THE JOURNAL OF BIOLOGICAL CHEMISTRY Vol. 278, No. 12, Issue of March 21, pp. 10208–10213, 2003
© 2003 by The American Society for Biochemistry and Molecular Biology, Inc. *Printed in U.S.A.*

Topology Models for 37 *Saccharomyces cerevisiae* Membrane Proteins Based on C-terminal Reporter Fusions and Predictions*

Received for publication, January 7, 2003
Published, JBC Papers in Press, January 10, 2003, DOI 10.1074/jbc.M300163200

Hyun Kim, Karin Melén, and Gunnar von Heijne‡

From the Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden

We provide experimentally based topology models for 37 integral membrane proteins from *Saccharomyces cerevisiae*. A C-terminal fusion to a dual Suc2/His4C topology reporter has been used to determine the location of the C terminus of each protein relative to the endoplasmic reticulum membrane, and this information is used in conjunction with theoretical topology prediction methods to arrive at a final topology model. We propose that this approach may be used to produce reliable topology models on a proteome-wide scale.

The topology of a membrane protein, *i.e.* a specification of its transmembrane segments and their in/out orientation relative to the membrane, is a basic structural characteristic that is a very powerful guide to experimental studies when no three-dimensional structure is available. Most proteins for which experimentally derived topology models exist are from bacteria; only a handful of topologies are known for yeast membrane proteins. As an example, the widely used MPtopo data base (1) currently holds a total of 92 experimentally determined topology models, of which only 3 are for yeast proteins.

The preponderance of known topologies for bacterial proteins is mainly a result of the relative ease with which they can be determined experimentally by making series of C-terminal-truncated versions of the target protein fused to topology reporters such as PhoA, LacZ, Bla, or green fluorescent protein (2, 3). This approach often yields clear-cut results, and reliable topology models can be proposed. Although a couple of topology reporters have also been developed for the yeast *Saccharomyces cerevisiae* (4), they seem to yield less definitive results than can be obtained in bacterial systems (5) and have not been much used.

We have recently shown that highly reliable topology predictions can be obtained for a subset of membrane proteins for which many different topology prediction methods (five in our case) give the same prediction (6, 7). When some limited experimental information (such as the in/out location of the C terminus of a protein) is available, consensus predictions are even more useful, and a combination of C-terminal reporter fusions and consensus predictions has been shown to make it possible to rapidly produce reliable topology models for *Escherichia coli* inner membrane proteins (3). We have further shown that this basic approach can be even more widely applied by using the experimentally determined location of the C terminus as a constraint on the topology predictions produced by the TMHMM program (8).

Considering that topology mapping based on reporter fusions to truncated target proteins may be problematic in yeast, it appeared to us that approaches based on a combination of C-terminal fusions to the full-length protein, which should be minimally disruptive to the structure of the proteins, and theoretical topology prediction should be particularly useful for yeast membrane proteins. In this pilot study, we report results for C-terminal reporter fusions to 40 *S. cerevisiae* membrane proteins. We find that only one of the 40 proteins that we cloned initially cannot be expressed using either of two vectors that we have tried, that the dual Suc2/His4C reporter (9) that we have used yields consistent results for 37 of the 39 expressed proteins, and that the location of the C terminus as predicted by our consensus method (6) is correct for 31 proteins of these 37. One of the two proteins for which the experimental results are inconsistent is a mitochondrial protein. We have also used the experimentally determined C-terminal locations to constrain the predictions from the TMHMM method, and we report TMHMM reliability scores for all proposed topology models. Our results suggest that large-scale topology mapping strategies where limited but reliable experimental information is combined with topology prediction will be successful in yeast.

MATERIALS AND METHODS

Construction of Plasmids—All plasmids were constructed by homologous recombination (10). Plasmid pJK90,[1] which contains the *OST4* gene fused to three hemagglutinin (HA)[2] epitopes, a part of the *SUC2* gene, and the *HIS4* gene, was treated with *Sma*I to linearize the vector between the end of TPI promoter and the start codon of *OST4*. The 5′-end homologous recombination region was selected to match the 3′-end of *Sma*I-digested pJK90, and the 3′-end homologous region was chosen to match the linker between the end of *OST4* and the start of the HA sequence. Each homologous region comprised 35 nucleotides, (5′-AGGTGGTTTGTTACGCATGCAAGCTTGATATCGAA-3′ and 5′-GATGGTCTAGAGGTGTAACCACTTGAGTTCTTAGG-3′). A gene of interest was amplified by PCR using genomic DNA as a template and two primers, a 5′-end primer complementing the start codon of the gene with the homologous region sequence and a 3′-end primer complementing the end of the gene excluding the stop codon with the homologous region sequence. Genomic DNA was isolated as described (11) from W303–1a (*MATa, ade2, can1, his3, leu2, trp1, ura3*) and from W303–1α (*MATα, ade2, can1, his3, leu2, trp1, ura3*). A yeast strain STY 50 (*MATa, his4–401, leu2, -3, and -112, trp1–1, ura3–52, HOL1–1, SUC2::LEU2*) (12) was transformed with the linearized pJK90 vector and the PCR product carrying the gene of interest flanked by the homologous region sequences. Transformation was carried out by the lithium acetate protocol (13). Transformants were selected on synthetic medium lacking uracil (SD−Ura). Plasmids were isolated and verified by PCR analysis and DNA sequencing. Plasmids were named as pJK92 (gene name) using gene names from the *Saccharomyces* genome data base (14).

* This work was supported by grants from the Swedish Research Council and the Swedish Cancer Foundation (to G. v. H.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed. Tel.: 46-8-16-25-90; Fax: 46-8-15-36-79; E-mail: gunnar@dbb.su.se.

[1] Kim *et al.*, submitted for publication.

[2] The abbreviations used are: HA, hemagglutinin; Endo H, endo-β-*N*-acetylglucosaminidase; SD−Ura, synthetic medium lacking uracil.

For construction of plasmids with an inducible Gal promoter, the fragment carrying the gene, HA epitope, *SUC2*, and *HIS4C* was amplified by PCR using pJK92(YEL059C) or pJK92(YJL028W) as a template. The two primers used in this PCR carried the homologous region sequences with the *Eco*RI-digested 424GALS (ATCC, Manassas, VA). The PCR product and the linearized 424GALS plasmid were transformed into strain STY50, and transformants were selected on −Trp plates. The correct construction of the plasmid was confirmed by yeast colony PCR.

Preparation of Whole-cell Lysates—Yeast transformants carrying TPI promoter plasmids were grown to OD_{600} 0.8 to 1 in 10 ml of SD−Ura. Harvested cell pellets were washed with 5 ml of dH_2O and left at −20 °C for at least 1 h. Frozen cells were resuspended in 200 μl of SDS sample buffer (50 mM Tris-HCl, pH 6.8, 10% glycerol, 2% SDS, 5% β -mercaptoethanol, 0.5 mM EDTA, 1 mM phenylmethylsulfonyl fluoride, protease inhibitor mixture (Roche Molecular Biochemicals), 0.0025% bromphenol blue), incubated at 60 °C for 10 to 15 min and centrifuged for 10 min at 13,000 rpm in an Eppendorf microfuge. Soluble fractions were transferred to new tubes and subjected to Endo H digestion. Transformants carrying the GALS promoter were grown to OD_{600} 1 to 2 in 5 ml of −Trp media. Cells were harvested by centrifugation and diluted to 4-fold with −Trp media supplemented with galactose instead of glucose as carbon source and grown for 5 h at 30 °C. Cell lysates were prepared as described above.

Deglycosylation by Endo H Digestion—Whole-cell lysates were supplemented with a final concentration of 80 mM potassium acetate, pH 5.6, and 2 μl of Endo H (1 unit/200 μl, Roche Molecular Biochemicals) was added. Samples were incubated at 37 °C for 1 to 2 h. Mock samples were treated and incubated in the same way but without Endo H.

Western Blot Analysis—Solublized proteins were separated on 7.5% SDS-polyacrylamide gels, transferred onto nitrocellulose membranes, and probed with anti-HA antibody (Babco, Richmond, CA).

Growth Assay—Transfomants carrying each fusion construct were streaked on SD−Ura medium lacking histidine but containing 6 mM histidinol. Plates were incubated at 30 °C for 3 to 4 days.

Computational Methods—All predicted *S. cerevisiae* open reading frames (15) were downloaded from genome-ftp.stanford.edu (version June 29, 2001). TMHMM1.0 (16) was used to identify putative membrane proteins with a minimum of two predicted transmembrane helices. From this set, 55 proteins were selected for which five different topology prediction methods, TMHMM1.0, HMMTOP2.0 (17, 18), MEMSAT1.8 (19), PHD2.1 (20), and TOPPRED1.0 (21, 22), all gave the same predicted topology. Three genes carrying introns (YDR376W, YML052W, YMR292W) were removed from this set, as were seven genes annotated as questionable open reading frames (YCL023C, YDR526C, YFL032W, YGL024W, YGL204C, YGR228W, YNL266W). A gene encoding a known mitochondrial protein (Q0275) was also excluded. The remaining 44 genes were cloned into the expression vectors described above.

The 37 proteins for which the location of the C terminus could be determined experimentally (Table I) were further analyzed using a new version of TMHMM (8) that calculates a reliability score for the predicted topology and also allows any part of the topology to be fixed to a given location *a priori*. The experimentally determined C-terminal locations were used as constraints in these predictions.

RESULTS

Selection of Target Proteins—To select *S. cerevisiae* membrane proteins for this study, we first searched all predicted open reading frames in the yeast genome (15) for membrane proteins for which five prediction methods (TOPPRED, TMHMM, HMMTOP, MEMSAT, and PHD) all give the same predicted topology. From our previous work (6, 7), we anticipated that the predicted topologies should be correct for a high proportion of these proteins. We further required that the TMHMM method predict at least two transmembrane helices in each protein because currently available bioinformatics tools cannot reliably distinguish between N-terminal signal-anchor sequences and cleavable signal peptides and thus may mistakenly identify secreted proteins as single-spanning, N-terminal anchored membrane proteins. This initial screen produced a list of 55 proteins.

Three genes carrying introns (YDR376W, YML052W, YMR292W) were excluded from the original list, as were seven genes annotated as questionable open reading frames (YCL023C, YDR526C, YFL032W, YGL024W, YGL204C, YGR228W, YNL266W). A gene encoding a known mitochondrial protein (Q0275) was excluded because the glycosylation assay cannot be used for mitochondrial proteins. Five additional proteins could not be analyzed. YNL323W was not amplified by PCR, the cloned sequence of YOL137W turned out to be different from the expected sequence, the cloned sequences of YDL196W and YNL101W contained frameshifts relative to the published sequences, and protein expression of YJL028W was not detected. We successfully made and expressed C-terminal reporter fusions to the remaining 39 proteins (Table I).

Experimental Determination of C-terminal Locations—For this study, we chose a 125-kDa dual Suc2/His4C topology reporter (4) to determine the location of a protein's C terminus in either the cytosol or the endoplasmic reticulum lumen (9, 12). The histidinol dehydrogenase activity of the His4C moiety converts histidinol to histidine only when it is localized in cytosol. Thus, only cells expressing fusion proteins with the reporter domain in the cytosol can grow on histidine-free media supplemented with histidinol. The part of the *SUC2* gene that is present in the reporter encodes a segment of invertase containing eight *N*-glycosylation acceptor sites. When this domain is localized in the lumen of the endoplasmic reticulum, the fusion protein becomes heavily glycosylated.

The cytosolic/non-cytosolic location of the C terminus of each of the 39 Suc2/His4C fusion proteins was determined by Endo H treatment (to identify a glycosylated, lumenally oriented reporter) Fig. 1 and growth on histidine-free media containing histidinol (to identify a cytosolically oriented reporter), Fig. 2. We did not observe any general growth defects of the yeast transformants carrying these fusion constructs, indicating that the addition of the reporter domain to the target proteins had no obvious harmful effects.

For 35 of the 39 fusion proteins, the results from the glycosylation and histidinol growth assays were entirely consistent (Table I). Some of these proteins are known to be localized to the membranes of secretory organelles and the plasma membrane, but most have no known localization or function annotated in the *Saccharomyces* genome data base (14) or in MIPS (23). Because these 35 fusion proteins were either glycosylated or had histidinol dehydrogenase activity, it is reasonable to assume that their natural locations are in the membranes along the secretory pathway, although we cannot completely rule out that some of the unglycosylated proteins are located in mitochondria with their C terminus facing either the cytosol or the intermembrane space.

The initial results from the glycosylation and histidinol growth assays were inconsistent for four proteins. Growth on histidinol was seen for YGR290W, despite the fact it was efficiently glycosylated. A small amount of unglycosylated protein possibly representing molecules where the reporter is cytosolically oriented was evident, however, and given the high level of expression seen for this protein this may be enough to allow growth on histidinol. We thus conclude that YGR290W has its C terminus in the endoplasmic reticulum lumen. We further found that YEL059W was not expressed from the constitutive TPI promoter. However, a low level of expression was seen when the inducible Gal promoter was used, Fig. 2. The fusion protein was sensitive to Endo H digestion, indicating that the C terminus of the protein was glycosylated and located in the lumen of the endoplasmic reticulum. Because YEL059W was only expressed from the inducible Gal promoter, growth on histidinol could not be assayed.

Finally, in the case of YKR065C and YER185W, the fusion proteins were expressed, but neither became glycosylated nor allowed growth on histidinol. We considered that a possible

Table I

Summary of the results for the 39 proteins analyzed in this study

Gene names in column 1 are from the *Saccharomyces* genome data base (14). Column 2 gives the length of the wild-type protein, column 3 a summary of the annotation found in the the *Saccharomyces* genome data base and in the MIPS database (23), column 4 gives a qualtitative measure of expression levels as judged visually from Western blots, column 5 shows whether or not the fusion protein is glycosylated, column 6 whether a *his4* strain transformed with a plasmid carrying the fusion protein can or cannot grow on histidinol, column 7 gives the location of the C-terminal end of the wild-type protein as deduced from the glycosylation and growth phenotypes, column 8 gives the consensus topology predicted by the five methods used here, column 9 gives the TMHMM S3-score (8) calculated for the topology in column 8, column 10 gives the expected accuracy (8) calculated from the score in column, 9, column 11 gives the topology predicted by TMHMM after inclusion of the experimentally determined location of the C-terminus (column 7), and columns 12 and 13 give the TMHMM S3-score and the expected accuracy for the topology prediction in column 11. The detailed TMHMM outputs corresponding to the predictions given in column 11 are provided as an Electronic Supplement.

Open reading frame	Length	Predicted or known function	Expression	Glyco	Growth	C-terminal	Consensus	Score	Accuracy	TMHMM (C)	Score (C)	Accuracy (C)
YALOO7C	215	p24 protein involved in membrane trafficking	Medium	No	Yes	In	2, N_{in}-C_{in}	0.44	0.55	2, N_{in}-C_{in}	0.58	0.71
YBR210W	142	Strong similarity to *Drosophila melanogaster* cornichon protein	High	Yes	No	Out	3, N_{in}-C_{out}	0.98	0.99	3, N_{in}-C_{out}	0.98	0.99
YDL212W	210	Chaperone	Medium	No	Yes	In	4, N_{in}-C_{in}	0.98	0.98	4, N_{in}-C_{in}	0.99	0.99
YDRO90C	310	Weak similarity to YRO2	Low	Yes	No	Out	7, N_{out}-C_{in}	0.34	0.47	7, N_{in}-C_{out}	0.65	0.76
YDR182W	491	Cell division control protein	Low	Yes	No	Out	3, N_{in}-C_{out}	1.00	1.00	3, N_{in}-C_{out}	1,00	1.00
YDR438W	370	Strong similarity to hypothetical protein YML018c	Medium	No	Yes	In	10, N_{in}-C_{in}	0.42	0.53	10, N_{in}-C_{in}	0.42	0.59
YDR525W-A	79	Similarity to PMP3/SNA1	Medium	No	Yes	In	2, N_{out}-C_{out}	0.42	0.54	2, N_{in}-C_{in}	0.76	0.83
YELO59W	102	Hypothetical protein	Medium	Yes	N/D	Out	2, N_{out}-C_{out}	0.84	0.87	2, N_{out}-C_{out}	0.88	0.91
YERO56C	533	Purine-cytosine permease	Medium	No	Yes	In	12, N_{in}-C_{in}	0.05	0.24	12, N_{in}-C_{in}	0.08	0.35
YER119C	448	Weak similarity to *Erwinia herbicola* tyrosine permease	Medium	Yes	No	Out	11, N_{in}-C_{out}	0.47	0.58	11, N_{in}-C_{out}	0.48	0.64
YER185W	303	Strong similarity to Rtm 1p	Medium	No	No	?	7, N_{out}-C_{in}	0.86	0.89			
YGR055W	574	High affinity methionine permease	Medium	No	Yes	In	12, N_{in}-C_{in}	0.20	0.36	12, N_{in}-C_{in}	0.20	0.44
YGR105W	77	ATPase assembly integral membrane protein	High	No	Yes	In	2, N_{in}-C_{in}	0.93	0.94	2, N_{in}-C_{in}	0.97	0.98
YGR121C	492	Ammonia permease	Low	No	Yes	In	11, N_{out}-C_{in}	0.37	0.50	11, N_{out}-C_{in}	0.45	0.62
YGR149W	432	Similarity to hypothetical protein SPBC776.05 *Schizosacchgromyces pombe*	Medium	No	Yes	In	8, N_{in}-C_{in}	0.34	0.47	8, N_{in}-C_{in}	0.55	0.69
YGR213C	317	Involved in 7-aminocholesterol resistance	Medium	No	Yes	In	7, N_{out}-C_{in}	0.98	0.99	7, N_{out}-C_{in}	0.99	0.99
YGR290W	147	Hypothetical protein	High	Yes	Yes	Out	2, N_{in}-C_{in}	0.91	0.93	2, N_{out}-C_{out}	0.63	0.74
YHRO26W	213	H+-ATPase 23 kDa subunit	Medium	Yes	No	Out	5, N_{in}-C_{out}	0.92	0.93	5, N_{in}-C_{out}	0.99	0.99
YHR140W	239	Hypothetical protein	Medium	No	Yes	In	6, N_{in}-C_{in}	0.91	0.93	6, N_{in}-C_{in}	0.91	0.94
YJL170C	183	Weak similarity to *Helicobacter pylori* endonuclease III	Medium	Yes	No	In	2, N_{in}-C_{in}	0.79	0.83	1, N_{in}-C_{out}	0.30	0.51
YKL119C	215	H+-ATPase assembly protein	Low	No	Yes	In	2, N_{in}-C_{in}	0.98	0.98	2, N_{in}-C_{in}	0.99	1.00
YKR044W	443	Hypothetical protein	Medium	No	Yes	In	2, N_{in}-C_{in}	0.90	0.92	2, N_{in}-C_{in}	0.91	0.94
YKR065C	197	Similarity to hypothetical protein *Schizosacchgromyces pombe*	Medium	No	No	?	2, N_{in}-C_{in}	0.99	0.99			
YLL028W	586	Polyamine transport protein	Medium	No	Yes	In	12, N_{in}-C_{in}	0.20	0.36	12, N_{in}-C_{in}	0.20	0.44

TABLE I—*continued*

Open reading frame	Length	Predicted or known function	Expression	Glyco	Growth	C-terminal	Consensus	Score	Accuracy	TMHMM (C)	Score (C)	Accuracy (C)
YLRO46C	270	Strong similarity to Rta1p and Rtm1p protein	Medium	No	Yes	In	6, N_{in}-C_{in}	0.51	0.60	6, N_{in}-C_{in}	0.59	0.71
YLR311C	115	Weak similarity to *Sauroleishmania tarentolae* cryptogene protein G4	Medium	Yes	No	Out	2, N_{in}-C_{in}	0.82	0.85	3, N_{in}-C_{out}	0.71	0.79
YLR404W	285	Hypothetical protein	Medium	No	Yes	In	2, N_{in}-C_{in}	0.81	0.85	2, N_{in}-C_{in}	0.99	0.99
YLR443W	448	Involved in cell wall biogenesis and architecture	Medium	No	Yes	In	4, N_{in}-C_{in}	0.52	0.62	4, N_{in}-C_{in}	0.75	0.82
YMRO40W	160	Strong similarity to Yet 1p	Medium	No	Yes	In	3, N_{out}-C_{in}	0.96	0.97	3, N_{out}-C_{in}	0.97	0.98
YMR148W	148	Hypothetical protein	Medium	No	Yes	In	2, N_{in}-C_{in}	0.54	0.63	2, N_{in}-C_{in}	0.63	0.74
YNL194C	301	Strong similarity to YDL222c and similarity to Sur7p	Medium	Yes	No	Out	4, N_{in}-C_{in}	0.90	0.92	4, N_{out}-C_{out}	0.58	0.70
YNR002C	282	Strong similarity to *Yarrowia lipolytica* GPR1	Medium	Yes	No	Out	6, N_{out}-C_{out}	0.23	0.38	6, N_{out}-C_{out}	0.32	0.53
YNR062C	327	Weak similarity to *Hemophilus influenzae* lctP homolog	Medium	No	Yes	In	8, N_{in}-C_{in}	0.88	0.90	8, N_{in}-C_{in}	0.88	0.92
YOLO79W	132	Similarity to NADH dehydrogenases	High	Yes	No	Out	3, N_{in}-C_{out}	0.34	0.48	3, N_{in}-C_{out}	0.37	0.56
YOL1O1C	312	Similarity to YOLOO2c and YDR492w	Low	Yes	No	Out	7, N_{in}-C_{out}	0.82	0.86	7, N_{in}-C_{out}	0.83	0.88
YOR376W	122	Hypothetical protein	Medium	No	Yes	In	2, N_{in}-C_{in}	0.72	0.78	2, N_{in}-C_{in}	0.89	0.92
YPL264C	353	Strong similarity to YMR253c	Medium	No	Yes	In	10, N_{in}-C_{in}	0.48	0.59	10, N_{in}-C_{in}	0.49	0.64
YPRO71W	211	Strong similarity to YIL029c	Medium	No	Yes	In	4, N_{in}-C_{in}	0.78	0.83	4, N_{in}-C_{in}	0.92	0.94
YPR192W	305	Similarity to water channel proteins	Medium	No	Yes	In	6, N_{in}-C_{in}	0.93	0.95	6, N_{in}-C_{in}	0.94	0.96

explanation for this observation might be that the proteins are localized to mitochondria with their C termini in the matrix space. YKR065C is strongly predicted to have an N-terminal mitochondrial targeting peptide both by TargetP (24) and a predictor specifically developed for yeast proteins (25). Indeed, YKR065C has recently been identified as a mitochondrial inner membrane protein with a cleavable, matrix-targeting presequence.[3] The location of YER185W is so far unknown.

We also roughly estimated the relative expression levels based on Western blotting with HA antibodies (Table I). It appears that small fusion proteins with few transmembrane helices tend to be better expressed than large proteins, but the correlation is not very strong.

Topology Models—As shown in Table I, the consensus predictions for the location of the C termini of the 37 proteins for which this could be deduced from the fusion protein data matched the experimental results in 31 cases (84%). Thus, for six proteins the consensus prediction does not yield a good topology model.

As a more direct way of integrating the experimental results into the final predictions, we took advantage of a recent improvement to the TMHMM program that allows predictions to be constrained by experimental information (8). This new version of TMHMM also calculates a reliability score for each prediction that correlates strongly with prediction accuracy. The TMHMM results with inclusion of the experimentally determined C-terminal location are shown in Table I, together with the corresponding reliability score and the estimated probability that the prediction is correct (the “expected accuracy”).

DISCUSSION

In contrast to the situation for bacterial inner membrane proteins, experimentally derived topology models are available for only a handful of yeast membrane proteins. This lack of data is aggravated by the fact that theoretical topology prediction methods seem to perform less well on yeast membrane proteins than on both bacterial and mammalian ones (8).[4]

In this study, we have applied and extended a strategy initially proposed for bacterial inner membrane proteins (3) to a set of 39 predicted membrane proteins from the yeast *S. cerevisiae*. The strategy is based on the premise that reliable topology models can be produced rapidly by combining limited experimental information with topology predictions. The experimental information generated is the cytosolic/non-cytosolic location of the C terminus of the target proteins, and we show that this can be easily and reliably obtained by fusion of the full-length protein to a C-terminal, dual topology reporter (9) composed of a hemagglutinin tag for immunodetection, a part of Suc2p that contains eight acceptor sites for *N*-linked glycosylation, and the His4p enzyme that converts histidinol to histidine. Because *N*-linked glycosylation can only be carried out in the endoplasmic reticulum lumen and histidinol cannot be transported into this compartment, the cytoplasmic/non-cytoplasmic location of the reporter (and thus of the C terminus of the target protein) is easily assayed by checking whether Endo H can digest any *N*-linked glycans and whether *his4* cells expressing the target protein-reporter fusion can grow on histidinol-containing media lacking histidine.

38 of the 39 fusion proteins could be expressed from the constitutive TPI promoter in amounts sufficient for analysis; one protein could be expressed only from the inducible Gal4 promoter (which precludes use of the histidinol growth assay). Only one protein that was included in our initial set could not

[3] N. Pfanner, personal communication.

[4] K. Melén, A. Krogh, and G. von Heijne (2003) *J. Mol. Biol.*, in press.

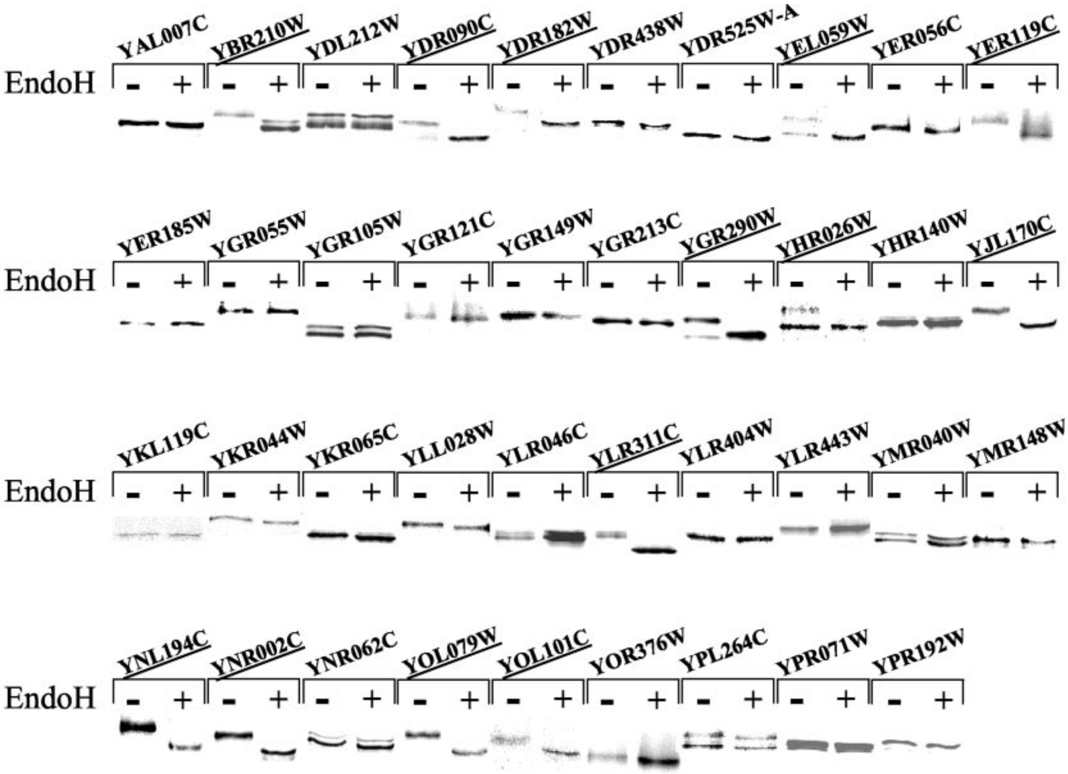

FIG. 1. **Endo H treatment of the 39 fusion proteins analyzed in this study.** Proteins were expressed as detailed under "Materials and Methods," and the samples were either treated (+) or not treated (−) with Endo H to remove *N*-linked glycans. After separation by SDS-PAGE, gels were blotted with an anti-HA antibody. Proteins for which Endo H treatment results in a reduction of the Mw are indicated by *underlining*.

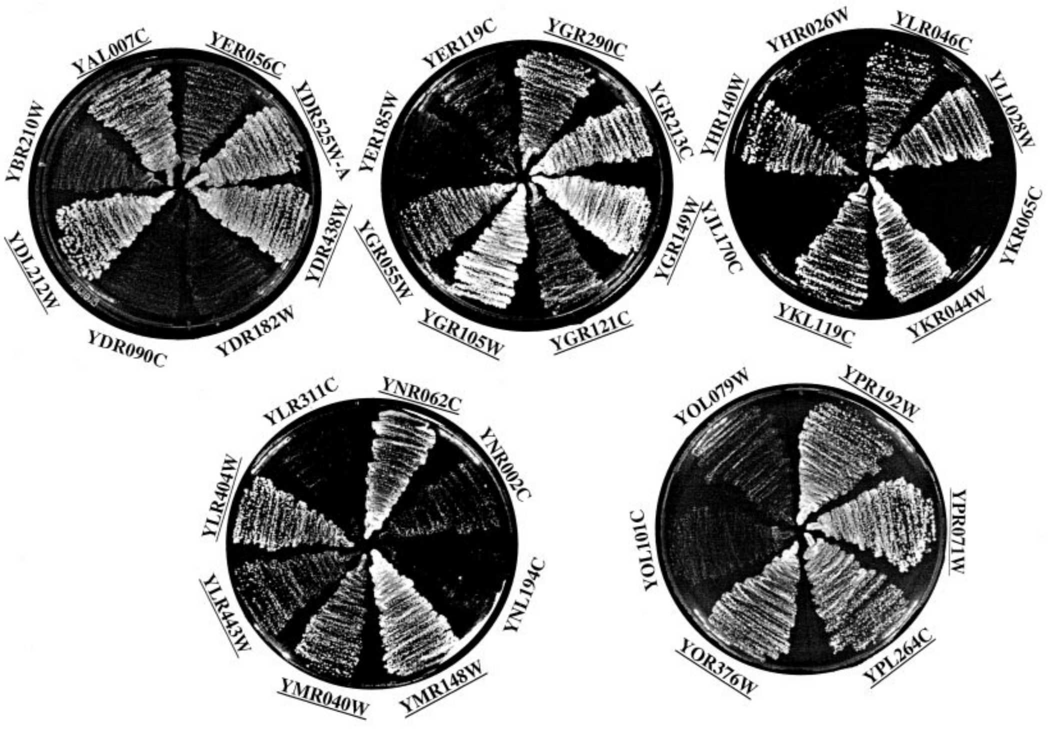

FIG. 2. **Growth of *his4* cells expressing the 39 fusion proteins analyzed in this study on a medium lacking histidine but including histidinol.** Strains that grow well are indicated by *underlining*.

be expressed from either promoter. It thus appears that nearly all membrane proteins in *S. cerevisiae* can be analyzed by our procedure.

The results from the two topology assays were entirely consistent for 35 of the 39 proteins, and the location of the reporter could be reliably inferred also for 2 of the remaining 4. In only two cases did we fail to observe both glycosylation and growth on histidinol; a possible explanation is that these proteins are not targeted to the endoplasmic reticulum but rather are imported into the mitochondrial inner membrane and have their C termini in the matrix space. In fact, one of the two proteins (YKR065C) has recently been identified as being located in the inner mitochondrial membrane,[5] although its membrane topology is not yet known.

When combined with theoretical predictions, the experimentally mapped C-terminal locations allow us to propose what we consider are reliable topology models for 37 yeast proteins for which no such information was previously available. To this end, we have used two approaches.

First, all our target proteins were selected from the full set of predicted yeast open reading frames in such a way that five different topology prediction methods all gave the same prediction for each protein. We have previously shown that such consensus predictions are highly reliable for bacterial inner membrane proteins (6, 7). The experimentally determined C-terminal location was the same as the predicted one for 31 of the 37 proteins, and we thus regard the topology models for these proteins as very likely to be correct.

Second, we constrained the TMHMM predictions by fixing the C-terminal end of each protein to the experimentally determined location, because this is known to substantially increase the prediction accuracy (8). We further calculated a reliability score for each predicted topology, both without and with a fixed C-terminal location, Table I. Because there is an approximately linear relationship between the reliability score and the probability that a particular prediction is correct (8), such probability values (expected accuracy) were also calculated. Most of the 37 proteins have high scores compared with the score distribution calculated for all predicted *S. cerevisiae* membrane proteins (8) (data not shown). This was expected, because the proteins in our set were selected based on the requirement that five different topology prediction methods should all give the same predicted topology. We also note that the changes in the reliability score for the 37 proteins seen after the inclusion of the experimentally determined C-terminal location in the prediction have a distribution that is very similar to the one derived for the much larger set of bacterial inner membrane proteins analyzed previously (8) (data not shown).

It is interesting to compare the 37 proteins studied here with TMHMM predictions for the whole *S. cerevisiae* membrane proteome (16). The overall distribution of proteins with different numbers of transmembrane helices is roughly the same for the set of 37 proteins and the whole proteome, with peaks at 2 helices and 10–12 helices. Proteins with an even number of predicted transmembrane helices are 1.8-fold more numerous than proteins with an odd number of helices among the 37 proteins and are 1.7-fold more numerous in the whole proteome (excluding the single-spanning proteins). There are 1.8 times more proteins with C_{in} as compared with C_{out} orientation in our set (1.7 times more in the whole proteome) and 4.3 times more proteins that are predicted to have N_{in} as compared with N_{out} orientation in our set (1.6 times more in the whole proteome). The proteins analyzed here thus seem to represent a rough cross-section of the whole proteome, except that their topologies are easier to predict and have higher TMHMM reliability scores than the proteome as a whole.

In summary, we have shown that reliable topology models for *S. cerevisiae* membrane proteins can be produced on a reasonably large scale by a combination of C-terminal reporter fusion analysis and theoretical prediction. This approach not only reduces the experimental efforts required but also avoids the pitfalls inherent to fusions between a truncated target protein and topology reporters.

Acknowledgment—We thank Johan Nilsson (Stockholm Bioinformatics Center) for bioinformatics analyses.

[5] N. Pfanner, personal communication.

REFERENCES

1. Jayasinghe, S., Hristova, K., and White, S. H. (2001) *Protein Sci.* **10,** 455–458
2. Manoil, C. (1991) *Methods Cell Biol.* **34,** 61–75
3. Drew, D., Sjöstrand, D., Nilsson, J., Urbig, T., Chin, C. N., de Gier, J. W., and von Heijne, G. (2002) *Proc. Natl. Acad. Sci. U. S. A.* **99,** 2690–2695
4. Sengstag, C. (2000) *Methods Enzymol.* **327,** 175–190
5. Green, N., and Walter, P. (1992) *Mol. Cell. Biol.* **12,** 276–282
6. Nilsson, J., Persson, B., and von Heijne, G. (2000) *FEBS Lett.* **486,** 267–269
7. Nilsson, J., Persson, B., and von Heijne, G. (2002) *Protein Sci.,* in press
8. Käll, L., and Sonnhammer, E. (2002) *FEBS Lett.* **532,** 415–418
9. Deak, R., and Wolf, D. (2001) *J. Biol. Chem.* **276,** 10663–10669
10. Oldenburg, K. R., Vo, K. T., Michaelis, S., and Paddon, C. (1997) *Nucleic Acids Res.* **25,** 451–452
11. Breeden, L. L. (1997) *Methods Enzymol.* **283,** 332–341
12. Strahl-Bolsinger, S., and Scheinost, A. (1999) *J. Biol. Chem.* **274,** 9068–9075
13. Ito, H., Fukuda, Y., Murata, K., and Kimura, A. (1983) *J. Bacteriol.* **153,** 163–168
14. Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., and Cherry, J. M. (2002) *Nucleic Acids Res.* **30,** 69–72
15. Goffeau, A., Aert, R., Agostini-Carbone, M., Ahmed, A., Aigle, M., Alberghina, L., Albermann, K., Albers, M., Aldea, M., Alexandraki, D., Aljinovic, G., Allen, E., Altmann, R., Alt-Mörbe, J., André, B., Andrews, S., Ansorge, W., Antoine, G., Anwar, R., Aparicio, A., Araujo, R., Arino, J., Arnold, W., Arroyo, J., Aviles, E., Backes, U., Baclet, M., Badcock, K., Bahr, A., Baladron, V., Ballesta, J., Bankier, A., Banrevi, A., Bargues, M., Baron, L., Barreiros, T., Barrell, B., Barthe, C., Barton, A., Baur, A., Bécam, A., Becker, A., Becker, I., Beinhauer, J., Benes, V., Benit, P., Berben, G., Bergantino, E., Bergez, P., Berno, A., Bertani, I., Biteau, N., Bjourson, A., Blöcker, H., Blugeon, C., Bohn, C., Boles, E., Bolle, P., Bolotin-Fukuhara, M., Bordonné, R., Boskovic, J., Bossier, P., Botstein, D., Bou, G., Bowman, S., Boyer, J., Brandt, P., Brandt, T., Brendel, M., Brennan, T., Brinkman, R., Brown, A., Brown, A., Brown, D., Brückner, M., Bruschi, C., Buhler, J., Buitrago, M., Bussereau, F., Bussey, H., Camasses, A., Carcano, C., Carignani, G., Carpenter, J., Casamayor, A., Casas, C., Castagnoli, L., Cederberg, H., Cerdan, E., Chalwatzis, N., Chanet, R., Chen, E., Chéret, G., Cherry, J., Chillingworth, T., Christiansen, C., Chuat, J., Chung, E., Churcher, C., Churcher, C., et al. (1997) *Nature* **387** (suppl.), 1–105
16. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. (2001) *J. Mol. Biol.* **305,** 567–580
17. Tusnady, G. E., and Simon, I. (1998) *J. Mol. Biol.* **283,** 489–506
18. Tusnady, G. E., and Simon, I. (2001) *Bioinformatics* **17,** 849–850
19. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1994) *Biochemistry* **33,** 3038–3049
20. Rost, B., Fariselli, P., and Casadio, R. (1996) *Protein Sci.* **5,** 1704–1718
21. von Heijne, G. (1992) *J. Mol. Biol.* **225,** 487–494
22. Claros, M. G., and von Heijne, G. (1994) *Comput. Appl. Sci.* **10,** 685–686
23. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. (2002) *Nucleic Acids Res.* **30,** 31–34
24. Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000) *J. Mol. Biol.* **300,** 1005–1016
25. Drawid, A., and Gerstein, M. (2000) *J. Mol. Biol.* **301,** 1059–1075