reference library of 'sequenced tagged markers'. Restriction profiles were obtained for all clones by using at least the four restriction enzymes *Bam*HI, *Sal*I, *Xba*I and *Xho*I.

Right telomere region of chromosome II. pEL19B2, a plasmid containing the right telomere of chromosome II, was constructed following the procedure as described by Louis (1994). In brief, DNA from URA$^+$ transformants of αS288C transformed with pEL61 was prepared for CHEF gel and Southern analysis. Transgenes that had integrated the vector by homologous recombination within the right telomere of chromosome II were identified by probing CHEF blots before and after diagnostic *Not*I restriction. The DNA from a right telomere integrant was digested with *Bam*HI and ligated at low DNA concentration. This ligation was transformed into *E.coli* strain HB101 pyrF$^-$ using electroporation. One transformant, pEL19B2, carrying an ~14 kb insert from the right arm of chromosome II, was selected by diagnostic Southern hybridizations.

Telomere mapping

Physical mapping of the telomeres was performed using the I-*Sce*I chromosome fragmentation procedure described by Thierry and Dujon (1992). Yeast strain FY73 and the 1.1 kb *Bam*HI fragment from pAF101 (the 'pAF cassette' containing the *URA3* gene and the I-*Sce*I site; Thierry *et al.*, 1990) was used. The cassette was engineered to be integrated into defined sites of the left and right terminal-most cosmids, respectively. DNA isolated from the transgenes obtained in this way was then analysed using I-*Sce*I and a number of other appropriate restriction enzymes, resolved by pulsed-field gel electrophoresis and the lengths of the terminal-most restriction fragments determined by hybridization with diagnostic probes (H.Feldmann *et al.*, manuscript in preparation).

Sequence assembly, sequence analysis and quality controls

Sequence assembly in the single contracting laboratories was performed by a variety of software program packages. Completed contigs submitted to the Martinsried Institute for Protein Sequences (MIPS) were stored in a data library and assembled using the GCG software package 7.2 for the VAX (Devereux *et al.*, 1984). Special software developed for the VAX by Dr S.Liebl at MIPS was used to locate and translate ORFs (ORFEX and FINDORF), to retrieve non-coding intergenic sequences (ANTIORFEX) and to display various features of the sequence(s) on graphic devices (XCHROMO; an interactive graphics display program, version 2.0). The sequence has been interpreted using the following principles. (i) All intron splice site/branch-point pairs detected using specially defined patterns (Fondrat and Kalogeropoulos, 1994; K.Kleine and H.Feldmann, unpublished results) were listed. (ii) All ORFs containing at least 100 contiguous sense codons and not contained entirely in a longer ORF on either DNA strand were listed (this includes partially overlapping ORFs, indicated by asterisks in Figure 1). (iii) The two lists were merged and all intron splice site/branch-point pairs occurring inside an ORF but in opposite orientations were disregarded. (iv) Centromere and telomere regions, as well as tRNA genes and Ty elements or remnants thereof, were sought by comparison with a previously characterized dataset of such elements (K.Kleine and H.Feldmann, unpublished results) including the database entries provided in a tRNA/tRNA gene library (Steinberg *et al.*, 1993; retrieved from the EMBL ftp server). All sequences submitted by collaborating laboratories to the MIPS data library were subjected to quality controls similar to those performed in the work on chromosome XI (Dujon *et al.*, 1994). Sequence verifications were obtained from (i) the original overlaps between 33 contiguous segments (total of 40 037 bp); (ii) resequencing of selected segments (209 bp to 14.6 kb long; 2255 bp on average; total of 58 635 bp); and (iii) resequencing of suspected segments from designed oligonucleotide pairs (210–1530 bp long; 511 bp on average; total of 6646 bp).

Searches for similarity of proteins to entries in the databanks were performed by FastA (Pearson and Lipman, 1988), BlastX (Altschul *et al.*, 1990) and FLASH (Califano and Rigoutsos, 1993), in combination with the Protein Sequence Database of PIR International (release 41) and other public databases. Protein signatures were detected using the PROSITE dictionary (release 11.1; Bairoch, 1989). ORFs were considered to be homologues or to have probable functions when the

alignments from FastA searches showed significant similarity and/or protein signatures were apparent; at this stage of analysis, FastA scores <150 were considered insufficient to confidently assign function. Compositional analyses of the chromosome (base composition; nucleotide pattern frequencies, GC profiles; ORF distribution profiles, etc.) were performed using the X11 program package (C.Marck, unpublished results). For calculations of CAI and GC content of ORFs, the algorithm CODONS (Lloyd and Sharp, 1992) was used. Comparisons of chromosome II sequence with databank entries (EMBL databank, release 39; GenBank, release 83) were based on a new algorithm developed at MIPS by K.Heumann.

Acknowledgements

The Laboratory Consortium operating under contracts with the European Commission was initiated and organised by A.Goffeau. This study is part of the second phase of the European Yeast Genome Sequencing Project carried out under the administrative coordination of A.Vassarotti (DG-XII) and the Université Catholique de Louvain, and under the scientific responsibility of H.Feldmann, as DNA coordinator, and H.W.Mewes, as Informatics coordinator. We thank P.Mordant for accounting; A.Thierry for supplying pAF101 and yeast strain FY73; E.Louis for preparing pEL19B2; P.Jordan and W.G.Peng for support in computing; K.Heumann for providing a new algorithm for database searches; our colleagues for help and discussions; and J.Svaren for comments on the manuscript. Datasets containing nucleic acid and protein sequences in different standard database formats, including annotations and detailed tables, will be available through anonymous ftp retrieval from the following computer nodes: ftp.ebi.ac.uk (/pub/databases/yeast_chrii); mips.embnet.org (|anonymous.yeast.chrII|); genome-ftp.stanford.edu (/pub/yeast/genome_seq/chrII); and ncbi.nlm.nih.gov (/repository/yeast/chrII). This work was supported by the EU under the BRIDGE Programme; the Région Wallone, the Fonds National de la Recherche Scientifique and La Région de Bruxelles Capitale; The Bundesminister für Forschung und Technologie and the Fonds der Chemischen Industrie; The Ministère de l'Education Nationale and the Ministère de la Recherche et de l'Espace; and The Ministère de la Recherche et Technologie; The Jumelage Franco-Polonais du CNRS.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
Amati,B.B. and Gasser,S.M. (1988) *Cell*, **54**, 967–978.
Bairoch,A. (1989) In *EMBL Biocomputing Technical Document 4*. EMBL, Heidelberg, Germany.
Baur,A., Schaaff-Gerstenschläger,I., Boles,E., Miosga,T., Rose,M. and Zimmermann,F.K. (1993) *Yeast*, **9**, 289–293.
Bécam,A.-M. *et al.* (1994) *Yeast*, **10**, S1–11.
Bell,S.P. and Stillman,B. (1992) *Nature*, **357**, 128–134.
Bernardi,G. (1993) *Gene*, **135**, 57–66.
Bussereau,F., Mallet,L., Gaillon,L. and Jacquet,M. (1993) *Yeast*, **9**, 797–806.
Califano,A. and Rigoutsos,I. (1993) In *Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology*. Bethesda, MD, pp. 56–64.
Daniels,D.L., Plunkett,G., Burland,V. and Blattner,F.R. (1992) *Science*, **257**, 771–778.
Delaveau,T., Jacq,C. and Perea,J. (1992) *Yeast*, **8**, 761–768.
Delaveau,T., Delahodde,A., Carvajal,E., Subik,J. and Jacq,C. (1994) *Mol. Gen. Genet.*, **243**, 501–511.
Démolis,N., Mallet,L., Busserau,F. and Jacquet,M. (1993) *Yeast*, **9**, 645–659.
Démolis,N., Mallet,L. and Jacquet,M. (1994) *Yeast*, **10**, in press.
Dershowitz,A. and Newlon,C.S. (1993) *Mol. Cell. Biol.*, **13**, 391–398.
Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
De Wergifosse,P., Jacques,B., Jonniaux,J.-L., Purnelle,B., Skala,J. and Goffeau,A. (1994) *Yeast*, **10**, in press.
Doignon,F., Biteau,N., Crouzet,M. and Aigle,M. (1993a) *Yeast*, **9**, 189–199.
Doignon,F., Biteau,N., Aigle,M. and Crouzet,M. (1993b) *Yeast*, **9**, 1131–1137.
Dujon,B. *et al.* (1994) *Nature*, **369**, 371–378.
Eisenberg,S., Civalier,C. and Tye,B.K. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 743–746.
Estes,H.G., Robinson,B.S. and Eisenberg,S. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 11156–11160.
Fangman,W.L. and Brewer,B.J. (1992) *Cell*, **71**, 363–366.
Feldmann,H. (1988) In Grunberg-Manago,M., Clark,B.F.C. and Zachau,H.G. (eds), *Evolutionary Tinkering in Gene Expression*. NATO ASI Series, Life Sciences A169, Plenum Press, New York, pp. 79–86.
Fondrat,C. and Kalogeropoulos,A. (1994) *Curr. Genet.*, **25**, 396–406.
Goffeau,A. (1994) *Nature*, **369**, 101–102.
Goffeau,A., Nakai,K., Slonimski,P.P. and Risler,J.L. (1993a) *FEBS Lett.*, **325**, 112–117.
Goffeau,A., Slonimski,P.P., Nakai,K. and Risler,J.L. (1993b) *Yeast*, **9**, 691–702.
Hauber,J., Stucka,R., Krieg,R. and Feldmann,H. (1988) *Nucleic Acids Res.*, **16**, 10623–10634.
Hohn,B. and Hinnen,A. (1980) In Setlow,J.K. and Hollaender,A. (eds), *Genetic Engineering*. Plenum Press, New York, pp. 169–183.
Holmstrøm,K., Brandt,T. and Kallesøe,T. (1994) *Yeast*, **10**, 47–62.
Honore,N. *et al.* (1993) *Mol. Microbiol.*, **7**, 207–214.
Ito,B., Fukuda,Y. and Kimura,A. (1983) *J. Bacteriol.*, **153**, 163–168.
Ji,H., Moore,D.P., Blomberg,M.A., Braiterman,L.T., Voytas,D.F., Natsoulis,G. and Boeke,J.D. (1993) *Cell*, **73**, 1007–1018.
Johnson,D.I., Jacobs,C.W., Pringle,J.R., Robinson,L.C., Carle,G.F. and Olson,M.V. (1987) *Yeast*, **3**, 243–253.
Klein,P., Kaneisa,M. and Delesi,C. (1985) *Biochim. Biophys. Acta*, **815**, 468–476.
Kunst,F. and Devine,K. (1991) *Res. Microbiol.*, **142**, 905–912.
Lloyd,A.T. and Sharp,P.M. (1992) *J. Hered.*, **83**, 239–240.
Logghe,M., Molemans,F., Fiers,W. and Contreras,R. (1994) *Yeast*, **10**, 1093–1100.
Louis,E.J. (1994) *Yeast*, **10**, 271–274.
Louis,E.J. and Haber,J.E. (1992) *Genetics*, **119**, 303–315.
Louis,E.J., Naumova,E.S., Lee,A., Naumov,G. and Haber,E.J. (1994) *Genetics*, **136**, 789–802.
Lundblad,V. and Blackburn,E.H. (1993) *Cell*, **73**, 347–360.
Marahrens,Y. and Stillman,B. (1992) *Science*, **255**, 817–823.
Mallet,L., Bussereau,F. and Jacquet,M. (1994) *Yeast*, **10**, 819–831.
Mannhaupt,G., Stucka,R., Ehnle,S., Vetter,I. and Feldmann,H. (1994) *Yeast*, **10**, 1363–1381.
Meyerowitz,E.M. and Pruitt,R.E. (1985) *Science*, **229**, 1214–1218.
Michels,C.A., Read,E., Nat,K. and Charron,M.J. (1992) *Yeast*, **8**, 655–665.
Milosavljevic,A. and Jurka,J. (1993) *CABIOS*, **9**, 409–411.
Miosga,T. and Zimmermann,F.K. (1993) *Yeast*, **9**, 1273–1277.
Mortimer,R.K., Contopoulou,R. and King,J.S. (1992) *Yeast*, **8**, 817–902.
Nasr,F., Bécam,A.-M., Grzybowska,E., Zagulski,M., Slonimski,P.P. and Herbert,C.J. (1994a) *Curr. Genet.*, **26**, 1–7.
Nasr,F., Bécam,A.-M., Slonimski,P.P. and Herbert,C.J. (1994b) *C.R. Acad. Sci. Paris/Life Sci.*, **317**, 607–613.
Nelböck,P. (1988) Ph.D. Thesis, University of Würzburg, Germany.
Newlon,C.S. (1988) *Microbiol. Rev.*, **52**, 568–601.
Oliver,S. *et al.* (1992) *Nature*, **357**, 38–46.
Palladino,F. and Gasser,S.M. (1994) *Curr. Opin. Cell Biol.*, **6**, 373–379.
Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
Ramezani Rad,M., Kirchrath,L. and Hollenberg,C.P. (1994) *Yeast*, **10**, 1217–1225.
Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Schaaff-Gerstenschläger,I., Mannhaupt,G., Vetter,I., Zimmermann,F.K. and Feldmann,H. (1993a) *Eur. J. Biochem.*, **217**, 487–492.
Schaaff-Gerstenschläger,I., Baur,A., Boles,E. and Zimmermann,F.K. (1993b) *Yeast*, **9**, 915–921.
Schaaff-Gerstenschläger,I., Schindwolf,T., Lehnert,T., Rose,M. and Zimmermann,F.K. (1994) *Yeast*, **10**, in press.
Scherens,B., El Bakkoury,M., Vierendeels,F., Dubois,E. and Messenguy,F. (1993) *Yeast*, **9**, 1355–1371.
Sharp,P.M. and Li,W.H. (1987) *Nucleic Acids Res.*, **15**, 1281–1295.
Sharp,P.M. and Lloyd,A.T. (1993) *Nucleic Acids Res.*, **21**, 179–183.
Skala,J., van Dyck,L., Purnelle,B. and Goffeau,A. (1992) *Yeast*, **8**, 777–785.
Skala,J., van Dyck,L., Purnelle,B. and Goffeau,A. (1994) *Yeast*, **10**, S13–24.
Smits,P., De Haan,M., Maat,C. and Grivell,L.A. (1994) *Yeast*, **10**, S75–80.
Steinberg,S., Misch,A. and Sprinzl,M. (1993) *Nucleic Acids Res.*, **21**, 3011–3015.

Stucka,R. (1992) Ph.D. Thesis, Ludwig-Maximilians-Universität Munich, Germany.
Stucka,R. and Feldmann,H. (1994) In Johnston,J. (ed.), *Molecular Genetics of Yeast: A Practical Approach*. Oxford University Press, Oxford, UK, pp. 49–64.
Thierry,A. and Dujon,B. (1992) *Nucleic Acids Res.*, **20**, 5625–5631.
Thierry,A., Fairhead,C. and Dujon,B. (1990) *Yeast*, **6**, 521–534.
van der Aart,Q.J.M., Barthe,C., Doignon,F., Aigle,M., Crouzet,M. and Steensma,H.Y. (1994) *Yeast*, **10**, 959–964.
Van Dyck,L., Purnelle,B., Skala,J. and Goffeau,A. (1992) *Yeast*, **8**, 769–776.
Van Dyck,L., Pearce,D. and Sherman,F. (1993) *J. Biol. Chem.*, **269**, 238–242.
Van Dyck,L., Jonniaux,J.-L., de Melo Barreiros,T., Kleine,K. and Goffeau,A. (1994) *Yeast*, **10**, in press.
Van Houten,J.V. and Newlon,S.M. (1990) *Mol. Cell. Biol.*, **10**, 3917–3925.
Vassarotti,A. and Goffeau,A. (1992) *Trends Biotechnol.*, **10**, 15–18.
Wilson,R. *et al.* (1994) *Nature*, **368**, 32–38.
Wolfe,K.H. and Lohan,A.J.E. (1994) *Yeast*, **10**, S41–46.
Zagulski,M., Bécam,A.-M., Grzybowska,E., Lacroute,F., Migdalski,A., Slonimski,P.P., Sokolowska,B. and Herbert,C.J. (1994) *Yeast*, **10**, 1227–1234.

Received on August 16, 1994; revised on September 21, 1994