Phylogenetic classification of the major superfamily of membrane transport facilitators, as deduced from yeast genome sequencing

Bart Nelissen[b], Philippe Mordant[a], Jean-Luc Jonniaux[a], Rupert De Wachter[b], André Goffeau[a,*]

[a] *Unité de Biochimie Physiologique, Faculté des Sciences Agronomiques, Université Catholique de Louvain, Place Croix du Sud 2-20, B-1348 Louvain-la-Neuve, Belgium*

[b] *Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1, B-2610 Antwerpen, Belgium*

Received 16 November 1995

Abstract **From the approximately 5000 open reading frames presently identified by systematic sequencing of the yeast genome, 100 *Saccharomyces cerevisiae* transport proteins belonging to the major facilitator superfamily (MFS), were assigned to 17 families on the basis of extensive database searches and binary comparisons. These families include multidrug resistance proteins and transport proteins for sugars, amino acids, uracil/allantoin, allantoate, phosphate, purine/cytosine, proteins, peptides, potassium, sulfate, and urea. Four new families of unknown function have been identified. For the sugar and amino acid transport proteins, alignments were made and phylogenetic trees were constructed allowing the identification of several clusters of proteins presumably exhibiting similar transport functions.**

Key words: Transport protein; Yeast genome; Major facilitator superfamily (MFS)

1. Introduction

Transport proteins can be classified into several superfamilies, the members of which are found in all living species from mycoplasma to man. One of these transport protein superfamilies, is the major facilitator superfamily or MFS [1], characterized by two structural units of a 6 transmembrane-spanning helical segment, connected by a cytoplasmic loop, resulting in proteins with about 500 to 600 amino acids and 12 transmembrane helices. The proteins of the MFS superfamily have been divided into six families [2]. Other transport protein families that are characterized by a structural motif of 12 transmembrane-spanning helical segments include the amino acid-polyamine-choline (APC) family, and the sodium:solute symporter (SSF) family [2].

The sequence of the *Saccharomyces cerevisiae* genome is almost completed [3–5]. Because this is the first complete eukaryote sequence becoming available, *Saccharomyces cerevisiae* is very well suited for a study of the function and classification of transport proteins, which may serve as a model for other eukaryotes.

In this paper we have classified the 12 transmembrane-spanning transport proteins of the major facilitator superfamily. A preliminary grouping has been based on database searches of 12 transmembrane-spanning query sequences. The consistency of these groupings into families has been investigated by binary comparisons of all retrieved amino acid sequences. Multiple alignments were made for the largest groups in order to study the relationships between the constituent proteins by tree construction.

2. Methods

2.1. Classification into families

The 1884 non-redundant open reading frames from the *Saccharomyces cerevisiae* chromosomes I, II, III, V, VIII, IX, XI and part of other chromosomes, available in March 1995, were retrieved from the EMBL, GenBank, PIR, SwissProt, MIPS, SYDB, and YPD databases. These sequences were first screened according to their number of transmembrane spans as predicted by the KKD algorithm [6], with the threshold value of 15 for the peripheral/integral odds as described by [3,4]. To be sure to include all 12 transmembrane-spanning proteins, all proteins with 8 or more predicted transmembrane spans were used.

A BLAST [7] search of all amino acid sequences with 8 or more predicted transmembrane spans was carried out by the BLAST e-mail server version 1.4 at the National Center for Biotechnology Information (Bethesda, MD). All sequences producing high-scoring segment pairs with a $P(N) < 10^{-9}$ were considered to be closely related. All query sequences that had at least one closely related sequence in common, were placed in the same family. Those families that did not belong to the major facilitator superfamily (MFS) as deduced from their function in the BLAST results, e.g. the ATP-binding cassette (ABC) superfamily, were excluded from further analysis. All closely related yeast sequences that did belong to the MFS families but that were not yet in our dataset were retrieved. Starting from this dataset, the validity of each family was investigated by binary comparison of all protein sequences with each other. These binary comparisons were done with PRSS, a program for testing the significance of a protein sequence similarity, which belongs to the FASTA [8,9] software package version 1.7. For each comparison, 100 shuffles were done. A protein sequence was assigned to a family when its PRSS P-value with at least one member of the family was below 10^{-9} (Goffeau et al., unpublished results). When suspected, frame shifts were detected and corrected with the software package DNA Strider version 1.2 (Centre d'Etudes Nucleaires de Saclay, France).

2.2. Alignment of amino acid sequences

The amino acid sequences of the sugar and amino acid permease family were aligned with the multiple alignment program PILEUP, which belongs to the Wisconsin Sequence Analysis Package [10], version 8.0.

2.3. Phylogenetic tree construction

On the basis of the alignments dissimilarity matrices were calculated. Dissimilarities were converted into distances, assuming [11,12] that the rate of amino acid substitution follows the Poisson distribution, using the equation $D_{AB} = -\ln(1-S)$, where D is the evolutionary distance between two proteins A and B, and S the fraction of different amino acids (dissimilarity) between two sequences. Phylogenetic trees were constructed using the neighbor-joining method [13]. Distance matrix calculation and tree construction were done with the software package TREECON for Windows [14] version 1.1.

*Corresponding author. Fax: (32) (10) 47 38 72.
E-mail: GOFFEAU@FYSA.UCL.AC.BE

SSDI 0014-5793(95)01380-6

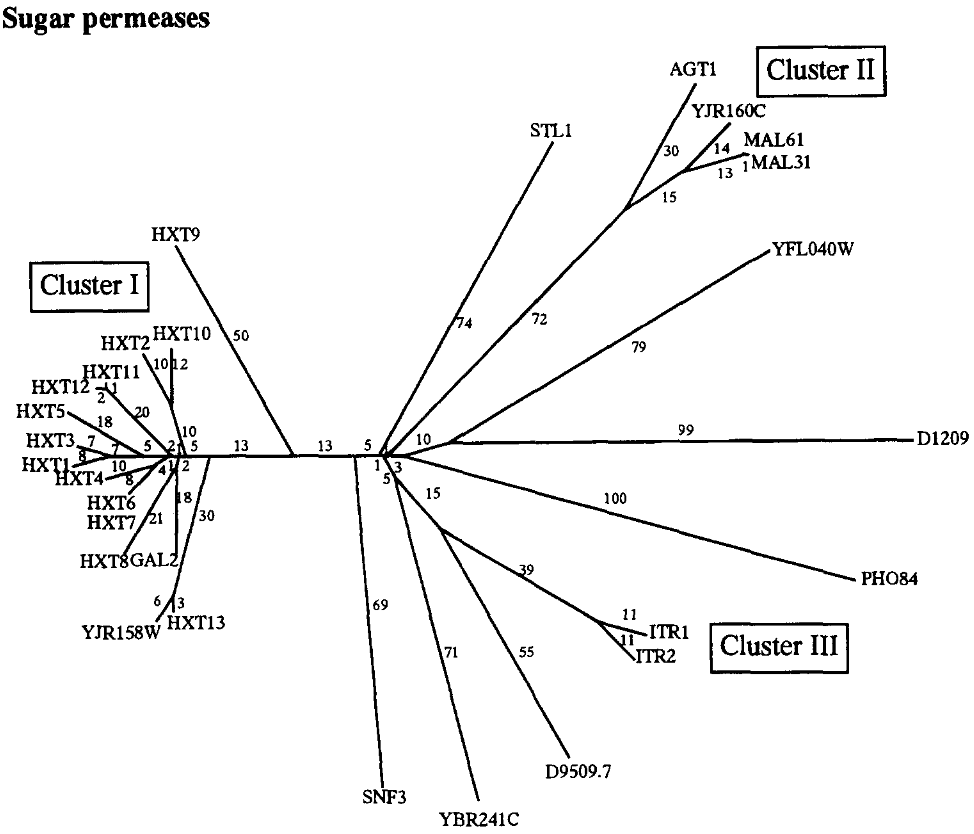

Fig. 1. Phylogenetic tree of the sugar permeases belonging to the MFS. Each number corresponds to the phylogenetic distance D multiplied by a factor 100. Proteins are considered to belong to the same cluster if $D \leq 0.9$ (arbitrary value). The exact composition of each cluster can be found in Table 1.

3. Results

3.1. Division into families

After prediction of the number of transmembrane spans, a BLAST search, and retrieval of the related proteins not yet in our dataset, 78 proteins belonging to the MFS were assigned to 16 families. Two frame shifts, probably the result of sequencing errors, were corrected. This resulted in joining ORFs YCL070C, YCL071C, and YCL073C into YCL070–73C, and in joining YIL170W and YIL171W into HXT12. Binary comparisons of all amino acid sequences finally resulted in 17 families, comprising 75 sequences.

In order to update the composition of these families, a new BLAST search and PRSS binary comparisons were carried out in October 1995. This resulted in the same 17 families comprising 100 sequences given in Table 1.

3.2. Phylogenetic trees

After alignment of the protein sequences of the sugar and amino acid permeases, phylogenetic trees were constructed as illustrated in Figs. 1 and 2.

4. Discussion

4.1. Sugar permeases

The family of sugar permeases comprises 28 representatives. On the basis of the phylogenetic tree, 21 representatives can be assigned to three different clusters, while the remaining representatives have no close relatives (Fig. 1, Table 1). Cluster I is the largest cluster and contains 15 proteins, mainly hexose/glucose permeases (HXT1-HXT13). This is surprising, even though glucose is an important substrate for *Saccharomyces cerevisiae*. The remaining representatives are a galactose permease (GAL2) and a protein (YJR158W) that is closely related to HXT13 and is thus probably a hexose/glucose permease. Cluster II contains 4 transport proteins, which include two maltose permeases (MAL31 and MAL61), one alpha-glucoside permease (AGT1), and a protein (YJR160C) that is related to the two maltose permeases. Cluster III contains 2 myo-inositol permeases (ITR1 and ITR2). The unclustered proteins consist mainly of permeases with an unknown substrate. Remarkably, the phosphate permease PHO84 belongs to this family of sugar permeases and not to another family that contains the phosphate permease PHO87.

4.2. Amino acid permeases

The family of amino acid permeases is the second largest family and comprises 19 representatives. As can be seen in the phylogenetic tree, 12 representatives can be assigned to 2 clusters (Fig. 2, Table 1). Cluster I contains 9 proteins, including a general amino acid permease (GAP1), branched amino acid permeases (BAP2 and YD9609.02), glutamine permeases (GNP1 and YCL025C), a histidine permease (HIP1), a tryptophan permease (SCM2), and a valine/leucine/isoleucine/tyrosine/tryptophan permease (TAT1). The functions of proteins YD6909.02 and YCL025C can be deduced from their relationship with BAP2 and GNP1 respectively, but L0555 is only loosely related to GAP1 and no function can be deduced. Cluster II contains 3 proteins that are basic amino acid permeases (APL1, CAN1, and LYP1). The unclustered proteins consist of a choline permease (CTR1), a proline permease (PUT4), and a GABA (4-aminobutyric acid) permease (UGA4). The remaining proteins (YBR132C, YD8358.14, YFL055W, and YKL174C) belong to the amino acid permease family, but their exact substrate is not known.

4.3. Multidrug resistance proteins

The multidrug resistance proteins (MDR) are subdivided

Table I
Families identified within the Major Facilitator Superfamily by BLAST and PRSS

Gene name (synonyms)[a]	Access. No[b]	Function
	SUGAR PERMEASES	
CLUSTER I		
GAL2 (IMP1)	P13181	galactose permease
HXT1 (YHR094C)	P32465	glucose permease, low-affinity
HXT2 (YM8270.15)	P23585	glucose permease, modulated affinity
HXT3	P32466	glucose permease, low-affinity
HXT4 (LGT1, RAG1, YHR092C)	P32467	glucose permease, moderate- to low- affinity
HXT5 (YHR096C)	P38695	hexose permease
HXT6	P39003	hexose permease, high-affinity
HXT7	P39004	hexose permease, high-affinity
HXT8 (YJL214W, HRA569)	P40886	similar to hexose permease HXT4
HXT9 (HXT 14, N0345)	P42833	hexose permease
HXT10 (YFL011W)	P43581	hexose permease
HXT11 (YJL219W, HRC567, LGT3)	P40885	glucose permease, low-affinity
HXT12 (YIL170W, YI9402.06B) (YIL171W, YI9402.06A)	P40441 P40440	similar to sugar permeases (frame shift: YIL170W and YIL171W joined)
HXT13 (YEL069C, HXT8)	P39924	hexose permease
YJR158W	Z49658x1	similar to sugar permeases
CLUSTER II		
AGT1	L47346x1	alpha-glucoside permease
MAL31 (MALK3T, YBR2116, YBR298C)	P38156	maltose permease
MAL61 (MAL6T)	P15685	maltose permease
YJR160C	Z49660x1	similar to sugar permeases
CLUSTER III		
ITR1	P30605	myo-inositol permease (major)
ITR2 (HRB612)	P30606	myo-inositol permease (minor)
UNCLUSTERED		
D1209	X83276x2	similar to sugar permeases
D9509.7	U32274x7	similar to ITR1
PHO84 (YM7056.03)	P25297	phosphate permease, high-affinity
SNF3	P10870	similar sugar permeases
STL1	P39932	sugar permease
YBR241C (YBR1625)	P38142	similar to sugar permeases
YFL040W	P43562	similar to sugar permeases
	AMINO ACID PERMEASES	
CLUSTER I		
BAP2 (YBR068C, YBR0629)	P38084	leucine / valine / isoleucine permease
GAP1 (YKR039W)	P19145	general amino acid permease
GNP1	U33057x14	glutamine permease, high-affinity
HIP1 (G7572)	P06775	histidine permease
L0555	Z47973x10	similar to GAP1
TAT2 (SCM2, TAP2, LTG3)	P38967	tryptophan permease, high-affinity
TAT1 (VAP1, TAP1, YBR710, YBR069C)	P38085	valine / leucine / isoleucine / tyrosine / tryptophan permease
YCL025C (YCC5)	P25376	similar to GNP1
PAP1 (YD9609.0)	P41815	similar to amino acid permeases
CLUSTER II		
ALP1 (APL1)	P38971	similar to basic amino acid permeases CAN1 and LYP1
CAN1 (YEL063C)	P04817	arginine / lysine / ornithine permease
LYP1	P32487	lysine permease, high-affinity
UNCLUSTERED		
CTR1 (HNM1)	P19807	choline permease
PUT4	P15380	proline permease, high-affinity
UGA4	P32837	GABA-specific permease, high-affinity
YBR132C (YBR1007)	P38090	similar to amino acid permeases
YD8358.14	Z50046x14	similar to amino acid permeases
YFL055W	P43548	similar to amino acid permeases
YKL174C (YKL639)	P36029	similar to CTR1 permease
	MULTIDRUG RESISTANCE PROTEINS, FAMILY 1	
HOL1	L42348x1	similar to YBR043C and YHR048C
P9584.7	U28371x3	similar to YBR008C
YBR008C (YBR0120)	P38124	similar to multidrug permeases
YBR043C (YBR0413)	P38227	similar to multidrug permeases
YBR180W (YBR1242)	P38125	similar to multidrug permeases
YHR048W	P38776	similar to multidrug permeases
YIL120W (I8277.09)	P40475	similar to multidrug permeases
YIL121W (I8277.08)	P40474	similar YIL120W
YNL1613	U12141x3	similar to multidrug permeases
	MULTIDRUG RESISTANCE PROTEINS, FAMILY 2	
ATR1 (SNQ1, YM83390.03)	P13090	aminotriazole resistance protein
ORF_886916	X87941x8	similar to multidrug permeases
SGE1 (NOR1, P9677.3)	P33335	crystal violet resistance protein
YBR293W (YBR2109)	P38358	similar to multidrug permeases
YCL069W	P25594	similar to bacterial multidrug resistance proteins
YCL070-73C (YCL070C, YCL071C, YCL073C)	P25596	similar to YKR106 (frame shift: YCL070C, YCL071C, and YCL073C joined)
YD9727.14	Z48758x14	similar to multidrug permeases
YEL065W	P39980	similar to multidrug permeases
YHL040C	P38731	similar to YKR106W
YHL047C	P38724	similar to YKR106W
YKR105C	P36172	similar to SGE1

Table I
(continued)

Gene name (synonyms)[a]	Access. No[b]	Function
YKR106W	P36173	similar to YCL070-73C
YM8021.05	Z49259x15	similar to multidrug permeases
YM9582.13	Z49259x15	similar to multidrug permeases
URACIL/ALLANTOIN PERMEASES		
DAL4 (YIR028W)	Q04895	allantoin permease
FUR4 (YBRO303, YBR021W)	P05316	uracil permease
L8083.2	U19027x14	similar to FUR4 and DAL4
YBL042C (YBL0406)	P38196	similar to FUR4 and DAL4
ALLANTOATE PERMEASES		
DAL5 (UREP1, YJR152W)	P15365	allantoate permease
L0578	Z47973x16	similar to DAL5
YAL067C	P39709	similar to DAL5
YCR028C	P25621	similar to DAL5
YIL166C (YI9402.09)	P40445	similar to DAL5
PHOSPHATE PERMEASES		
N2052	P27514	similar to PHO87
PHO87 (YCR524, YCR037C)	P25360	phosphate permease
YJL198W (J0336)	P39535	similar to PHO87
PURINE/CYTOSINE PERMEASES		
FCY2 (YER056C)	P17064	cytosine / purine permease
YER060W	P40039	similar to FCY2
PROTEIN PERMEASES		
SEC61 (L3502.5)	P32915	component of ER protein-translocation complex
YBR283C (YBR2020)	P38353	similar to SEC61
PEPTIDE PERMEASES		
PTR2 (YKR413C, YKR093W)	P32901	peptide permease
POTASSIUM PERMEASES		
TRK1 (YJL129C)	P12685	potassium permease, high affinity
TRK2 (RPD2, YKR050W)	P28584	potassium permease, moderate affinity
SULFATE PERMEASES		
SUL1 (SFP, YBR2110, YBR294W)	P38359	sulfate permease, high-affinity
YP9723.03 (LPZ3C)	Z48951x3	similar to high affinity sulfate transporter
UREA PERMEASES		
DUR3 (YHL016C)	P33413	urea permease
UNKNOWN FUNCTION, FAMILY 1		
SYG1 (YIL047C)	P40964	similar to N2052
UNKNOWN FUNCTION, FAMILY 2		
PTM1 (YKL252, YKL039W)	P32857	similar to YHL017W
YHL017W	P38745	similar to PTM1
UNKNOWN FUNCTION, FAMILY 3		
YBL089W (YBL0703)	P38176	similar to YER119C
YEL064C	P39981	similar to YBL089W
YER119C	P40074	similar to YBL089W
YIL088C (I9910.08)	P40501	similar to YBL089W
UNKNOWN FUNCTION, FAMILY 4		
JEN1 (YKL217W)	P36035	similar to bacterial proline / betaine and mammalian Na^+/carboxylic acid permeases

Families are based on BLAST and PRSS, clusters within families are based on phylogenetic trees. [a] Gene names and synonyms are according to the YPD database at URL http://www.proteome.com/YPDhome.html . [b]Accession numbers are from SwissProt if started by P or Q, otherwise from GenBank

into two families: MDR 1 and MDR 2 (Table 1), which comprise 9 and 24 representatives respectively. This slightly modifies the conclusions of a recent study of Goffeau et al. (unpublished results), in which all multidrug resistance proteins are in one family, divided into 3 clusters. Taking into account the number of predicted transmembrane spans (Goffeau et al., unpublished results), which is 12 for MDR 1 and 14 for MDR 2, it seems that the assignment of the multidrug resistance proteins to two families instead of one family is correct.

4.4. *Other permease families with known function*

As can be seen in Table 1, the uracil/allantoin permease family comprises 4 representatives. The allantoin permease (DAL4) and the uracil permease (FUR4) are more closely related to each other than to YBL042W and L8083.2 (unpublished results). The allantoate permease family contains 5 representatives. The allantoate permeases DAL5 and L0578 are more related to each other than to the other members, and so are YCR028C and YAL067C (unpublished results). The phosphate permease family contains 3 representatives, N2052, PHO87, and YJL198W, but not PHO84 which is a member of the sugar permease family. Based on the BLAST results, SYG1 also belongs to the phosphate permease family, but it was excluded on the basis of the PRSS results and put in a separate family with unknown function. The purine/cytosine, protein, potassium, and sulfate permease families contain only 2 representatives each, while the peptide and urea permease families consist of only one member each.

4.5. *Permease families with unknown function*

The families listed as unknown function bare no similarity to proteins with a known function in yeast or other organisms. Four such families are listed in Table 1 with 1, 2, 4, and 1 member(s).

5. Conclusions

The present work demonstrates the power of the phylogen-

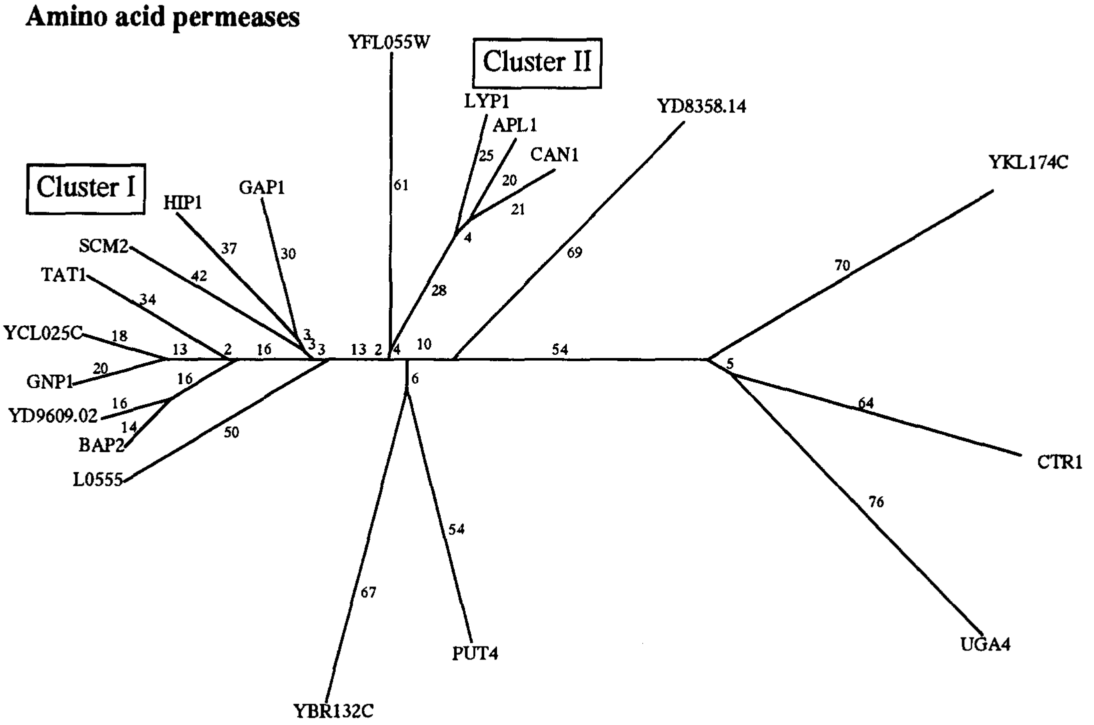

Fig. 2. Phylogenetic tree of the amino acid permeases belonging to the MFS. Conventions as in Fig. 1.

etic analysis of membrane proteins, pioneered by [2], as applied to the data generated by the systematic sequencing of the yeast genome [15]. This approach has allowed to distinguish 17 families within the yeast members of the MFS proteins. This is a considerable increase in the number of MFS families which so far was estimated to be 6 for all species combined [1]. At completion of this study (October 1995) approximately 5000 yeast ORFs were available, whereas the complete genome is estimated to comprise 6400 ORFs [3]. Taking into account that 100 MFS proteins were identified in the present study, the total number in the yeast genome can be estimated at 128. The additional members still to be revealed will most probably belong to the 17 families presently assigned. Our approach has allowed us to suggest functions by clustering, e.g. YJR158W which clusters with the hexose/glucose permeases in the sugar permease family. Interestingly, 4 families have been found with an unknown function (Unknown 1–4 in Table 1). Even within families with a known function, it has not been possible to suggest a function for all proteins by clustering, e.g. YFL040W in the sugar permease family.

While this work was in progress we became aware of a classification of yeast transport proteins by Bruno André (personal communication).

Acknowledgements: This research was supported by the Programme on Interuniversity Poles of Attraction of the Office for Scientific, Cultural and Technical Affairs of the Belgian State (contracts no. 18 and 23).

References

[1] Marger, M.D. and Saier, M.H. (1993) Trends Biochem. Sci. 18, 13–20.
[2] Saier, M.H. (1994) Microbiol. Rev. 58, 71–93.
[3] Goffeau, A., Slonimski, P., Nakai, K. and Risler, J.-L. (1993) Yeast 9, 691–702.
[4] Goffeau, A., Nakai, K., Slonimski, P. and Risler, J.-L. (1993) FEBS Lett. 325, 112–117.
[5] Williams, N. (1995) Science 268, 1560–1561.
[6] Klein, P., Kanehisha, M. and DeLisi, C. (1985) Biochim. Biophys. Acta 815, 468–476.
[7] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.
[8] Pearson, W.R. (1990) Methods Enzymol. 183, 63–68.
[9] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA 85, 2444–2448.
[10] Wisconsin Sequence Analysis Package, Version 8 (1994) Program Manual, Genetics Computer Group, 575 Science Drive, Madison, WI 53711, USA.
[11] Zuckerkandl, E. and Pauling, L. (1965) in: Evolving genes and proteins (Bruson, V. and Vogel, H.J., Eds.) pp. 97–166, Academic Press, New York.
[12] Dickerson, R.E. (1971) J. Mol. Evol. 1, 26–45.
[13] Saitou, N. and Nei, M. (1987) Mol. Biol. Evol. 4, 406–425.
[14] Van de Peer, Y. and De Wachter, R. (1994) Comput. Applic. Biosci. 10, 569–570.
[15] Goffeau, A. (1994) Nature 369, 101–102.