Genomic Exploration of the Hemiascomycetous Yeasts: 3. Methods and strategies used for sequence analysis and annotation

Fredj Tekaia[a,]*, Gaëlle Blandin[a], Alain Malpertuy[a], Bertrand Llorente[a], Pascal Durrens[b], Claire Toffano-Nioche[c], Odile Ozier-Kalogeropoulos[a], Elisabeth Bon[d], Claude Gaillardin[d], Michel Aigle[b], Monique Bolotin-Fukuhara[c], Serge Casarégola[d], Jacky de Montigny[e], Andrée Lépingle[d], Cécile Neuvéglise[d], Serge Potier[e], Jean-Luc Souciet[e], Micheline Wésolowski-Louvel[f], Bernard Dujon[a]

[a] *Unité de Génétique Moléculaire des Levures (URA 2171 CNRS and UFR927 Univ. P.M. Curie), Institut Pasteur, 25, Rue du Dr Roux, F-75724 Paris Cedex 15, France*

[b] *Laboratoire de Biologie Cellulaire de la Levure, IBGC, 1 rue Camille Saint-Saens, F-33077 Bordeaux Cedex, France*

[c] *Institut de Génétique Moléculaire (CNRS/UPS UMR 8621), Batiment 400, Université de Paris Sud, F-91405 Orsay, France*

[d] *Collection de Levures d'Intérêt Biotechnologique, Laboratoire de Génétique Moléculaire et Cellulaire (INRA UMR 216, CNRS URA 1925) INA-PG, BP 01, F-78850 Thiverval-Grignon, France*

[e] *Laboratoire de Génétique et Microbiologie (ULP/CNRS UPRES-A 7010), Institut de Botanique, 28 rue Goethe, Strasbourg Cedex, France*

[f] *Microbiologie et Génétique (CNRS/UCB/INSA ERS 2009), Bat. 405 R2, Université Lyon I, F-69622 Villeurbanne Cedex, France*

Received 9 November 2000; accepted 11 November 2000

First published online 29 November 2000

Edited by Horst Feldmann

Abstract The primary analysis of the sequences for our Hemiascomycete random sequence tag (RST) project was performed using a combination of classical methods for sequence comparison and contig assembly, and of specifically written scripts and computer visualization routines. Comparisons were performed first against DNA and protein sequences from *Saccharomyces cerevisiae*, then against protein sequences from other completely sequenced organisms and, finally, against protein sequences from all other organisms. *Blast* alignments were individually inspected to help recognize genes within our random genomic sequences despite the fact that only parts of them were available. For each yeast species, validated alignments were used to infer the proper genetic code, to determine codon usage preferences and to calculate their degree of sequence divergence with *S. cerevisiae*. The quality of each genomic library was monitored from contig analysis of the DNA sequences. Annotated sequences were submitted to the EMBL database, and the general annotation tables produced served as a basis for our comparative description of the evolution, redundancy and function of the Hemiascomycete genomes described in other articles of this issue. © 2000 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Alignment; Contig; Mitochondrion; Ty; Plasmid; rDNA; tRNA; Protein; Genetic code; RSCU

1. Introduction

In the present program, we have tried to obtain a maximum of biologically relevant data from a minimal sequencing effort. For reasons that will become more apparent later [1–5], this strategy relies upon the existence of the complete genome sequence of *Saccharomyces cerevisiae* to serve as the universal reference for comparisons of all other yeast species despite their different phylogenetic distances. We have, therefore, first compared all our RSTs to the predicted *S. cerevisiae* gene products, and only subsequently to the other completely sequenced organisms and to protein sequences of all other organisms. Because we have deliberately limited our sequencing of each yeast species to low genome coverage (from 0.2 to 0.4 genome equivalents) we reasoned that all sequence comparisons could be done using the *blast* algorithm only and that the application of more complete search algorithms was premature. Yet, in order to limit the effect of alignment size variations due to the fact that we examined random single sequence reads and not complete gene sequences, we introduced a manual validation step which was largely facilitated by the use of a number of scripts and routines specifically written for this project or adapted from existing sources. The present paper describes the strategy used for the interpretation of the 49 199 original DNA sequences, totalling over 45 millions nucleotides, obtained from the 13 Hemiascomycete yeast species selected in this work [6].

2. Materials and methods

2.1. Databases

Systematic comparisons of all our Hemiascomycete RSTs [7] were conducted against two different types of databases.

First, a series of seven *S. cerevisiae* data files comprising: (1) the 6213 predicted protein sequences as described in [1]; (2) the intergenic and subtelomeric DNA sequences [4]; (3) the 2 μm plasmid DNA sequence [8]; (4) the rDNA sequences extracted from the chromosome XII sequence [9]; (5) the 52 distinct tRNA gene sequences as defined in [1]; (6) the Ty elements DNA sequences [10]; and (7) the mitochondrial DNA sequence [11] and its translation products.

Second, a compilation of 124 456 protein sequences, named *GPROTEOME* and containing the 66 091 predicted protein sequences from the first 23 completely sequenced organisms plus the *Schizosaccharomyces pombe* partial sequence (as defined by Table 1), merged with a 'filtered' SwissProt [34] version comprising 58 365 entries. The 'filtered' SwissProt version was constructed from the *sprot.dat* and *sprot.fas* files downloaded on November 3rd, 1999 from the *ftp.expasy.ch* server and each containing 81 851 entries (*databases/*

*Corresponding author. Fax: (33)-140-61 34 56.
E-mail: tekaia@pasteur.fr

PII: S0014-5793(00)02274-2

sp_tr_nrdb/, and *databases/fasta/* directories, respectively). All entries corresponding to *S. cerevisiae* and to the organisms listed in Table 1 (identified by the OS line in the *sprot.dat* file which contains annotations and sequences) were eliminated from the *sprot.fas* file which only contains the sequences in the *fasta* format (note that sequences corresponding to *Deinococcus radiodurans* were also eliminated because the complete sequence of this species was released at the time we constructed the 'filtered' SwissProt, but not included in *GPROTEOME*). The use of *GPROTEOME* introduces over 89 000 novel protein sequences, not homologous to *S. cerevisiae* proteins.

2.2. Annotation strategy

The general strategy applied to the analysis of all RSTs is diagrammed by Fig. 1. It uses standard published procedures of sequence comparisons as well as *perl* and *sh shell* scripts specifically developed for this work (Table 2). The procedure involves the following steps, applied independently to each yeast species.

2.2.1. Step 1: contig assembly. For each yeast species, contigs were assembled from the sequence electropherograms using the *phred/phrap* programs version 0.99.03.19 [35,36] with their default options except as otherwise indicated [37–49]. The sequence of the cloning vector used was added to the vector sequence database distributed with the *phred/phrap* programs.

2.2.2. Step 2: comparisons of RSTs to S. cerevisiae rDNA, tRNA genes, Ty elements, plasmid and mitochondrial sequences. Comparisons of all RSTs with *S. cerevisiae* DNA sequences were performed using *blastn* (default parameters) to search for rDNA, tRNA genes, plasmid or mitochondrial DNA sequences, using *tblastx* for Ty elements, and using *blastx* for mitochondrial gene translation products. The *blastx* and *tblastx* searches were made using the *seg* filter [50] and the *pam250* substitution matrix (unless otherwise indicated in respective articles). The *blastx* searches against mitochondrial sequences were made using the yeast mitochondrial genetic code.

2.2.3. Step 3: identification of genetic elements other than protein-coding genes in contigs and single RSTs. Contigs matching *S. cerevisiae* rDNA repeat unit or Ty elements were examined to reconstitute the rDNA repeat organization or the structure of the putative retrotransposons in the species of interest. For each yeast species, new long terminal repeats (LTRs) identified by their vicinity to Ty ORFs were used in turn as *blastn* queries to search the complete RST set for solo LTRs. Nuclear tRNA genes were validated after examination of the anticodon stem and determination of the anticodon. The possible existence of introns in tRNA genes was also examined.

2.2.4. Step 4: blast comparisons to protein sequences. Comparisons of RSTs from each yeast species with the *S. cerevisiae* proteome were performed using *blastx* [51] version 2.0.10 with the *seg* filter [50] and either the *pam250* substitution matrix (*Candida tropicalis*, *Kluyveromyces lactis*, *Kluyveromyces marxianus* var. *marxianus*, *Kluyveromyces thermotolerans*, *Pichia angusta*, *Pichia sorbitophila* and *Zygosaccharomyces rouxii*) or the *blosum62* substitution matrix (*Debaryomyces hansenii* var. *hansenii*, *Saccharomyces bayanus* var. *uvarum*, *Saccharomyces exiguus*, *Saccharomyces servazzii*, *Saccharomyces kluyverii* and *Yarrowia lipolytica*). Comparisons with *GPROTEOME* were done using the same procedure except that *pam250* was used throughout.

All *blast* searches were automatically launched using the script '*blastallgenomes*' which gives as output the detailed *blast* results for each RST. Matching segments and relevant descriptive figures were automatically extracted from the *blast* outputs using the script '*readblast*' in order to construct the working annotation table.

Note that all *blastx* searches were made using the 'universal' genetic code (see Sections 2.4 and 3.2).

2.2.5. Step 5: expert homology validation for S. cerevisiae comparisons. Because the genome fragments sequenced in the RSTs fall at random with respect to actual gene limits, a manual validation step was introduced taking into account, in addition to *blast* scores, the size of the aligned segment, the amino acid identity score, and the positions of the aligned segment with respect to the homologous gene. This step was needed because a short gene fragment falling at the edge of an RST may have a low *blast* score although sharing a high degree of sequence similarity to a homolog (Fig. 2, left). In addition, the visual validation step was used to distinguish between RST segments having a single clearcut homolog from those having several possible homologs as a result of the existence of gene families in *S. cerevisiae*. The first were denoted '**o**', the second '**oo**' (Fig. 2, right). Finally, some *blast* alignments having a significant expected value were nevertheless discarded because they were found to correspond to dispersed low similarities, short motifs or to overlap other possible alignments of higher quality.

Table 1
Protein sequences from completely sequenced organisms used for comparisons

Organism	Protein sequence data extracted from server	Ref.	Tot. prot.
Bacteria			
Aquifex aeolicus	ncbi.nlm.nih.gov/genbank/genomes/bacteria/Aquae/aquae.faa and aquae.ptt	[12]	1 522
Bacillus subtilis	ftp.pasteur.fr/GenomeDB/Subtilist/FlatFiles/SLR14.2_prot	[13]	4 100
Borrelia burgdorferi	ftp.tigr.org/pub/data/b_burgdorferi/GBB	[14]	1 639
Campylobacter jejuni	ftp.sanger.ac.uk/pub/pathogens/cj/CJ	[15]	1 731
Chlamydia pneumoniae	ncbi.nlm.nih.gov/genbank/genomes/bacteria/CP/cpneu.faa and cpneu.ptt	[16]	1 052
Chlamydi atrachomatis	ncbi.nlm.nih.gov/genbank/genomes/bacteria/Ctra/ctra.faa and ctra.ptt	[17]	877
Escherichia coli	ftp.genetics.wisc.edu/pub/sequence/m52.fap	[18]	4 290
Haemophilus influenzae	ftp.tigr.org/pub/data/h_influenzae/GHI	[19]	1 713
Helicobacter pylori	ftp.tigr.org/pub/data/h_pylori/GHP	[20]	1 577
Mycobacterium tuberculosis	www.pasteur.fr/Bio/TubercuList/TB_protein	[21]	3 924
Mycoplasma genitalium	ftp.tigr.org/pub/data/m_genitalium/GMG	[22]	479
Mycoplasma pneumoniae	www.zmbh.uni-heidelberg.de/M_pneumoniae/genome/Get_orf.html	[23]	677
Rickettsia prowazekii	evolution.bmc.uu.se/∼thomas/Rickettsia/dataRPaa.fas	[24]	837
Synechocystis sp.	ftp.kazuza.or.jp/pub/cyano/cyano.p.aa	[25]	3 168
Thermotoga maritima	ftp.tigr.org/pub/data/t_maritima/BTM	[26]	1 849
Treponema pallidum	ftp.tigr.org/pub/data/t_pallidum/GTP	[27]	1 031
Archaea			
Aeropirum pernix K1	ftp.bio.nite.go.jp/pub/a_pernix/apepep.fasta	[28]	2 694
Archaeoglobus fulgidus	ftp.tigr.org/pub/data/a_fulgidus/GAF	[29]	2 409
Methanobacterium thermoautotrophicum	www.genomecorp.com/ftp/sequences/methanobacter/mth_proteins.tfa	[30]	1 871
Methanococcus jannaschii	ftp.tigr.org/pub/data/m_jannaschii/GMJ	[31]	1 771
Pyrococcus abyssi	www.genoscope.cns.fr/Pab/		1 765
Pyrococcus horikoshii	ftp.bio.nite.go.jp/pub/ot3pep.fasta	[32]	2 061
Eukaryota			
Caenorhabditis elegans	ftp.sanger.ac.uk/pub/C_elegans_sequences/SCIENCES98/October_Proteins	[33]	19 099
S. pombe	ftp.sanger.ac.uk/pub/yeast/sequences/pombe/pompep		3 955

For each organism sequenced (listed in column 1), the table gives the ftp or http server where the sequences were downloaded from (column 2), the corresponding publication (column 3), and the total number of protein sequences (last column). All data were downloaded from servers on April 19th, 1999, except for *P. abyssi* (May 5th, 1999), *T. maritima* (May 28th, 1999), *A. pernix* (July 23rd, 1999) and *S. pombe* (October 9th, 1999). The *S. pombe* data set corresponds to ca. 70% of the total genome (V. Wood, personal communication).

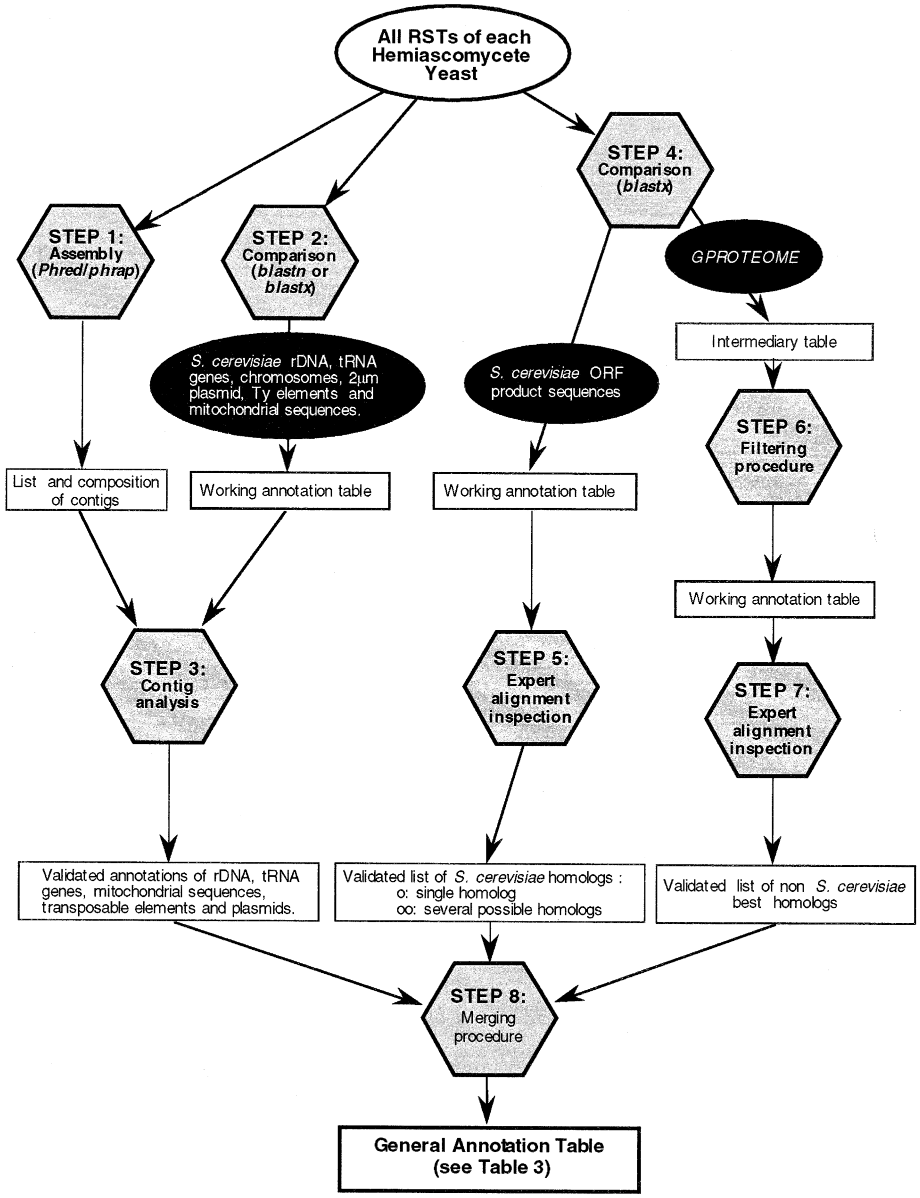

All RSTs of each Hemiascomycete Yeast
STEP 1: Assembly (Phred/phrap)
STEP 2: Comparison (blastn or blastx)
STEP 4: Comparison (blastx)
GPROTEOME
S. cerevisiae rDNA, tRNA genes, chromosomes, 2µm plasmid, Ty elements and mitochondrial sequences.
S. cerevisiae ORF product sequences
Intermediary table
STEP 6: Filtering procedure
List and composition of contigs
Working annotation table
Working annotation table
Working annotation table
STEP 3: Contig analysis
STEP 5: Expert alignment inspection
STEP 7: Expert alignment inspection
Validated annotations of rDNA, tRNA genes, mitochondrial sequences, transposable elements and plasmids.
Validated list of S. cerevisiae homologs : o: single homolog oo: several possible homologs
Validated list of non S. cerevisiae best homologs
STEP 8: Merging procedure
General Annotation Table (see Table 3)

Fig. 1. Flow chart of sequence analysis steps. All action steps (hexagons) and databases used for comparisons (ovals) refer to Section 2. The resulting tables and files are indicated by rectangles. Not shown: (1) a *S. cerevisiae* gene family table, constructed by *blastp* comparisons of the predicted *S. cerevisiae* translation products, and used for expert alignment inspection at step 5; (2) a *S. cerevisiae–GPROTEOME* relation table used for results filtration at step 6.

In order to facilitate the visual inspection of the alignments, we have adapted to our data set the original *blast2html* script of Kate Robinson (krobinson@nucleus.harvard.edu) that converts regular *blast* output files to the HTML format ('*html'ized*'). In addition, a graph was inserted above the descriptive lines (see Fig. 2) showing alignments colored according to their similarity score with the RST query. This graph is based on the original *PaintBlast* program of Alessandro Guffanti (http://hercules.tigem.it/Biomodules/PaintBlast.pm).

In order to facilitate the annotation of our RSTs, the working annotation table was *html'ized* (using the script '*segmatch2html*') to create links with (i) the output files of *blast* searches of each RST against the *S. cerevisiae* database, (ii) the RSTs and (iii) the hit sequences. Using such a procedure, we could immediately visualize all homologs to the various regions of a given RST and complete the working annotation table with the '**o**' or '**oo**' annotations (Table 3, column 15).

2.2.6. Step 6: filtering RST comparison results to GPROTEOME relatively to comparison results to S. cerevisiae. This procedure is needed to reduce the number of RSTs to be visually inspected such as to enable experts to solely concentrate on new hits. Filtration was done using the script '*cleanreads*' which eliminates all hits (**P**) of an RST (**R**) to *GPROTEOME* if that RST (**R**) already has a validated hit to *S. cerevisiae* (**Y**) which itself has the same hit (**P**) in *GPROTEOME.*

if (**R** → **P**) and (**R** → **Y** and **Y** → **P**) then (**R** → **P**) is ignored.

For this calculation, a table of comparison results between *S. cerevisiae* and *GPROTEOME*, listing all **Y**→**P** relations, has been set up. This table contains the 202 960 record lines of all significant *blastp* hits (see [52,53] for threshold definition) between the 6213 *S. cerevisiae* predicted ORF products and *GPROTEOME.*

2.2.7. Step 7: expert homology validation for GPROTEOME comparisons. Validation was performed as in step 5 except for the addition of a link to the *html'ized blast* comparisons of RSTs to *S. cerevisiae* and for the fact that only the best match was retained. Note that, at this step, a threshold of 25% amino acid identity was placed

Table 2
List of *perl* or *sh shell* scripts used for this work

Script name	Function
blastallgenomes	launches systematic *blastx* comparisons
blast2html	converts *blast* output files in html format and constructs graphs of the aligned segments with the query RST
cleanreads	filters comparison results of RSTs to GPROTEOME (see Section 2)
codonusageh	computes occurrences of all amino acids from *S. cerevisiae* sequences with their corresponding codons in the validated alignments of the RST
extractpartseq	extracts DNA sequences of RST segments corresponding to validated alignments
readblast	extracts descriptive figures from *blast* output alignments and produces a working annotation table (see Table 3, columns 1–14)
readblastxalign	extracts amino acid sequence segments of the translated RST and of *S. cerevisiae* gene products corresponding to validated alignments
segmatch2html	converts the working annotation table in html format and makes links to the *html-ized blast* output files as well as to RST and hit sequences

All scripts can be downloaded from http://www-alt.pasteur.fr/∼tekaia/HYG/scripts.html. They were written by F. Tekaia (except for *blast2html*, see Section 2) and may need adaptation for different environments.

for alignments longer than 100 amino acids whereas, for shorter alignments, a minimum of 50% amino acid identity was demanded.

2.2.8. Step 8: merging the comparison results. This step results in the completion of a general annotation table whose logic is exemplified by Table 3. This table is used for all subsequent analyses described in this series of publications. For each RST, the annotation submitted to EMBL was extracted from columns 1, 2, 9, 11, 12 and 16. Note that in cases of sequencing frameshift errors, only the two extreme coordinates of all alignments with the same homolog are given. Functional annotation was added for the *S. cerevisiae* homologs (extracted from MIPS (http://www.mips.biochem.mpg.de) and/or YPD (http://www.proteome.com/databases/index.html), and for other homologs (extracted from SwissProt).

2.3. Minimum and maximum number of genes identified in each yeast species by comparison to S. cerevisiae

When several (x) distinct RSTs share homology to the same *S. cerevisiae* gene, two cases were distinguished depending on whether or not the RST fragments are similar to the same region of the *S. cerevisiae* sequence. In the first case, the number of genes in the yeast species studied was deduced from the fact that the RST fragments can be assembled into a contig (indicating that they are part of the same gene) or not (indicating that they belong to different genes). In the second case, as it is not possible to decide upon the existence of one or several genes in the studied species, we considered 1 as the minimum number of genes and x as the maximum number.

2.4. Determination of the genetic code and codon usage

For each '**o**' validated *blastx* alignment of an RST segment with a *S. cerevisiae* protein sequence were extracted (i) the corresponding DNA sequence of the RST using the script '*extractpartseq*' which reads the first and last positions as indicated in the general annotation table (Table 3, columns 11 and 12, respectively), and (ii) the corresponding segment of amino acid sequence of the *S. cerevisiae* gene product (defined by columns 13 and 14 of Table 3) using the script '*readblastxalign*'. For each yeast species, the number of occurrences of each of the 20 amino acids was then computed for each of the 64 codons using the script '*codonusageh*'.

Relative synonymous codon usage (RSCU) values were calculated as in [54].

2.5. Contig distribution and genome size

In random genome sequencing programs, the frequency of contigs and singletons evolves with increasing genome coverage as described by [55]. Because our sequence data originate from random clones that were, in their large majority, sequenced from both ends (see [7]), the mathematical analysis of [56] should be applied. Estimation of genome sizes for each species can be found in original articles of this series [37–49].

To estimate the randomness of the genomic libraries the population of RSTs was first filtered to eliminate non-chromosomal or known repeated elements. For each species, the total number of RSTs included in the calculation (N) was determined as follows:

$$N = N_s - (N_r + N_m + N_y + N_w + N_x + P)$$

where N_s is the total number of RSTs, and N_r, N_m, N_y, N_w and N_x the number of RSTs corresponding to, respectively, rDNA, mitochondrial DNA, Ty elements, other repeated elements or eliminated for various reasons (size, sequence, quality, etc.). P is the number of cases in which the two RSTs from a same insert overlap each other (short inserts).

2.6. Comparisons of RSTs to intergenic and subtelomeric DNA sequences from S. cerevisiae

All nuclear DNA sequences from *S. cerevisiae* which are not part of the predicted protein-coding genes, tRNA genes, rDNA or other RNA-coding genes, Ty elements or centromeres (a total of 3 165 016 bp or ca. 24% of the *S. cerevisiae* nuclear genome, see [1]) were compared to all RSTs using *tblastx*. Results were analyzed individually by visual inspection. Additional *S. cerevisiae* genes discovered during this analysis are described in [1] and their corresponding homologs are indicated in [37–49].

3. Results and discussion

3.1. Basic annotation

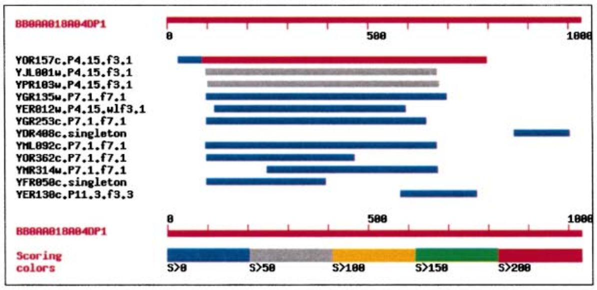
Given the overall sequencing strategy of this program, namely a random exploration of yeast genomes at low coverage, and despite the exceptional quality and length of the sequencing reads [7], the interpretation and annotation of sequences can only rely on comparisons, not on gene prediction. For reasons that will become apparent in [2–5,37–49], we have chosen to give the priority to comparisons to *S. cerevisiae* and only subsequently to comparisons to other organisms, giving again the priority to the completely sequenced genomes over the piecemeal sequences of general databases. No systematic comparisons have been made with other yeasts such as *S. pombe* or *Candida albicans*, whose genomes are extensively studied, because their DNA sequencing is still incomplete.

After data processing as explained by Fig. 1 and Section 2,

→

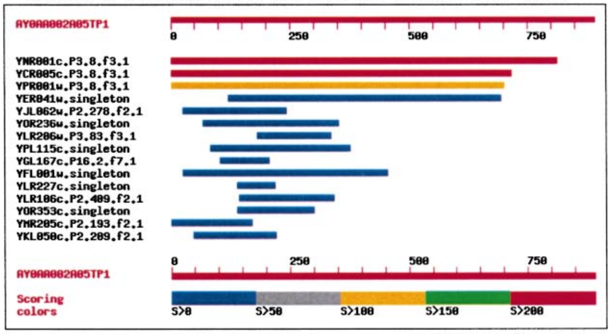
Fig. 2. HTML-formatted *blast* output files used for alignment validation. These files were produced using the script *blast2html* (Table 2 and Section 2) and used with the *Netscape* browser to facilitate the expert validation step of the proposed *blast* alignments. Note that, on the alignment graphs (inserts), the word 'singleton' or a 'partition number' (e.g. P4.15.f3.1) is associated to each *S. cerevisiaee* ORF name to help decide whether the hits are members of a same *S. cerevisiae* gene family or not (the partitioning process refers to [1,53], the designation P4.15.f3.1, for example, refers to the fact that YOR157c is a member of a three-members family number 1, which is included in the four-members part number 15). Left: alignments of an RST showing: (i) an example of an expert validation of a single homolog validated **o** (*YOR157c*) among three members of the same *S. cerevisiae* gene family (P4.15.f3.1); (ii) an example of several alignments with the same gene (*YOR157c*) due to a sequencing frameshift error; and (iii) an example of expert validation of a short gene fragment falling at the edge of the RST (*YDR408c*). Right: alignments of another RST showing an example of ambiguity between two possible homologs each validated as **oo** (*YNR001c* and *YCR005c*) of the same *S. cerevisiae* three-members gene family (P3.8.f3.1).

```
BLASTX 2.0.8 [Jan-05-1999]

Query= BB0AA018A04DP1 (1032 letters) (Pichia angusta)
Database: S. cerevisiae genome, 6213 sequences; 2,850,294 total letters
```



```
                                                               Score     E
Sequences producing significant alignments:                   (bits)  Value

YOR157c PUP1 20S proteasome subunit (beta2)                     269   2e-74
YJL001w PRE3 20S proteasome subunit (beta1)                      81   2e-16
YPR103w PRE2 20S proteasome subunit (beta5)                      54   4e-08
YGR135w PRE9 20S proteasome subunit Y13 (alpha3)                 46   8e-06
YER012w PRE1 20S proteasome subunit C11(beta4)                   45   2e-05
YGR253c PUP2 20S proteasome subunit(alpha5)                      42   1e-04
YDR408c ADE8 phosphoribosylglycinamide formyltran...             38   0.003


>YOR157c PUP1 20S proteasome subunit (beta2), Length = 261
 Score =  269 bits (936), Expect = 2e-74
 Identities = 185/236 (78%), Positives = 212/236 (89%), Frame = +3

Query: 90  DPAAISTGTTIVGCKFKDGVVIAADTRATAGPIVADKNCEKLHRLAPRIWCAGAGTAADT 269
             P A STGTTIVG KF +GVVIAADTR+T GPIVADKNC KLHR++P+IWCAGAGTAADT
Sbjct: 22  QPKATSTGTTIVGVKFNNGVVIAADTRSTQGPIVADKNCAKLHRISPKIWCAGAGTAADT 81

Query: 270 EMVTQLVQSNLELHSMSLNREPRVSSALQMLKQHLFKYQGHIGAYLIVAGVDPKGAHLFS 449
            E VTQL+ SN+ELHS+  +REPRV SALQMLKQHLFKYQGHIGAYLIVAGVDP G+HLFS
Sbjct: 82  EAVTQLIGSNIELHSLYTSREPRVVSALQMLKQHLFKYQGHIGAYLIVAGVDPTGSHLFS 141

Query: 450 IHAHGSTDIGFYQSLGSGSLAAMAVLERDWKEDLTKEEAMKLCADAIEAGIWNDLGSGSN 629
            IHAHGSTD+G+Y SLGSGSLAAMAVLE  WK+DLTKEEA+KL +DAI+AGIWNDLGSGSN
Sbjct: 142 IHAHGSTDVGYYLSLGSGSLAAMAVLESHWKQDLTKEEAIKLASDAIQAGIWNDLGSGSN 201

Query: 630 VDLCVMEIGKDAQLYRNFLTPNVREAKARNYKFERGTTAILKESIYNLCEVEEVRV 797
            VD+CVMEIGKDA+  RN+LTPNVRE K ++YKF RGTTA+LKESI N+C+++E +V
Sbjct: 202 VDVCVMEIGKDAEYLRNYLTPNVREEKQKSYKFPRGTTAVLKESIVNICDIQEEQV 257

 Score = 26.9 bits (78), Expect = 2e-74
 Identities = 15/24 (62%), Positives = 19/24 (78%), Frame = +1

Query: 28 MAGLSFDNFQRNQFLSKNGVQTPQ 99
           MAGLSFDN+QRN FL++N    P+
Sbjct: 1  MAGLSFDNYQRNNFLAENSHTQPK 24


>YPR103w PRE2 20S proteasome subunit (beta5), Length = 287
 Score = 54.0 bits (174), Expect = 4e-08
 Identities = 55/192 (28%), Positives = 99/192 (50%), Gaps = 2/192 (1%), Frame = +3

Query: 102 ISTGTTIVGCKFKDGVVIAADTRATAGPIVADKNCEKLHRLAPRIWCAGAGTAADTEMVT 281
            I+ GTT +  +F+ G+++A D+RATAG  VA +  +K+  + P +    AG AAD +
Sbjct: 72  IAHGTTTLAFRFQGGIIVAVDSRATAGNWVASQTVKKVIEINPFLLGTMAGGAADCQFWE 131

Query: 282 QLVQSNLELHSMSLNREPRVSSALQMLKQHLFKYQGH--IGAYLIVAGVDPKGAHLFSIH 455
              + S   LH +        V++A ++L   +++Y+G        +I       +G  ++ +
Sbjct: 132 TWLGSQCRLHELREKERISVAAASKILSNLVYQYKGAGLSMGTMICGYTRKEGPTIYYVD 191

Query: 456 AHGSTDIGFYQSLGSGSLAAMAVLERDWKEDLTKEEAMKLCADAIEAGIWNDLGSGSNVD 635
            + G+   G    +GSG   A  VL+ ++K DL+ E+A+ L   +I A    D  SG +V+
Sbjct: 192 SDGTRLKGDIFCVGSGQTFAYGVLDSNYKWDLSVEDALYLGKRSILAAAHRDAYSGGSVN 251

Query: 636 LCVMEIGKDAQLYR 677
            L    + +D  +Y
Sbjct: 252 LY--HVTEDGWIYH 263


>YDR408c ADE8 phosphoribosylglycinamide formyltransferase (GART), Length = 214
 Score = 37.6 bits (116), Expect = 0.003
 Identities = 25/46 (54%), Positives = 38/46 (82%), Gaps = 2/46 (4%), Frame = -3

Query: 1003 VDKGTPLIVKEIDVK--KESLEEWEARIHALEHEAIVEGTIEVLKQLN 866
             VDKG PL+VK++++    +E+LE++E R+H  EH AIVE T +VL+QL+
Sbjct: 166  VDKGEPLVVKKLEIIPGEETLEQYEQRVHDAEHIAIVEATYKVLQQLH 213
```

```
BLASTX 2.0.8 [Jan-05-1999]

Query= AY0AA002A05TP1 (891 letters) (Kluyveromyces thermotolerans)
Database: S. cerevisiae genome, 6213 sequences; 2,850,294 total letters
```



```
                                                           Score     E
Sequences producing significant alignments:               (bits)   Value

YNR001c CIT1 citrate (si)-synthase, mitochondrial           301   7e-83
YCR005c CIT2 citrate (si)-synthase, peroxisomal             274   1e-74
YPR001w CIT3 citrate (si)-synthase, mitochondrial           146   6e-36

>YNR001c CIT1 citrate (si)-synthase, mitochondrial, Length = 479
 Score =  301 bits (1051), Expect = 7e-83
 Identities = 198/270 (73%), Positives = 234/270 (86%), Frame = -3

Query: 811 ASTRNTLARGLLQQSNSGRASVLLTVGGARLYSNGEKTLKXSFAEIIPAKAEQIKAXRQE 632
             ++++ L+RG  +Q  + + ++    +       S  E+TLK  FAEIIPAKAE+IK  ++E
Sbjct: 7   TTSKSFLSRGSTRQCQNMQKALFALLNARHYSSASEQTLKERFAEIIPAKAEEIKKFKKE 66

Query: 631 HGSTVIGEVVLNQAYGGMRGIKGLVWEGSVLDPDEGIRFRNRTIPDIQKELPKGAGGTEP 452
            HG TVIGEV+L QAYGGMRGIKGLVWEGSVLDP+EGIRFR RTIP+IQ+ELPK  G TEP
Sbjct: 67  HGKTVIGEVLLEQAYGGMRGIKGLVWEGSVLDPEEGIRFRGRTIPEIQRELPKAEGSTEP 126

Query: 451 LPEALFWLLLTGETPTESQVKALSADLASRSELPEHVSQLLDSLPKDLHPMAQFSIAVTA 272
            LPEALFWLLLTGE PT++QVKALSADLA+RSE+PEHV QLLDSLPKDLHPMAQFSIAVTA
Sbjct: 127 LPEALFWLLLTGEIPTDAQVKALSADLAARSEIPEHVIQLLDSLPKDLHPMAQFSIAVTA 186

Query: 271 LESESKFSKAYAQGVSKKDYWNYAFEDSMDLIGKLPVIASKIYRNVFKDGKLGSVDPNAD 92
            LESESKF+KAYAQGVSKK+YW+Y FEDS+DL+GKLPVIASKIYRNVFKDGK+ S DPNAD
Sbjct: 187 LESESKFAKAYAQGVSKKEYWSYTFEDSLDLLGKLPVIASKIYRNVFKDGKITSTDPNAD 246

Query: 91  FGKNLANLLGFKNDEFVELMRLYLTIHADH 2
            +GKNLA LLG++N +F++LMRLYLTIH+DH
Sbjct: 247 YGKNLAQLLGYENKDFIDLMRLYLTIHSDH 276


>YCR005c CIT2 citrate (si)-synthase, peroxisomal, Length = 460
 Score =  274 bits (955), Expect = 1e-74
 Identities = 174/238 (73%), Positives = 206/238 (86%), Frame = -3

Query: 715 SNGEKTLKXSFAEIIPAKAEQIKAXRQEHGSTVIGEVVLNQAYGGMRGIKGLVWEGSVLD 536
            S+ EKTLK  F+EI P  A+ ++   +EHG T I +V+L Q YGGMRGI G VWEGSVLD
Sbjct: 20  SSQEKTLKERFSEIYPIHAQDVRQFVKEHGKTKISDVLLEQVYGGMRGIPGSVWEGSVLD 79

Query: 535 PDEGIRFRNRTIPDIQKELPKGAGGTEPLPEALFWLLLTGETPTESQVKALSADLASRSE 356
            P++GIRFR RTI DIQK+LPK  G ++PLPEALFWLLLTGE PT++QV+ LSADL SRSE
Sbjct: 80  PEDGIRFRGRTIADIQKDLPKAKGSSQPLPEALFWLLLTGEVPTQAQVENLSADLMSRSE 139

Query: 355 LPEHVSQLLDSLPKDLHPMAQFSIAVTALESESKFSKAYAQGVSKKDYWNYAFEDSMDLI 176
            LP HV QLLD+LPKDLHPMAQFSIAVTALESESKF+KAYAQG+SK+DYW+Y FEDS+DL+
Sbjct: 140 LPSHVVQLLDNLPKDLHPMAQFSIAVTALESESKFAKAYAQGISKQDYWSYTFEDSLDLL 199

Query: 175 GKLPVIASKIYRNVFKDGKLGSVDPNADFGKNLANLLGFKNDEFVELMRLYLTIHADH 2
            GKLPVIA+KIYRNVFKDGK+G VDPNAD+ KNL NL+G K+++FV+LMRLYLTIH+DH
Sbjct: 200 GKLPVIAAKIYRNVFKDGKMGEVDPNADYAKNLVNLIGSKDEDFVDLMRLYLTIHSDH 257


>YPR001w CIT3 citrate (si)-synthase, mitochondrial, Length = 486
 Score =  146 bits (500), Expect = 6e-36
 Identities = 93/233 (39%), Positives = 142/233 (60%), Gaps = 19/233 (8%),
 Frame = -3

Query: 700 TLKXSFAEIIPAKAEQIKAXRQEHGSTVIGEVVLNQAYGGMRGIKGLVWEGSVLDPDEGI 521
            TLK +    +IP K + +K  +   +GST +G + ++   GGMRG + + W+G+ LDP+ GI
Sbjct: 28  TLKEALENVIPKKRDAVKKLKACYGSTFVGPITISSVLGGMRGNQSMFWQGTSLDPEHGI 87

Query: 520 RFRNRTIPDIQKELPK-GAGGTEPLPEALFWLLLTGETPTESQVKALSADLASRS-ELPE 347
            +F+   TI + Q  LP  G  G   LPE++ WLL+TG  PT  Q  +   +LA R  +LP
Sbjct: 88  KFQGLTIEECQNRLPNTGIDGDNFLPESMLWLLMTGGVPTFQQAASFRKELAIRGRKLPH 147

Query: 346 HVSQLLDSLPKDLHPMAQFSIAVTALESESKFSKAYAQG-VSKKDYWNYAFEDSMDLIGK 170
            +  ++L SLPKD+HPM Q +I + ++   S F+  Y +G + K ++W    EDS++LI
Sbjct: 148 YTEKVLSSLPKDMHPMTQLAIGLASMNKGSLFATNYQKGLIGKMEFWKDTLEDSLNLIAS 207

Query: 169 LPVIASKIYRNVFKDGK-LGSVDPNADFGKNLANLLGFKND---------------EFVE 38
            LP++  +IY N+  +G  LG       D+  N+ +LLG  N               +F+
Sbjct: 208 LPLLTGRIYSNITNEGHPLGQYSEEVDWCTNICSLLGMTNGTNSSNTCNLTSQQSLDFIN 267

Query: 37  LMRLYLTIHADH 2
            LMRLY  IH DH
Sbjct: 268 LMRLYTGIHVDH 279
```

Table 3
Structure and logic of the general annotation table of RSTs

RST	Size RST (nuc)	Match	Size match (a.a.)	*Blastx* E-value	% Iden.	% Sim.	% Gaps	Fr./str.	Seg. size	Coord./RST		Coord./match		Val. code	Primary annotation
										Beg. (nuc)	End (nuc)	Beg. (a.a.)	End (a.a.)		
BB0AA001A02D1	1019														no similarity found
BB0AA001A01D1	943	*YDL006w*	281	4.00E-08	76	88		2	39	794	910	21	59	**o**	similar to *S. cerevisiae YDL006w*
BB0AA018A04DP1	1032	*YOR157c*	261	2.00E-74	62	78		1	24	28	99	1	24	**o**	similar to *S. cerevisiae YOR157c*
BB0AA018A04DP1	1032	*YOR157c*	261	2.00E-74	78	89		3	236	90	797	22	257	**o**	similar to *S. cerevisiae YOR157c*
BB0AA018A04DP1	1032	*YDR408c*	214	0.003	54	82	4	−3	46	1003	866	166	213	**o**	similar to *S. cerevisiae YDR408c*
AY0AA002A05TP1	891	*YNR001c*	479	7.00E-83	73	86		−3	270	811	2	7	276	**oo**	similar to *S. cerevisiae YNR001c*
AY0AA002A05TP1	891	*YCR005c*	460	1.00E-74	73	86		−3	238	715	2	20	257	**oo**	similar to *S. cerevisiae YCR005c*
BB0AA001B03T1	987	Q45515	471	9.00E-37	37	55		−1	314	945	4	104	393	**o**	similar to *B. stearothermophilus* Q45515
BB0AA004H06D1	993	tC_TGC_tRNA								778	813			**oo**	similar to *S. cerevisiae* C(TGC) tRNA genes
BB0AA004F11D1	1035	putative LTR element								805	1035			**o**	putative *P. angusta* LTR element
BB0AA015F08DP1	656	part of Ty element												**o**	similar to *S. cerevisiae* Ty element
AY0AA001B08DP1	968	rDNA												**o**	part of *K. thermotolerans* rDNA repeats
BB0AA001C06D1	541	mitochondrial DNA												**o**	part of *P. angusta* mitochondrial DNA
AR0AA003C07CP1	977	plasmid DNA (ZrpSR1)												**o**	part of *Z. rouxii* pSR1 plasmid

Each RST (column 1, in the actual tables, they are listed in alphanumerical order for each species) is represented by as many lines as necessary to describe all validated matches (see validation code in column 15). When no validated match is found, the RST is represented by a single line where columns 3 to 15 are left empty (example line 1). In all other cases, each line indicates the best homolog retained for annotation and corresponding to a given segment of the RST. In cases of protein sequence alignments (examples lines 2–8), the alignment parameters were extracted from the *blastx* outputs using the script *readblast* and are given respectively in columns 3 (accession number or gene name if *S. cerevisiae*), 4 (size in amino acids), 5 (expected value), 6 and 7 (% amino acid identity and similarity in the aligned segment), 8 (% gaps), 9 (strand (+ or −) and frame (1, 2 or 3)), 10 (segment size in amino acids), 11 and 12 (first and last positions (in nuc.) of the aligned segment in the RST), 13 and 14 (first and last positions (in a.a.) of the aligned segment in the homologous protein). In cases of homologs to *S. cerevisiae* tRNA genes (example line 9) or other short elements (line 10), the coordinates of the segment are indicated in columns 11 and 12 (the orientation of the tRNA gene in the RST is indicated in column 9). Columns 4–14 are left empty for RSTs corresponding to fragments of rDNA (example line 12), Ty elements (line 11), or to non-chromosomal replicons such as mtDNA (example line 13) or plasmid DNA (example line 14). Validation codes (column 15) are **o**: a single homolog clearly identified, or **oo**: several possible homologs of nearly equivalent probabilities. Note that in cases of sequencing frameshift errors in RSTs, each aligned fragment corresponding to the same homolog is represented on an independent line (e.g. lines 3 and 4). Similarly, in cases where several alignments of an RST to distinct homologs are validated (several genes in the same RST), alignment figures corresponding to each homolog are represented on an independent line (e.g. lines 3, 4 and 5). A few cases of actual tandem gene duplication in the same RSTs were detected in this work and annotated accordingly in column 16 (not shown) to distinguish them from sequencing frameshift errors. In cases of ambiguous assignments, each possible homolog is represented on an independent line (e.g. lines 6 and 7).

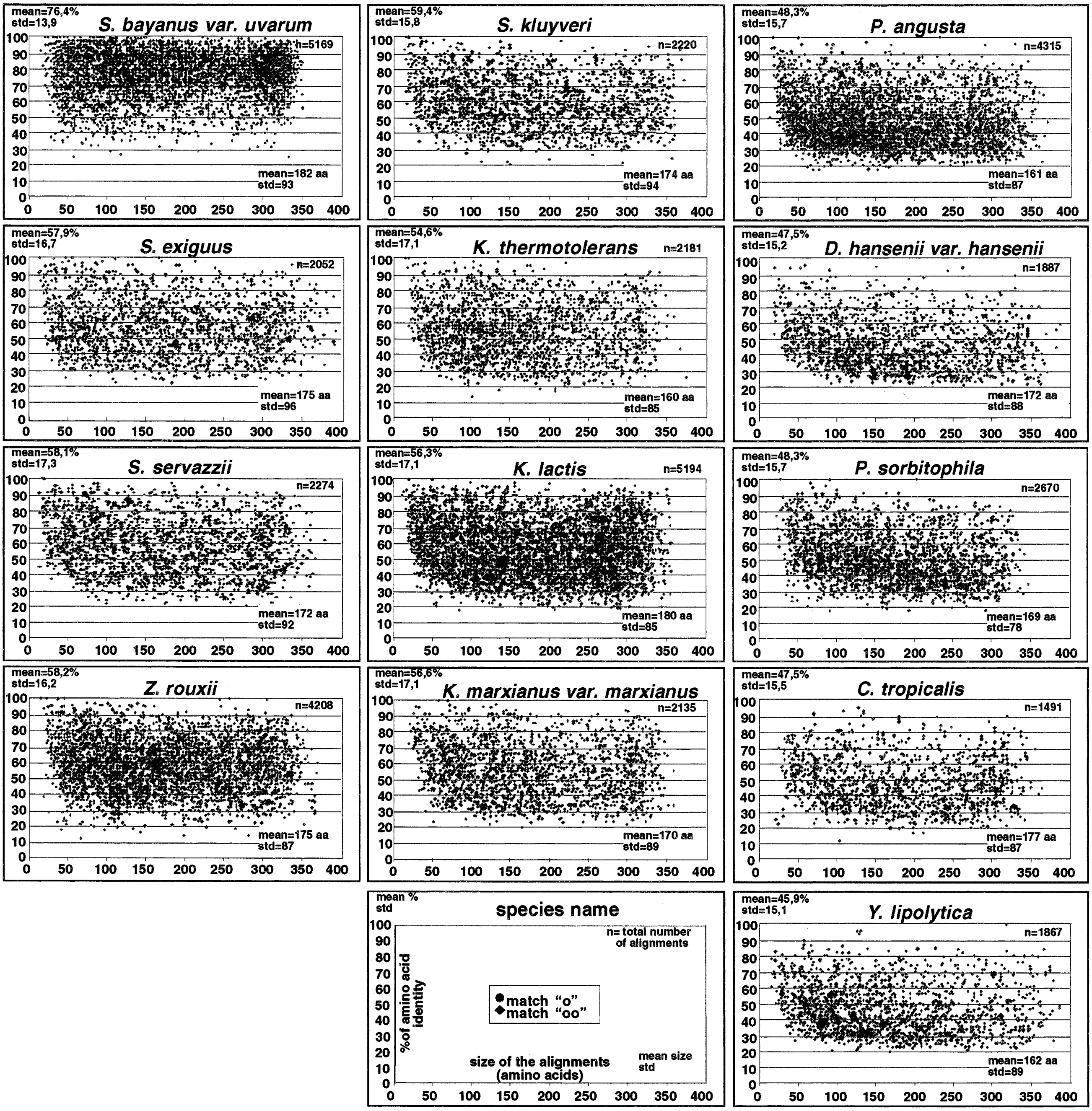


Fig. 3. Distributions of sizes and identities in validated *blast* alignments. For each yeast species, the figure represents the number of amino acids (abscissa) and % of amino acid identities (ordinates) of all o (blue) and oo (red) validated *blastx* alignments. In cases of multiple *blastx* alignments between a given RST and the same *S. cerevisiae* protein sequence due to frameshift sequencing error, the alignment with the lowest E-value was considered.

a general annotation table of all RSTs of each yeast species was produced (Table 3). Such tables, which contain the complete list of genetic elements found in each RST by similarity comparisons, were used to extract the annotations submitted to EBI, to construct Table 1, and for all subsequent interpretations given in the articles of this series.

In order to build the general annotation tables, *blast* comparisons were used and instead of trying to define a significance threshold for each species based on expected values, we have individually inspected the alignments to take into account the fact that RSTs fall at random relative to gene limits and may contain base addition/omission resulting in frameshift errors. Based on quality, size and positions of the alignments relative to the limits of the RST and to the position in the putative homolog, the homology was validated or rejected. In some cases, the validated alignment was not the one with the best expected value or the highest identity score. In other cases, several short segments of low significance were nevertheless validated if corresponding to successive sections of a same gene due to sequencing frameshifts in the RST.

In annotating our RSTs by comparison to *S. cerevisiae*, an additional complication originated from the existence of gene

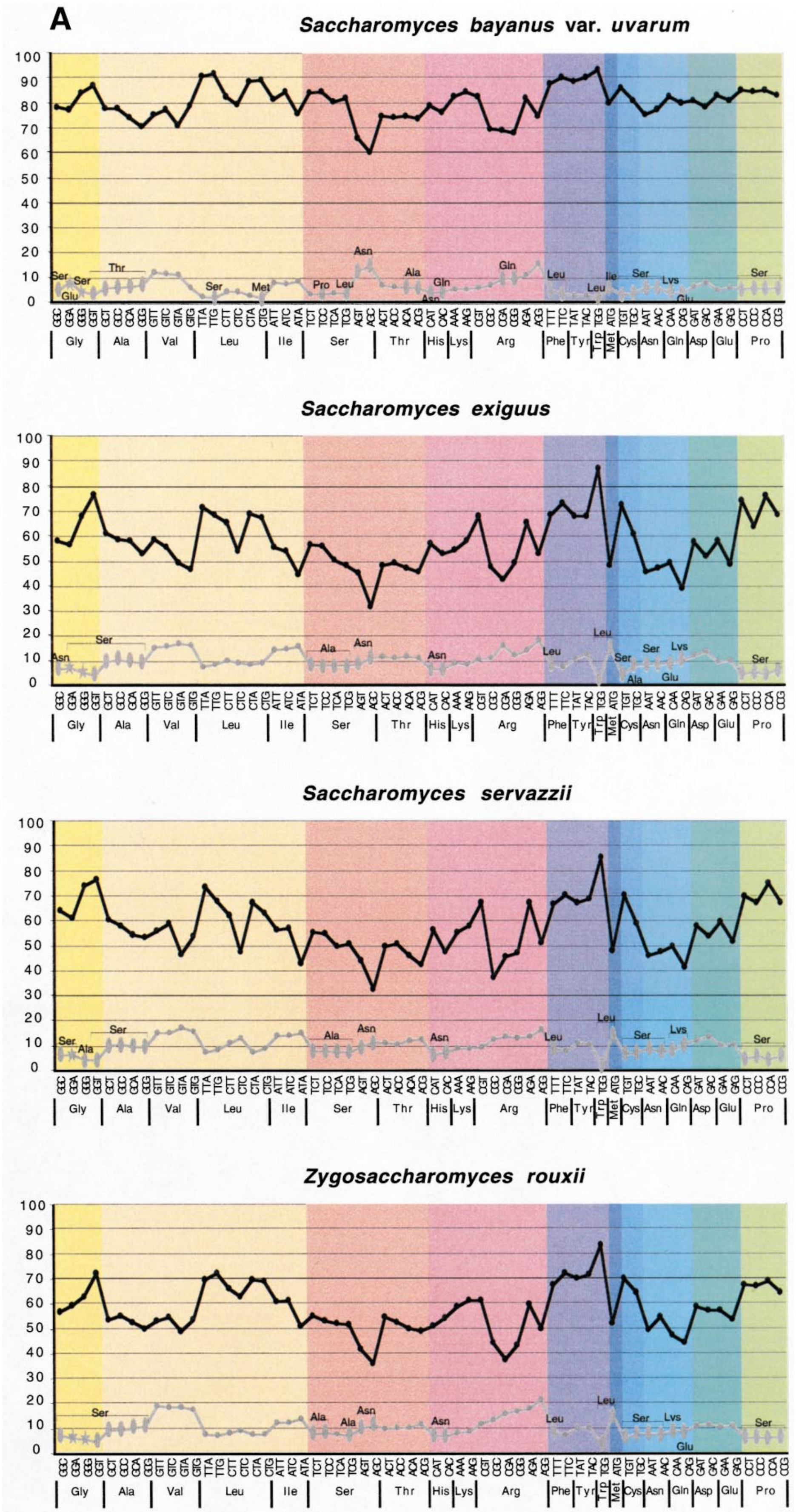

A
Saccharomyces bayanus var. uvarum
Saccharomyces exiguus
Saccharomyces servazzii
Zygosaccharomyces rouxii
100
90
80
70
60
50
40
30
20
10
0
GGC GGA GGG GGT GCT GCC GCA GCG GTT GTC GTA GTG TTA TTG CTT CTC CTA CTG ATT ATC ATA TCT TCC TCA TCG AGT AGC ACT ACC ACA ACG CAT CAC AAA AAG CGT CGC CGA CGG AGA AGG TTT TTC TAT TAC TGG ATG TGT TGC AAT AAC CAA CAG GAT GAC GAA GAG CCT CCC CCA CCG
Gly
Ala
Val
Leu
Ile
Ser
Thr
His
Lys
Arg
Phe
Tyr
Trp
Met
Cys
Asn
Gln
Asp
Glu
Pro
Ser
Glu
Thr
Met
Pro
Leu
Asn
Ala
Gln
Lys
Ile
Asn
Ala
Leu
Ser
Asn

Fig. 4.

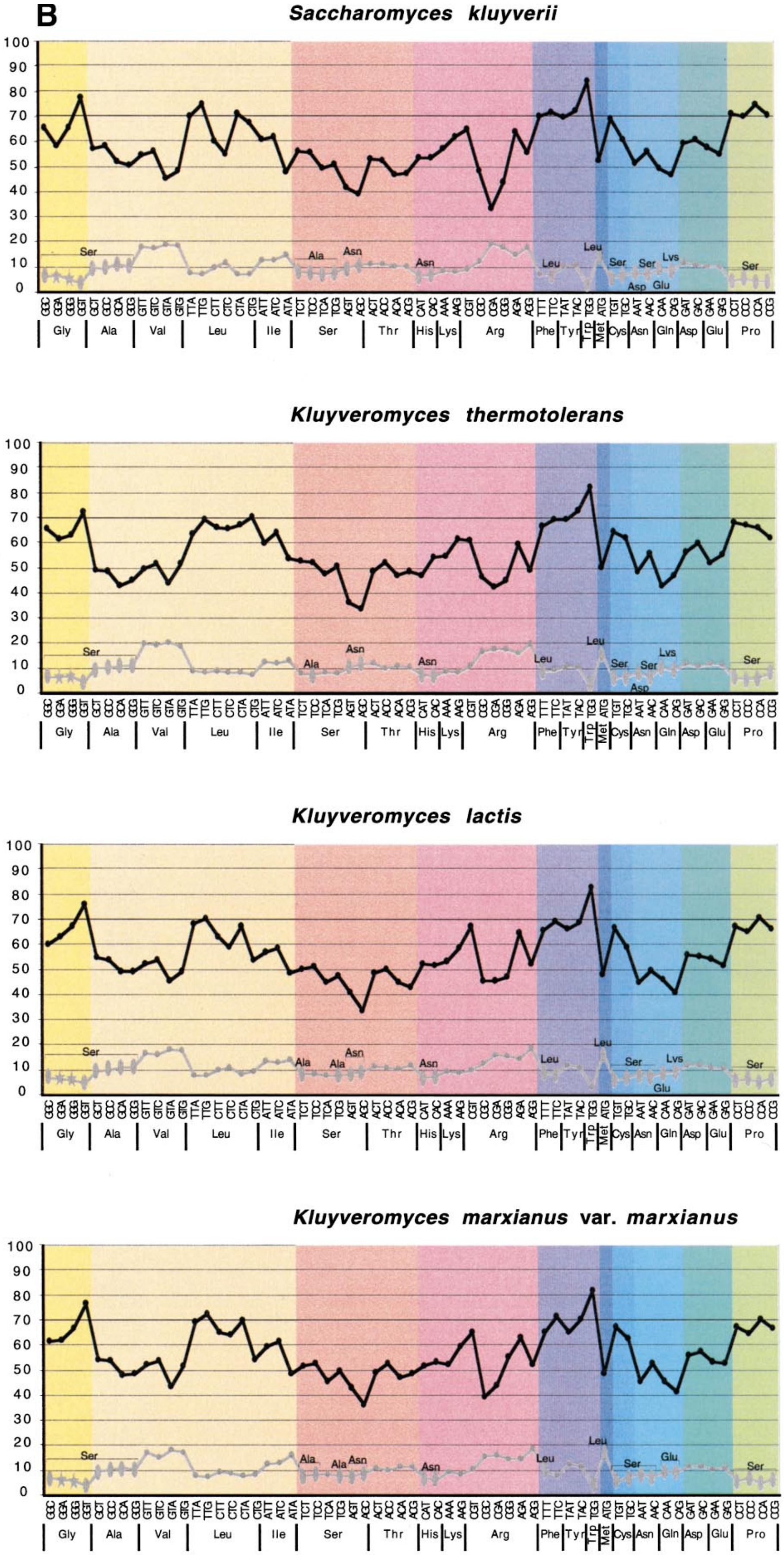


Fig. 4 (*continued*).

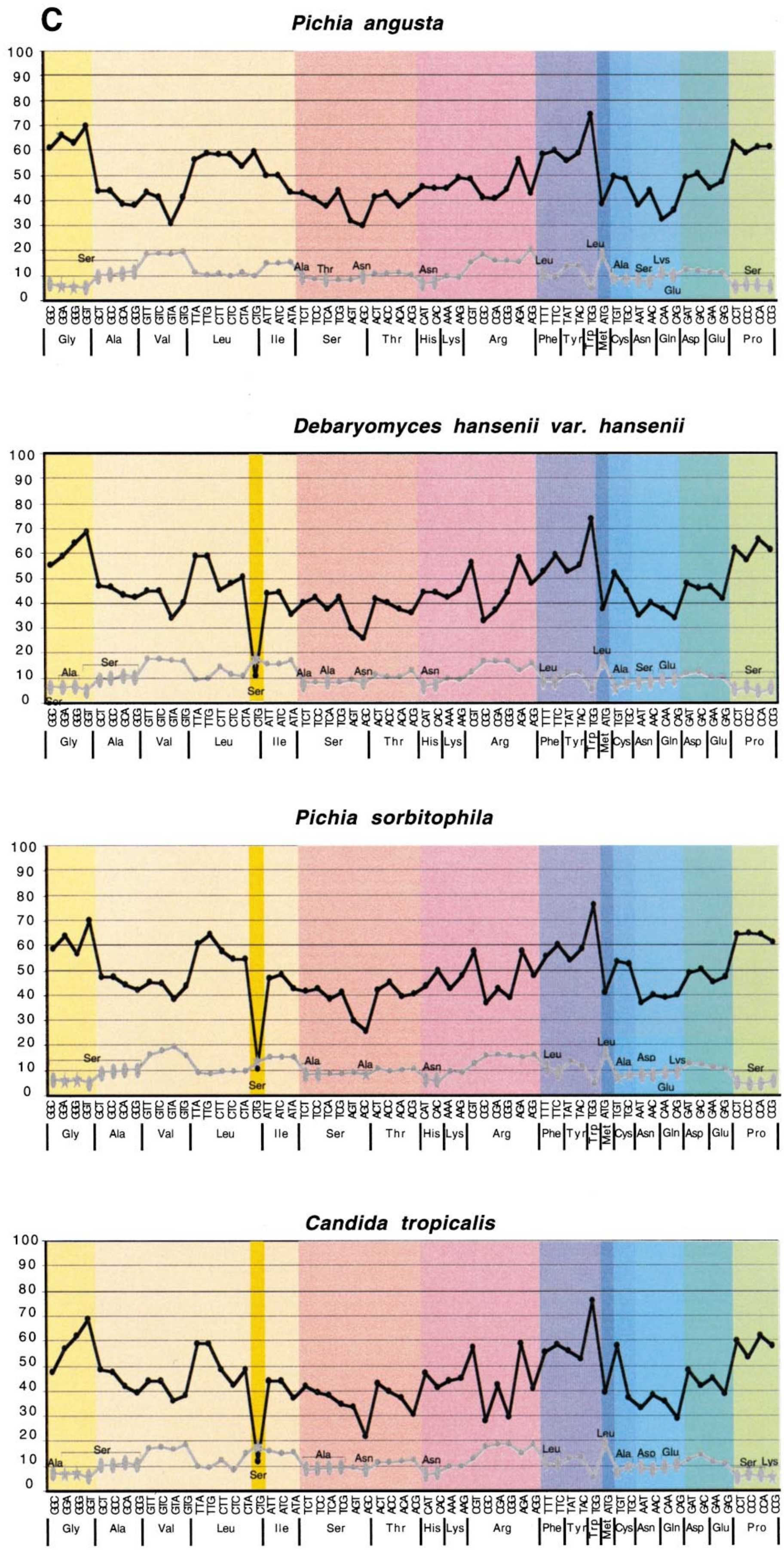


Fig. 4 (*continued*).

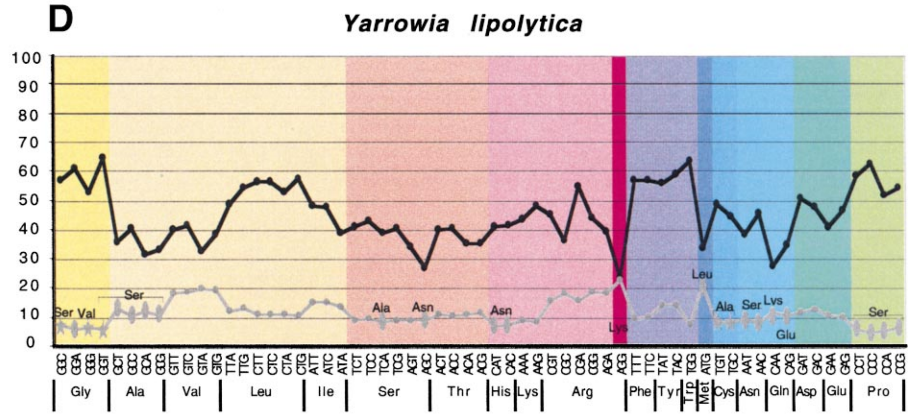


Fig. 4. Inferring the genetic code from the frequencies of *S. cerevisiae* amino acids corresponding to each codon of the studied yeast species. For each yeast species, the figure represents for each of the 61 sense codons, the frequency (in % of total of the corresponding *S. cerevisiae* segments aligned) of the corresponding amino acid according to the universal code (black) and the frequency and nature of the most frequent other amino acid (gray). All '**o**' validated *blastx* alignments were considered for computation. The frequencies of other amino acids are generally comprised between 0 and 5% of total, with few exceptions. Gold vertical bars indicate the non-universal CUG codon found in some species. The magenta vertical bars indicate the frequent arginine/lysine substitutions observed between *S. cerevisiae* and *Y. lipolytica*.

families in *S. cerevisiae*. If, in some cases, the sequence divergence was sufficient to decide which of the members of the family was the likely homolog (designated '**o**'), in other cases it was difficult to decide among several members (example in Fig. 2). We, therefore, used the annotation '**oo**' to designate each of the possible homologs.

Because the expert validation step was done by different labs and on yeast species of various evolutionary distances to *S. cerevisiae*, it is interesting to examine a posteriori the distributions of sizes and identities of the validated *blastx* alignments. This is shown by Fig. 3. It can be seen that, for all 13 species, the validated alignments range from less than 20 amino acids to more than 350 with average values close to 170 amino acids. Note that this result is the best demonstration of the high quality of our sequences because on average, segments of ca. 510 nucleotides in length are found without base addition/omission. Similarly, identity scores within the validated alignments range from ca. 25% to over 90%. In this case, however, the average values and distributions vary according to the species studied. For *S. bayanus* var. *uvarum*, for example, the closest relative to *S. cerevisiae* in the present program, the average identity score is clearly higher than for other yeast species. The converse is true for the five more distant species: *C. tropicalis*, *D. hansenii* var. *hansenii*, *P. angusta*, *P. sorbitophila* and *Y. lipolytica*. Overall, except for the fact that very short fragments with low identity scores are obviously under-represented, there is no clearcut relationship between the sizes and identity scores of the validated alignments, as expected for a good random genomic sequencing program. In particular, we do not notice a significant tendency for long fragments with low identity scores to dominate over shorter ones, which was one of our concerns in deciding the present strategy. Examination of Fig. 3 also shows that the severity applied in deciding between validation and rejection of the *blast* alignments was not very different for the different yeast species despite the fact that the analysis was done in different laboratories.

Despite the strategy used, it is important to stress that one cannot be certain that the homologies retained as '**o**' correspond to actual orthologous genes in all cases. In the example given by Fig. 2, the alignment with *YPR001w* (the *CIT3* gene) was rejected because the same RST segment shows a better alignment with either *YNR001c* (the *CIT1* gene) or *YCR005c* (the *CIT2* gene) both validated with the '**oo**' sign. If, by chance or otherwise, the last two genes were not present in the *S. cerevisiae* database, the alignment with *YPR001w* (*CIT3*) would have been validated as '**o**'. The absence of a gene in *S. cerevisiae* may, therefore, lead to an artifactual validation of another gene for homolog. In the example above, the misassignment would not lead to a biological misinterpretation because the three *S. cerevisiae* genes belong to the same family (the citrate synthases). But other cases may exist in which an RST may be considered to contain a homolog to a *S. cerevisiae* gene only for the absence of another true homolog.

This is one of the reasons why comparisons of our RSTs to organisms other than *S. cerevisiae* were so important. In order to be able to confront the results of such comparisons with those obtained for *S. cerevisiae*, we decided to apply a similar expert validation strategy to the *blast* alignments. *Blastx* searches were done against a compilation (designated *GPROTEOME*) of the protein sequences from the 23 first sequenced organisms (16 bacteria, six archaea and one eukaryote, see Table 1) to which were added partial data from *S. pombe* and data from SwissProt filtered from the above (see Section 2). Note that this compilation contains solely the protein sequences as offered by authors of the original sequence (Table 1). No attempt was made to reinterpret the original DNA sequences by comparisons to our RSTs to find possible additional genes not annotated. Also note that, consistent with our general strategy of giving priority to *S. cerevisiae*, a homolog to a given protein in *GPROTEOME* was not retained if the same RST fragment already had a homolog to a *S. cerevisiae* protein which itself is homologous to the *GPROTEOME* entry (see Section 2 and Fig. 1). The complete lists of additional genes found in the RST set of each yeast species are given in the respective articles of this series [37–49].

Note that, for the general annotation of our RSTs, we did not consider public protein databases such as PIR, TrEMBL or genpept and that, consequently, some homologies to sequenced genes of a variety of organisms may have been missed. Similarly, our systematic comparisons did not include the sequences of *Neisseria meningitidis* [57], *D. radiodurans* [58], *Drosophila melanogaster* [59], chromosomes 2 and 4 from *Arabidopsis thaliana* [60,61], chromosomes 2 and 3 from *Plasmodium falciparum* [62,63], and chromosomes 21 and 22 from *Homo sapiens* [64,65] that were recently reported. For the few cases of homologies to these organisms, refer to [37–49].

One of the limitations of a low coverage random sequencing program is the fact that the actual number of genes in the species of interest can only be estimated between minimum and maximum limits due to the possible existence of gene families that may not always be precisely defined. As indicated in Section 2 and discussed in details in [4], when two distinct RST segments from a given yeast species, not overlapping each other, share the same *S. cerevisiae* homolog, they may originate from the same gene or from distinct genes of the same family. The diploidy of some of the yeast species studied (see [6]) further complicates the problem by the fact that even if two partially overlapping RST segments differ in sequences, they may represent two heteroalleles of a single locus or two different genes of a conserved family. At the present stage of our work, this problem was not fully addressed.

Another limitation of our strategy concerns the genes that result from the fusion of distinct genes or modules found in other organisms. In the comparisons to *S. cerevisiae*, this possibility has been properly dealt with because all putative homologous segments were simultaneously examined during the expert validation step. But the hierarchical strategy used for the comparison to other organisms would misinterpret a gene made of two modules (A and B) in a given RST, if a gene having one module (e.g. B) is present in *S. cerevisiae* while no gene having the other module (e.g. A) is present. In such a case, the filtering strategy used would eliminate a gene in *GPROTEOME* having the two modules A and B, solely because B was already found in *S. cerevisiae*. Although this possibility exists, in practice it does not seem to have played an important role in the interpretation of our results.

3.2. Determination of the genetic code

A precise definition of the genetic code is of utmost importance in all comparative programs, such as ours, in which comparisons are made between protein sequences deduced from DNA sequences which are the only actual data. Deviations from the universal code were first reported in the mitochondria of a variety of organisms including man [66] and *S. cerevisiae* [67] and subsequently in bacteria (reviewed by [68]). Other deviations from the universal code were reported for the nucleo-cytoplasmic machinery of various Ciliates where the UAG and UAA codons were found to encode glutamine [69] or UGA to encode cysteine [70]. With regard to yeasts, it was originally discovered using cell-free translation experiments, that the standard leucine codon CUG encodes a serine in *Candida cylindracea* [71]. The same was subsequently found to apply to a variety of other *Candida* species, including *C. albicans* and *C. tropicalis* [72], while a larger exploration of the *Candida* genus showed 66 species using the CUG codon for serine and 11 others for leucine [73]. This last study included *Candida famata*, the anamorph of *D. hansenii* used in our program.

In order to infer the nuclear genetic code of each yeast species studied in this program, we have examined the frequency of each of the 20 amino acids at positions of the *S. cerevisiae* proteins corresponding, in validated *blastx* alignments, to each of the 64 codons of the DNA sequence from the yeast species of interest (see Section 2). Despite the fact that such frequency distributions must necessarily be noisy because of possible misalignments and of natural sequence divergence between the species, results are very clearcut (Fig. 4). For nine yeast species, a single amino acid corresponding to the universal code was clearly found, for each of the 61 significant codons, at frequencies above background defined by other amino acids. In such cases, the second-most frequent amino acid observed almost always corresponds to a conservative amino acid change. For the remaining four species, *C. tropicalis*, *D. hansenii*, *P. sorbitophila* and *Y. lipolytica*, the same holds true for all codons except CUG for the first three species and AGG for the last one. In the case of the CUG codons, the most frequently observed amino acid is not the leucine, as expected from the universal code, but a serine. Leucine comes in second for frequency. In the case of the AGG codon of *Y. lipolytica*, two amino acids are found with similar frequencies, arginine, as predicted from the universal code, and its conservative replacement, lysine.

We conclude from this computation that, consistent with previous results, the CUG codon encodes a serine in *C. tropicalis* and *D. hansenii*, and that the same is true for *P. sorbitophila*, not previously studied. This is in agreement with the phylogenetic position of this species proposed in this work [6]. For *Y. lipolytica*, there is no indication that the code may not be universal because the ambiguity observed in our computation results from the frequent replacement of the AAG codons (lysine) found in *S. cerevisiae* by AGG codons (arginine), is consistent with the overall high GC content of *Y. lipolytica*.

Now, all sequence comparisons of this program were made after translation using the universal code for all 13 yeast species. To be rigorous, *blastx* comparisons should be repeated in *C. tropicalis*, *D. hansenii* and *P. sorbitophila* using properly

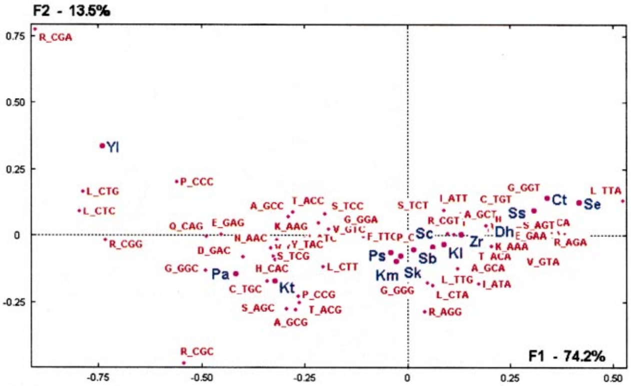


Fig. 5. Correspondence analysis of codon usage. RSCU values, calculated for each yeast species from '**o**' validated sequence alignments with *S. cerevisiae*, were analyzed using correspondence analysis. Yeast species, codons and corresponding amino acids are plotted on the first and second axis, which together represent nearly 88% of the total information of the matrix. Average GC content of protein-coding sequences of each yeast species, calculated from '**o**' validated sequence alignments with *S. cerevisiae*, are: *Y. lipolytica*: 53.0% (Yl); *P. angusta*: 48.5% (Pa); *K. thermotolerans*: 47.3% (Kt); *K. marxianus* var. *marxianus*: 42.3% (Km); *P. sorbitophila* (Pa) and *S. kluyverii*: 41.5% (Sk); *K. lactis* (Kl) and *S. bayanus*. var. *uvarum*: 40.2% (Sb); *Z. rouxii*: 39.5% (Zr); *D. hansenii* var. *hansenii*: 36.5% (Dh), *S. servazzii*: 34.7% (Ss); *C. tropicalis*: 34.6% (Ct) and *S. exiguus*: 33.0% (Se). For comparison, the average GC content of protein-coding sequences in *S. cerevisiae* is 39.6% (Sc).

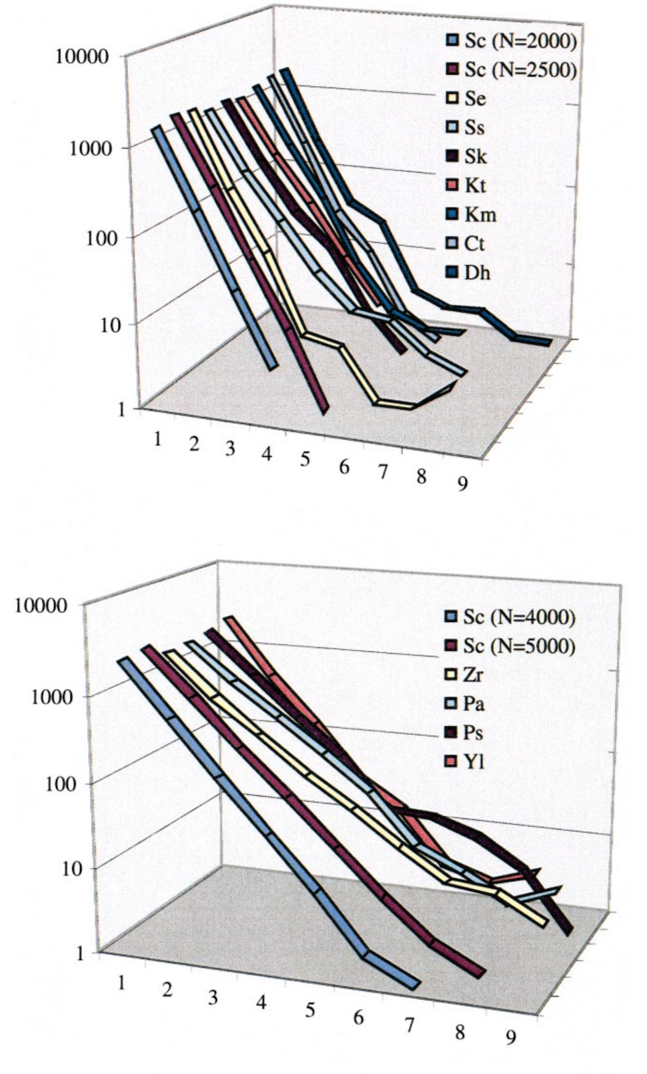


Fig. 6. Distribution of contigs from the genomic libraries. According to [55], the expected number of 'islands' consisting of j RSTs ($j \geq 1$) is given by: $I_j = N\mathrm{e}^{-2c\sigma}(1-\mathrm{e}^{-c\sigma})^{j-1}$ where N is the number of RSTs, c is the relative genome coverage ($c = NL/G$) and σ depends upon the fraction of the sequence needed to detect an overlap (islands are contigs of two members or more plus the singletons ($j = 1$)). The curves shown were calculated for the genome size of *S. cerevisiae* ignoring rDNA ($G = 12\,069\,298$ nucleotides), for an average RST length of 910 nucleotides (see [2]) and for a minimum overlap detection of 45 nucleotides ($\sigma = 0.95$) and for 2000 and 2500 sequences (upper part) or for 4000 and 5000 sequences (lower part). Other curves represent observed results for other yeast species in which ca. 2500 (upper part) or ca. 5000 (lower part) RSTs were determined (abbreviations as in Fig. 5). Ordinates: number of islands (log scale), abscissae: number of RST per island. Species name abbreviations as in Fig. 5.

translated sequences. However, the CUG codon is rare enough in all three species not to significantly contribute to misalignments and loss of detectable homology (respectively, 0.3% of the 12 221 serine codons used in *C. tropicalis*). This was directly confirmed, in the case of *C. tropicalis*, by a *blastx* search using the alternative yeast genetic code [48].

3.3. Codon usage and general base composition

We used the relative synonymous codon usage (RSCU) parameter of [54] to characterize the codon usage of each yeast species studied in this work. This parameter indicates, for each amino acid, codon preference among the synonyms. For each of the 13 surveyed yeast species, a matrix of 59 RSCU values (all codons except stops and those encoding Met and Trp) was computed from the validated *blast* alignments, and compared to those of *S. cerevisiae*. For *C. tropicalis*, *D. hansenii* and *P. sorbitophila*, consistent with results of the previous paragraph, five synonymous codons were considered for leucine and seven codons for serine. The universal code was considered otherwise. Correspondence analysis [74] was used to analyze the information from this matrix. The first factorial axis obtained (Fig. 5) indicates a clustering of the yeast species that coincides reasonably well with the average GC content of their coding regions as deduced from the validated *blast* alignments. Namely, *Y. lipolytica*, *P. angusta* and *K. thermotolerans*, which have the highest GC content tend to favor GC-rich codons, *S. exiguus*, *C. tropicalis*, *S. servazzii* and *D. hansenii* var. *hansenii* which have the lowest GC content tend to favor AT-rich codons, and the six remaining species present intermediate values, similar to that of *S. cerevisiae*.

3.4. Randomness of the genomic libraries

One of the major concerns for the type of analysis performed in the subsequent articles of this series, is the randomness of the genomic libraries used. Biased libraries not only would result in a significant loss of efficiency of our sequencing program, but are prone to incorrect interpretations in terms of yeast-specific genes [3], gene family distributions [4], functional analysis [5], or map comparisons [2]. For each of the 13 yeast species, we have, therefore, compared the number of contigs obtained from the set of RSTs to the theoretical distribution predicted from the work of [55,56]. This comparison is complicated by the fact that our genomic libraries were made from total yeast DNA (thus containing mitochondrial and plasmid DNA which occur in multiple copies) and by the existence, in genuine chromosomal DNA, of repeated sequences such as rDNA or transposons. Therefore, for each yeast species we have first classified the contigs or the single RSTs that belong to any of the above and eliminated them from the computation (see Section 2). The theoretical distribution of contigs, calculated for the *S. cerevisiae* genome, is given by Fig. 6, along with actual data for the other yeast species. It can be seen that, in general, the distribution of contigs observed fits reasonably well the theoretical distribution of *S. cerevisiae*, indicating that the genomic libraries used are not significantly biased.

Acknowledgements: We thank J.C. Aude, J.J. Codani, A. Perrin, and M. Termier for fruitful discussions and H. Feldmann and A. Goffeau for careful reading of the manuscript. This work was supported in part by grants from CNRS (Programme Génome 1997–99), and by a contract from the European Commission (BIOCT97-2294). B.D. is a member of Institut Universitaire de France.

References

[1] Blandin et al. (2000) FEBS Lett. 487, 31–36 (this issue).
[2] Llorente et al. (2000) FEBS Lett. 487, 101–112 (this issue).
[3] Malpertuy et al. (2000) FEBS Lett. 487, 61–65 (this issue).
[4] Llorente et al. (2000) FEBS Lett. 487, 71–75 (this issue).
[5] Gaillardin et al. (2000) FEBS Lett. 487, 134–149 (this issue).
[6] Souciet, J.-L. et al. (2000) FEBS Lett. 487, 3–12 (this issue).
[7] Artiguenave et al. (2000) FEBS Lett. 187, 13–16 (this issue).

[8] Hartley, J.L. and Donelson, J.E. (1980) Nature 286, 860–865.
[9] Johnston, M., Hillier, L., Riles, L., Albermann, K. and André, B. et al. (1997) Nature 387 (Suppl.), 87–90.
[10] Hani, J. and Feldmann, H. (1998) Nucleic Acids Res. 26, 689–696.
[11] Foury, F., Roganti, T., Lecrenier, N. and Purnelle, B. (1998) FEBS Lett. 440, 325–331.
[12] Deckert, G., Warren, P.V., Gaasterland, T., Young, W.G. and Lenox, A. et al. (1998) Nature 392, 353–358.
[13] Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M. and Alloni, G. et al. (1997) Nature 390, 249–256.
[14] Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G. and Clayton, R.A. et al. (1997) Nature 390, 580–586.
[15] Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M. and Churcher, C. et al. (2000) Nature 403, 665–668.
[16] Kalman, S., Mitchell, W., Marathe, R., Lammel, C. and Fan, J. et al. (1999) Nat. Genet. 21, 385–389.
[17] Stephens, R.S., Kalman, S., Lammel, C., Fan, J. and Marathe, R. et al. (1998) Science 282, 754–759.
[18] Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T. and Burland, V. et al. (1997) Science 277, 1453–1462.
[19] Fleischman, R.D., Adams, M.D., White, O., Clayton, R.A. and Kirkness, E.F. et al. (1995) Science 269, 496–512.
[20] Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A. and Sutton, G.G. et al. (1997) Nature 388, 539–547.
[21] Cole, S.T., Brosch, R., Parkhill, J., Garnier, T. and Churcher, C. et al. (1998) Nature 393, 537–544.
[22] Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D. and Clayton, R.A. et al. (1995) Science 270, 397–403.
[23] Himmelreich, H.R., Plagens, H., Hilbert, H. and Hermann, R. (1996) Nucleic Acids Res. 24, 4420–4449.
[24] Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T. and Alsmark, U.C. et al. (1998) Nature 396, 133–140.
[25] Kaneko, T., Sato, S., Kotani, H., Tanaka, A. and Asamizu, E. et al. (1996) DNA Res. 3, 109–136.
[26] Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L. and Dodson, R.J. et al. (1999) Nature 399, 323–329.
[27] Fraser, C.M., Norris, S.J., Weinstock, G.M., White, O. and Sutton, G.G. et al. (1998) Science 281, 375–388.
[28] Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y. et al. (1999) DNA Res. 6, 83–101, 145–152.
[29] Klenk, H.-P., Clayton, R.A., Tomb, J.-F., White, O. and Nelson, K.E. et al. (1997) Nature 390, 364–370.
[30] Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H. and Dubois, J. et al. (1997) J. Bacteriol. 179, 7135–7155.
[31] Bult, C.J., White, O., Olsen, G.J., Zhou, L. and Fleischmann, R.D. et al. (1996) Science 273, 1058–1073.
[32] Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y. and Hino, Y. et al. (1998) DNA Res. 5, 55–76.
[33] The *C. elegans* sequencing consortium (1998) Science 282, 2012–2018.
[34] Bairoch, A. and Apweiler, R. (2000) Nucleic Acids Res. 28, 45–48.
[35] Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Genome Res. 8, 175–185.
[36] Ewing, B. and Green, P. (1998) Genome Res. 8, 186–194.
[37] Bon et al. (2000) FEBS Lett. 487, 37–41 (this issue).
[38] Bon et al. (2000) FEBS Lett. 487, 42–46 (this issue).
[39] Casaregola et al. (2000) FEBS Lett. 487, 47–51 (this issue).
[40] de Montigny, J. et al. (2000) FEBS Lett., (this issue).
[41] Neuvéglise et al. (2000) FEBS Lett. 487, 56–60 (this issue).
[42] Malpertuy et al. (2000) FEBS Lett. 487, 52–55 (this issue).
[43] Bolotin-Fukuhara, M. et al. (2000) FEBS Lett. 487, 66–70 (this issue).
[44] Llorente et al. (2000) FEBS Lett. 487, 71–75 (this issue).
[45] Blandin et al. (2000) FEBS Lett. 487, 76–81 (this issue).
[46] Lépingle et al. (2000) FEBS Lett. 487, 82–86 (this issue).
[47] de Montigny et al. (2000) FEBS Lett. 487, 87–90 (this issue).
[48] Blandin et al. (2000) FEBS Lett. 487, 91–94 (this issue).
[49] Casaregola et al. (2000) FEBS Lett. 487, 95–100 (this issue).
[50] Wootton, J.C. and Federhen, S. (1993) Comput. Chem. 17, 149–163.
[51] Altschul, S.F. et al. (1997) Nucleic Acids Res. 25, 3389–3402.
[52] Tekaia, F. and Dujon, B. (1999) J. Mol. Evol. 49, 591–600.
[53] Tekaia, F., Lazcano, A. and Dujon, B. (1999) Genome Res. 9, 550–557.
[54] Sharp, P.M. and Li, W.H. (1987) Nucleic Acids Res. 15, 1281–1295.
[55] Lander, E.S. and Waterman, M.S. (1988) Genomics 2, 231–239.
[56] Port, E., Sun, F., Martin, D. and Waterman, M.S. (1995) Genomics 26, 84–100.
[57] Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C. and Nelson, K.E. et al. (2000) Science 287, 1809–1815.
[58] White, O., Eisen, J.A., Heidelberg, J.F., Hickey, E.K. and Peterson, J.D. et al. (1999) Science 286, 1571–1577.
[59] Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A. and Gocayne, J.D. et al. (2000) Science 287, 2185–2195.
[60] Lin, X., Kaul, S., Rounsley, S.D., Shea, T.P. and Benito, M.-I. et al. (1999) Nature 402, 761–768.
[61] Mayer, K., Schuller, C., Wambutt, R., Murphy, G. and Volckaert, G. et al. (1999) Nature 402, 769–777.
[62] Gardner, M.J., Tettelin, H., Carucci, D.J., Cummings, L.M. and Aravind, L. et al. (1998) Science 282, 1126–1132.
[63] Bowman, S., Lawson, D., Basham, D., Brown, D. and Chillingworth, T. et al. (1999) Nature 400, 532–538.
[64] Dunham, I. et al. (1999) Nature 402, 489–495.
[65] Hattori, M. et al. (2000) Nature 405, 311–319.
[66] Barrell, B.G., Bankier, A.T. and Drouin, J. (1999) Nature 282, 189–194.
[67] Macino, G., Coruzzi, G., Nobrega, F.G., Li, M. and Tzagoloff, A. (1979) Proc. Natl. Acad. Sci. USA 76, 3784–3785.
[68] Osawa, S., Muto, A., Ohama, T., Andachi, Y., Tanaka, R. and Yamao, F. (1990) Experientia 46, 1097–1106.
[69] Horowitz, S. and Gorovsky, M.A. (1985) Proc. Natl. Acad. Sci. USA 82, 2452–2455.
[70] Caron, F. and Meyer, E. (1985) Nature 314, 185–188.
[71] Kawaguchi, Y., Honda, H., Taniguchi-Morimura, J. and Iwasaki, S. (1989) Nature 341, 164–166.
[72] Ueda, T., Suzuki, T., Yokogawa, T., Nishikawa, K. and Watanabe, K. (1994) Biochimie 76, 1217–1222.
[73] Sugita, T. and Nakase, T. (1999) Syst. Appl. Microbiol. 22, 79–86.
[74] Greenacre, M. (1984) Theory and Application of Correspondence Analysis, Academic Press, London.