Volume 325, number 1,2, 112–117 FEBS 12433 June 1993

Minireview

The membrane proteins encoded by yeast chromosome III genes

A. Goffeau[a,b], Kenta Nakai[c], Piotr Slominski[d] and Jean-Loup Risler[d]

[a]*Unité de Biochimie Physiologique, Université de Louvain, Place Croix du Sud, 2–20, B-1348 Louvain-la-Neuve, Belgium,* [b]*Division of Biotechnology, DG XII - F-2, Commission of the European Communities, Rue de la Loi 200, B-1049 Bruxelles, Belgium,* [c]*National Institute for Basic Biology, 38 Nishigonaka, Myoda, Okazaki 444, Japan* and [d]*Centre de Génétique Moléculaire, Centre National de la Recherche Scientifique, 91198 Gif-sur-Yvette, France*

Received 29 March 1993

Examples are given for the analysis of the 68 putative membrane proteins encoded among the 170 predicted genes identified by the systematic sequencing of yeast chromosome III [(1992) Nature 357, 38–56].

Genome project; Chromosome III; Yeast; Gene product; Membrane protein

1. INTRODUCTION

On the 7th of May 1992, a team of 147 scientists from 13 countries and 39 laboratories published in *Nature* the entire sequence of chromosome III from the yeast *Saccharomyces cerevisiae* [1]. This is the first complete chromosome sequenced ever and the longest DNA contingent (314 kbp) submitted to databases so far (April 1993).

The sequencing of chromosome III was funded by the Biotechnology Action Programme from the European Community. Six other yeast chromosomes are presently being sequenced using the same network organisation principles within the frame work of the BRIDGE and BIOTECH programmes from the European Community. Four other chromosomes are presently being tackled in Japan, USA, Canada and UK (Table I).

Several publications have reported on operational and deontological aspects of this work [2–8]. The present paper will focus on some aspects of the analysis of membrane protein sequences predicted to be encoded by genes identified on chromosome III. Another paper covers some related topics such as targeting signals and comparison with the randomly obtained yeast protein sequences [9].

2. THE GENES DISCOVERED ON CHROMOSOME III

The non-Ty sequence of chromosome III is 309 kbp long [1]. Originally 214 putative open reading frame (ORFs) longer than 100 amino acids were identified [1]. Among them, 31 ORFs are totally included in a longer one encoded by the complementary DNA strand. In addition 18 pairs of partially overlapping ORFs have been identified. Detailed examination of these ORFs has trimmed the number of possibly expressed open reading frames to 170 [9,10]. This number is in general agreement with the estimated 156 transcripts from vegetative growing yeast cells mapped on chromosome III by Yoshikawa and Isono [11].

Correspondence address. A. Goffeau, Unité de Biochimie Physiologique, Université de Louvain, Place Croix du Sud, 2–20, B-1348 Louvain-la-Neuve, Belgium. Fax: (32) (10) 473 614.

3. THE KKD ALGORITHM FOR PREDICTION OF MEMBRANE PROTEINS

The 170 ORFs from chromosome III have been analyzed for prediction of transmembrane α helices. Among the algorithms tested that of Klein, Kanehisa and Delisi [12] was estimated to provide efficient and rather reliable results [13,16]. This method (abbreviated as KKD) computes like the Kyte and Doolittle one [14] an average hydrophobic index which characterizes spans of seventeen neighbouring amino acids (H17). When the value of the empirical Eqn. I is lower than 10^0 a given membrane span is considered to be a putative 'integral' (*I*) membrane span. When the value is between 10^0 to 10^2 the span is considered to be 'possibly integral'. Values above 10^2 correspond to 'peripheral' (*P*) domains. The result of Eqn. I is therefore called *P*:*I* odds.

$$10^{1.05(H17)^2 - 12.30(H17) + 17.49} = P{:}I \text{ odds} \quad (1)$$

Published by Elsevier Science Publishers B.V.

The validity of Eqn. I has been tested on 102 integral and peripheral proteins. However it must be pointed out that among them, the exact topography of only one protein, bacteriorhodopsin, was characerized by electron microscopical analysis (for discussion and further references, see [13] and [15]). A particular problem of the KKD method is the establishment of the threshold *P:I* odds above which an hydrophobic stretch is not membrane integrated but peripheral.

Table II reports examples of *P:I* odds encountered by using the KKD algorithm on two different open reading frames from yeast chromosome III.

YCLO69W and YCR106W contain both a high number of "integral" plus "possibly integral" membrane spans. YCL069W which contains 9 "integral" and 2 "possibly integral" membrane spans is likely to correspond to a membrane protein since it shows significant homology (FASTA score of 184) to the tetracycline transporter from *E coli.* It is not clear whether YCR106W corresponds to a membrane protein even though the total number of "integral" plus "possible integral" membrane spans is almost as high than that predicted for YCL069W. Indeed its best homology (FASTA score 143) is with CYP1, a Zn finger regulatory protein which is not likely to be membrane-bound. From the analysis of the predictions made on the photosynthetic complex of *Rhodopseudomonas viridis* [13] and of 17 H^+-ATPases [16] it was proposed that a threshold value of $1.5 \cdot 10^1$ for the *P:I* odds might be optimal [10]. If this threshold was adopted one would speculate that YCL069W and YCR106W comprise 10 and 7 membrane spans, respectively. If the threshold *P:I* odds was chosen to be $3.5 \cdot 10^1$, the two proteins would be predicted to comprise 10 membrane spans each. For a *P:I* threshold of 10^{-1}, YCL069W would be considered to comprise 7 membrane spans whereas YCR106W would have only one. These examples illustrate the difficulty of adopting a single threshold value for the *P:I* odds of ORFs for which no well-characterized homologs are available. Integration of multispanning proteins in membranes is not yet understood and parameters others than the individual hydrophobicity of each membrane stretches are likely to be involved in this process (see [17,18]).

4. EXAMPLES OF PREDICTIONS FOR MEMBRANE TOPOGRAPHIES

The KKD algorithm provides a prediction of the borders of each putative membrane α helix but does not give information on the orientation of the transmembrane span respective to the outer or inner side of the

Table I

Yeast genome systematic sequencing

Chromosomes undertaken (Funding approved in March 1993)

Expected termination date	Chromosome	Approximate length (in kb)	Approximate number of ORFs	DNA coordinators
1992	III	320	170	S. Oliver (EC)
1993	XI	630	340	B. Dujon (EC)
1993	VI	280	150	Y. Murakami (Japan)
1993	II	830	450	H. Feldmann (EC)
1994	I	220	120	H. Bussey (Canada)
1994	IX	440	240	B. Barrell (UK)
1994	V	620	330	D. Botstein (USA) R. Davis (USA)
1994	X	760	410	F. Galibert (EC)
1994	XIV	800	430	P. Philippsen (EC)
1995	XV	1100	600	B. Dujon (EC)
1996	VII	1200	650	A. Goffeau (EC)

Chromosomes under negociation (March 1993)

	IV	1600	870	(EC)
	VIII	550	300	(EC, USA, Japan)
	XVI	960	520	(Canada)

Free chromosomes (March 1993)

	XIII	920	500	
	XII	2200 (rDNA)	600	

Table II

Examples of the predictions made on membrane spanning segments from two ORFs of unknown function from yeast chromosome III

YCLO69W Protein length: 458 amino acids	*P:I* odds	Membrane part
Integral	$5.56 \cdot 10^{-6}$	160–189
Integral	$1.03 \cdot 10^{-4}$	34–62
Integral	$1.25 \cdot 10^{-3}$	131–154
Integral	$2.76 \cdot 10^{-3}$	248–271
Integral	$6.45 \cdot 10^{-3}$	272–298
Integral	$6.45 \cdot 10^{-3}$	205–227
Integral	$1.16 \cdot 10^{-2}$	68–88
Integral	$2.14 \cdot 10^{-1}$	307–323
Integral	$3.27 \cdot 10^{-1}$	9–25
Possibly integral	$1.88 \cdot 10^{0}$	419–435
Possibly integral	$5.62 \cdot 10^{1}$	230–246
YCR106W **Protein length: 832 amino acids**	***P:I* odds**	**Membrane part**
Integral	$6.32 \cdot 10^{-2}$	104–121
Integral	$1.83 \cdot 10^{-1}$	616–632
Integral	$3.64 \cdot 10^{-1}$	275–292
Possibly integral	$1.17 \cdot 10^{0}$	542–558
Possibly integral	$1.99 \cdot 10^{0}$	699–715
Possibly integral	$6.39 \cdot 10^{0}$	398–414
Possibly integral	$1.21 \cdot 10^{1}$	767–783
Possibly integral	$1.85 \cdot 10^{1}$	130–146
Possibly integral	$1.85 \cdot 10^{1}$	314–330
Possibly integral	$3.49 \cdot 10^{1}$	203–219

The algorithm of Klein et al. [12] has been used for *P:I* odds and borders of the membrane parts. The two ORFs YCL069W and YCR106W are described in [1] and [10].

membrane. However, other methods allow such predictions to be made.

Fig. 1 illustrates a variety of situations encountered with ORFs from yeast chromosome III which have in common that the KKD method predicts the presence of a single 'integral' membrane span with a *P:I* score below $3 \cdot 10^{-2}$.

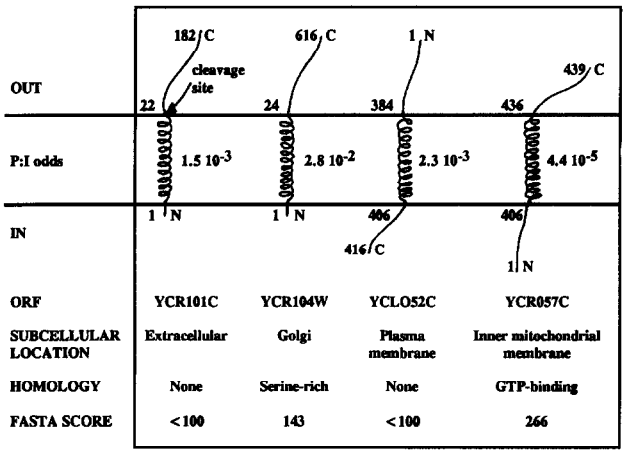

Fig. 1. Examples of predicted membrane topographies of four ORFs from yeast chromosome III containing only one putative integral membrane span. The *P:I* odds and location of membrane spanning α-helices are predicted according to Klein et al. [12]. The numbers given at the top of the figure refer to the amino acid positions of the N and C termini and of the end of the predicted integral membrane spans. The subcellular location and cleavage sites are predicted according to the PSORT programme from Nakai and Kanehisa [19].

The predicted 'integral' membrane span of YCR104W and YCR101C is in their N terminus whereas that of YCL52C and YCR57C is in their C terminus. Using the procedure proposed by Nakai and Kanehisa [19] which combines the methods from McGeoch [20] and von Heijne [21] a cleavable endoplasmic reticulum signal peptide is predicted in YCR101C. Therefore the mature YCR101C is likely not to be membrane-bound but corresponds to a soluble extracellular protein. Based on the charge differences between the amino acid residues flanking both sides of the predicted membrane span [22] it is possible to predict the orientation of the membrane spans from YCR104W, YCL52C and YCL57C. These results are illustrated in Fig. 1 which also provides the results from the PSORT programme for prediction of the intracellular location of proteins known only by their sequence [19] as well as the best homology FASTA score [10].

5. A LISTING OF THE 68 PREDICTED MEMBRANE PROTEINS FROM CHROMOSOME III

Table III lists 68 ORFs longer than 100 amino acids encoded by chromosome III which are predicted to con-

Table III

Listing of 68 predicted membrane proteins from chromosome III

Name	Best homology or identity*	Number of membrane spans	
		According to KKD	According to NK
YCL016C	None	1	0
YCL027W	*Fusion protein	1	1
YCL050C	*Tetraphosphohydrolase	1	0
YCL052C	None	1	1
YCL060C	Serine dehydratase	1	0
YCL073C	None	1	1
YCL007C	None	1	1
YCR101C	None	1	0
YCR104W	Serine rich protein	1	1
YCR012W	*Phosphoglycerate kinase	1	0
YCR001W	None	1	0
YCR043C	None	1	1
YCR057C	GTP binding	1	1
YCR005C	Citrate synthase	1	0
YCR069W	Cyclophilin	1	0
YCR085W	None	1	1
YCL018W	*Isopropyl malate dehydrogenase	2	0
YCL021W	None	2	1
YCL023C	None	2	1
YCL024W	Protein kinase	2	0
YCL048W	Sporulation protein	2	0
YCL071C	None	2	2
YCR026C	Plasma membrane glycoprotein	2	1
YCR059C	None	2	1
YCR068W	None	2	0
YCR077C	Glutamine rich protein	2	1
YCR007C	NADH deshydrogenase	2	2
YCR094W	None	2	2
YCR089W	a-Agglutinin	2	1
YCL014W	Lysine-rich protein	3	2
YCL001W	None	3	3
YCL002C	None	3	0
YCL045C	None	3	1
YCL053C	Chloride channel	3	1
YCL056C	None	3	0
YCL065W	None	3	3
YCR013C	None	3	0
YCR041W	None	3	3
YCR042C	*Temperature sensitive mutation	3	0
YCR054C	None	3	0
YCR067C	Secretory protein	3	1
YCL041C	None	4	1
YCL005W	None	4	1
YCR105W	Alcohol dehydrogenase	4	2
YCL058C	None	5	4
YCR073C	Protein kinase	5	1
YCR010C	None	6	6
YCR032W	Unknown human protein	6	1
YCR034W	None	6	4
YCR044C	None	6	2
YCR075C	Endoplasmic reticulum protein	6	2
YCR087W	Leucine rich protein	6	3
YCR106W	Zn-finger regulatory protein	7	1
YCR021C	Heat shock protein	7	6
YCL070C	Aminotriazole resistance	8	6
YCR081W	None	8	0
YCR048W	None	9	9
YCR061W	Methionine rich protein	9	7
YCR028C	Allantoate permease	9	7
YCL069W	Tetracycline resistance	10	9
YCR023C	Tetracycline resistance	10	8
YCL038C	None	11	10
YCR011C	White protein (ABC protein)	11	8
YCR098C	Citrate transporter	11	7
YCL025C	General amino acid permease	12	11
YCR037C	None	13	12
YCR093W	None	15	2
YCR017C	None	18	15

The numbers of membranes spans are predicted according to KKD [12] using *P:I* threshold values of $1.5 \cdot 10^{1}$ or according to NK [19] using *P:I* odds threshold values of 10^{-2} and $5 \cdot 10^{-1}$. *Best homologies are taken from [9].

tain at least one α helix transmembrane span. A more complete set of the predicted properties of these proteins is reported in [9] and [10].

The ORFs are classified accordingly to the number of membrane spans per protein as predicted by KKD using either a single threshold *P:I* odds of $1.5 \cdot 10^{1}$ or the number of membrane spans as predicted by the PSORT programme. The latter method uses a linear conversion form of the KKD algorithm but in a more stringent manner: it uses the *P:I* threshold value of $5 \cdot 10^{-1}$ only for proteins which have at least one membrane span with a *P:I* value below 10^{-2}. The assumption which justifies this choice is that a nascent polypeptide has to be anchored in the membrane by at least one span of high hydrophobicity whereas additional transmembrane spans require lower hydrophobicity to be integrated. Moreover PSORT does not consider the predicted cleaved signal peptides.

From Table III it appears that several proteins considered to be soluble, such as tetraphosphohydrolase, serine dehydratase, phosphoglycerate kinase, citrate synthase, isopropyl malate dehydrogenase, are predicted to contain one or even two membrane spans when a single *P:I* threshold of $1.5 \cdot 10^{1}$ is used. In contrast these sequences are predicted to be not membrane-bound by the PSORT programme. This might justify the use of the latter programme for prediction of membrane proteins with a low number of membrane spans even though the number of membrane spans in proteins comprising a large number of relatively amphiphilic membrane spans, such as the seven transmembrane seg-

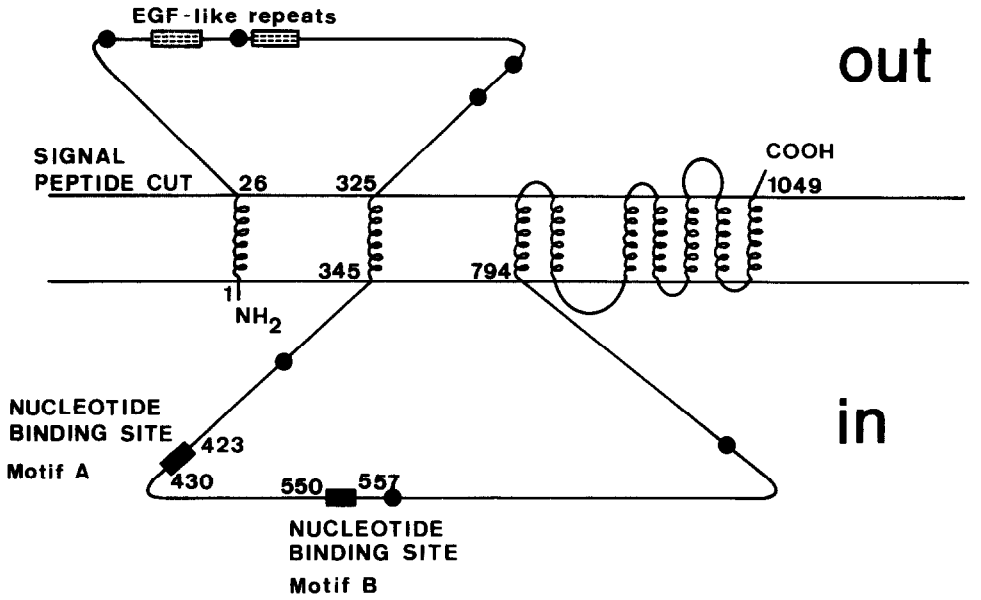

Fig. 2. The predicted membrane topography from YCR11C. This ORF is also termed ATP-dependent permease or ADP1. Black dots represent putative glycosilation sites. Taken with permission from [30].

ments in the rhodopsin family or proteins from the inner mitochondrial membrane, might be underestimated by this method [19].

A second conclusion drawn from Table III is that among the 68 membrane proteins predicted to be encoded on yeast chromosome III, a vast majority (64%) show no homology to any other protein sequence known today. These EEC (esoteric, elusive but conspicuous genes according to P. Slonimski nomenclature) seem more frequent for membrane than soluble proteins [10].

Several interesting homologies of chromosome III-encoded membrane proteins are reported in detail by Slonimski and Brouillet [10] such as those between YCL053C and a chloride channel from *Torpedo mormorate*, between YCR026C and mouse or human plasma membrane glycoprotein or between YCR098C and bacterial anion (citrate, α-ketoglutarate or phosphate) transporter. We wish to make some additional comments on the four new putative multiple drug transporters identified on chromosome III.

6. MULTIPLE DRUG TRANSPORTERS

Mutations conferring resistance to a variety of structurally and functionally unrelated drugs such as cycloheximide, oligomycin and about 20 other inhibitors are frequent in yeast where they define at least 6 different pleiotropic drug resistance (PDR) loci [23]. The best described gene responsible for this phenotype is *PDR1*, a transcriptional regulator [24] which was proposed to regulate the expression of a network of membrane-bound transporters similar in structure and function to the mammalian MDR (or P-glycoprotein) gene products [23]. These membrane proteins belong to the superfamily of ABC proteins which are characterized by a 'two-times six' α helix structure separating two hydrophilic domains each containing a conserved ATP-binding cassette [25]. *STE6*, the first gene found to encode an ABC protein in yeast, does not seem to correspond to a pleiotropic drug resistance loci but was shown to be responsible for the efflux of the *a* mating type hormone [26,27]. Very recently PDR5, a multiple drug resistance loci [28], was found to encode a typical ABC protein [29]. The expression levels of both STE6 and PDR5 are both controlled by the transcription factor encoded by PDR1 [29].

Yeast chromosome III contains another member of this family, YCR011C, which however contains only one ATP binding cassette and 'one-times six' helix structure [30]. The C-terminal half of this ORF, called ATP-dependent permease (ADP1), has strong homology to the *Drosophila* white protein believed to be involved in the transport of eye pigment precursors. The N-terminal part of the proteins contains an additional membrane span and a putative signal peptide with a cleavable site giving the predicted topography illustrated in Fig. 2. Furthermore a repeated epidermal growth factor (EGF) motif made of cysteine-rich repeats is detected in the N-terminal part. This suggest some laminin B-like or adhesion function of the external part of the protein. This arrangement is unique so far among the more than 50 ABC type of proteins reported from bacteria to man [25]. The function of the yeast YCR011C is still unknown today even though strains with a disrupted gene are under analysis.

Another family of proteins catalyzing the efflux of various drugs such as quinolone, tetracycline, methylenemycin A, antiseptics and multiple drugs from bacterial cells has been identified. These drug-resistance proteins make a family within the more than 50 members of the major facilitator superfamily (MFS) characterized by a common structural motif of 12 transmembrane α-helices [31]. Like the ABC proteins the MFS proteins consist of 'two-times six' α-helices but they do not contain ATP binding cassettes. Instead the central cytoplasmic domain of MFS proteins is dispensable. Designation as a member of the drug resistance protein family is obtained not only by the phenotype indicating a drug-efflux function but also by binary comparisons of the amino acid sequences indicating high homology scores.

So far only one eukaryote sequence, the yeast aminotriazole resistance ATR1 ORF had been reported to belong to this drug resistance family which comprises 14 bacterial sequences [31]. Yeast chromosome III reveals three new members of this drug-resistance family: YCL069W and YCR023C which show homology to the bacterial tetracycline resistance sequence, and YCL070C which shows homology to the yeast ATR1. Detailed alignments of these bacterial and yeast sequence are given in Slonimski and Brouillet [10].

The KKD predictions for membrane spans of YCL069W are given in Table II. Nine 'integral' segments and two 'possibly integral segments' are predicted. The predictions for YCR023C are very similar to those for YCL069W whereas for YCL070C only 6 'integral' and 3 'possibly integral' are predicted. Thus the KKD algorithm does not support a 12 membrane spanner structure even though the homologies suggest membership to the 'multiple facilitator' superfamily.

7. EXTRAPOLATION TO THE ENTIRE YEAST GENOME

The yeast clean (non-Ty) chromosome III contains one ORF longer than 100 amino acids per 1.8 kb DNA sequences. Extrapolated to the 12,390 kb estimated length of the complete clean (non-Ty, non-rDNA) yeast genome, this value would predict the existence of about 6,800 ORFs in yeast. It is no longer unreasonable to assume that the vast majority of yeast genes might be duplicated. If so the total number of 'original' genes in yeast might be of the order of no more than 4,000, of which 35–40% might correspond to membrane proteins [9].

Within one year of publication of the present paper, the sequences of chromosome II (220 kb, Canada), chromosome VI (280 kb, Japan), chromosome XI (630 kb, EC) and chromosome II (830 kb, EC) will become available. This corresponds to more than 1,000 additional gene sequences and will allow a reliable assessment of the exact frequency of duplication of the yeast genes as well as statistics on the percentage of membrane proteins of different classes in yeast cells (e.g. classified by the number of membrane spans per protein). The systematic deletion and overexpression of these new membrane proteins encoding genes is underway (Slonimski and Vassarotti, personal communication) and will provide unique tools for function searches. Systematic sequencing of the yeast genome is thus rapidly opening new fields for experimental and theoretical research on membrane proteins.

Acknowledgements: Support from the EC BRIDGE programme, as well as from the Fonds National de la Recherche Scientifique, the Services de la Politique Scientifique and the Région Wallone are gratefully acknowledged. Information and help of various nature have been provided by A. Vassarotti (EC), E. Balzi (UCL, Belgium), S. Oliver (UMIST, UK), W. Mewes and J. Sgouros (MIPS, Germany).

REFERENCES

[1] Oliver, S.G. and 146 al. (1992) Nature 357, 38–46.
[2] Vassarotti, A., Dujon, B., Feldmann, H., Mewes, W. and Goffeau, A. (1993) J. Biotechnol., in press.
[3] Dujon, B. (1990) Biofutur 94, 52–55.
[4] Grivell, L.A. and Planta, R.J. (1990) Trends Biotechnol. 8, 241–243.
[5] Vassarotti, A., Goffeau, A., Magnien, E., Loder, B. and Fasella, P. (1990) Biofutur 94, 84–90.
[6] Goffeau, A. and Vassarotti, A. (1991) Res. Microbiol. 142, 901–903.
[7] Dujon, B. (1992) Current Biology 2, 279–281.
[8] Vassarotti, A. and Goffeau, A. (1992) Trends Biotechnol. 10, 15–18.
[9] Goffeau, A., Slonimski, P., Nakai, K. and Risler, J.L. (1993) Yeast (in press).
[10] Slonimski, P. and Brouillet, S. (1993) Yeast, in press.
[11] Yoshikawa, A. and Isono, K. (1990) Yeast 6, 383–401.
[12] Klein, P., Kanehisa, M. and Delisi, C. (1985) Biochim. Biophys. Acta 815, 468–476.
[13] Fasman, G.D. and Gilbert, W.A. (1990) Trends Biol. Sci. 15, 89–92.
[14] Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105–132.
[15] Jähnig, F. (1990) Trends Biol. Sci. 15, 93–95.
[16] Wach, A., Schlesser, A. and Goffeau, A. (1992) J. Bioenerg. Biomembr. 24, 309–317.
[17] Sanders, S.L. and Schekman, R. (1992) J. Biol. Chem. 267, 13791–13794.
[18] Rapoport, T. (1992) Science 258, 931–936.
[19] Nakai, K. and Kanehisa, M. (1992) Genomics 14, 897–911.
[20] McGeoch, D.J. (1985) Virus Res. 3, 271–286.
[21] von Heijne, G. (1986) Nucleic Acids Res. 14, 4683–4690.
[22] Hartmann, E., Rapoport, T.A. and Lodish, H.F. (1989) Proc. Natl. Acad. Sci. USA 86, 5786–5790.
[23] Balzi, E. and Goffeau, A. (1991) Biochim. Biophys. Acta 1073, 241–252.
[24] Balzi, E. and Goffeau, A. (1987) J. Biol. Chem. 262, 16871–16879.
[25] Higgins, C.F. (1992) Annu. Rev. Cell. Biol. 8, 67–113.
[26] Kuchler, K., Steine, R.E. and Thorner, J. (1989) EMBO J. 8, 3973–3984.
[27] McGrath, J.P. and Varshavsky, A. (1989) Nature 340, 400–404.
[28] Meyers, S., Schauer, W., Balzi, E., Wagner, M., Goffeau, A. and Golin, J. (1992) Curr. Genet. 21, 431–436.
[29] Balzi, E., Wang, M., Leterme, S., Van Dyck, L., and Goffeau, A. (in preparation).
[30] Purnelle, B., Skala, J. and Goffeau, A. (1991) Yeast 7, 876–872.
[31] Marger, M.D. and Saier, M.H. (1993) Trends Biol. Sc. 18, 13–20.