A global topology map of the *Saccharomyces cerevisiae* membrane proteome

Hyun Kim*[†], Karin Melén*[†][‡], Marie Österberg*[†], and Gunnar von Heijne*[‡][§]

*Center for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University, and [‡]Stockholm Bioinformatics Center, AlbaNova University Center, SE-106 91 Stockholm, Sweden

Communicated by Roderick MacKinnon, The Rockefeller University, New York, NY, May 17, 2006 (received for review March 27, 2006)

The yeast *Saccharomyces cerevisiae* is, arguably, the best understood eukaryotic model organism, yet comparatively little is known about its membrane proteome. Here, we report the cloning and expression of 617 *S. cerevisiae* membrane proteins as fusions to a C-terminal topology reporter and present experimentally constrained topology models for 546 proteins. By homology, the experimental topology information can be extended to ≈15,000 membrane proteins from 38 fully sequenced eukaryotic genomes.

membrane proteins | membrane proteomics | yeast

Subsequent to the determination of the *Saccharomyces cerevisiae* genome sequence (1), a wide variety of genomics and proteomics studies have been carried out, and there is now ample information available on, e.g., gene expression under different conditions (2, 3), gene dispensability (4, 5), protein expression profiles (6), organellar proteomes (7), global protein localization patterns (8, 9), and protein–protein interaction networks (10, 11).

However, the yeast integral membrane proteins are generally underrepresented in the proteomics studies and are even less well represented in the Protein Data Bank (12). In the absence of high-resolution structural data, good topology models provide a necessary background to all structure–function studies of membrane proteins, but, also here, *S. cerevisiae* lags far behind (13).

There is currently no experimental method that makes it possible to derive full-topology models in a high-throughput mode, and one generally has been forced to resort to sequence-based prediction methods to study membrane protein topology on a proteome-wide scale. We have shown that much improved topology models can be achieved by a combination of large-scale experimental mapping of the location of the C terminus of membrane proteins and topology prediction constrained by this information (14). In a first application of this approach, we recently presented experimentally constrained topology models for 601 *Escherichia coli* inner membrane proteins (15) and were able to extend this information to >50,000 bacterial proteins by sequence homology (16).

As a first step toward the exploration of the *S. cerevisiae* membrane proteome, we now report the construction and experimental analysis of a clone collection of >600 predicted membrane proteins. Based on the experimental data, we assign the location of the C termini and produce constrained topology models for 546 polytopic membrane proteins and for ≈15,000 homologous membrane proteins from other eukaryotes. We find that topologies with both the N and C termini of the protein in the cytosol (N_{in}-C_{in}) predominate and that the overall distribution of membrane protein topologies is surprisingly similar between *S. cerevisiae* and *E. coli*.

Results

The most direct way to characterize the *S. cerevisiae* membrane proteome is by constructing strain collections in which each strain expresses a suitably tagged membrane protein. The choice of tag and expression strategy (expression from endogenous promoters on the chromosome or from a plasmid) is dictated by the intended use of the strain collection. Thus, GFP-tagged proteins have been used for subcellular localization studies (9), and chromosomally encoded TAP-tagged proteins have been used to assess endogenous expression levels (6). Here, our focus is on membrane protein topology, and we have consequently used a C-terminal topology reporter tag and expression from a multicopy plasmid.

Experimental Determination of the C-Terminal Location. To determine the location of the C termini of the yeast membrane proteins, we chose the previously characterized HA/Suc2/His4C chimeric protein (17, 18) as a topology reporter, Fig. 1*A*. The Suc2 part contains eight, and the His4C part four, consensus acceptor sites for N-linked glycosylation that will be glycosylated only if the reporter is translocated to the lumen of the endoplasmic reticulum (ER), whereas the His4C catalytic domain of the His4p histidinol dehydrogenase can act on its substrate, histidinol, only if located in the cytosol. The location of the C terminus of any membrane protein–HA/Suc2/His4C fusion can thus be determined by a combination of endoglycosidase H (Endo H) digestion to assess the glycosylation status of the Suc2/His4C part and growth of a *his4*$^-$ strain transformed with the fusion gene on plates lacking histidine but containing histidinol to assess the localization of the His4C domain (18). The hemagglutinin (HA) tag is included to allow identification of the expressed fusion protein by Western blotting. Earlier reports using various C-terminal tags to study global protein expression, localization, and complex formation in *S. cerevisiae* (6, 9, 19) suggest that the stability, localization, and function of most proteins are not compromised by C-terminal fusions.

Eight hundred forty-eight *S. cerevisiae* ORFs predicted by the program TMHMM (20) to encode proteins with at least two transmembrane helices were initially selected for study (see *Methods*). Proteins with a single predicted transmembrane helix were not included, because current prediction programs cannot reliably distinguish between soluble proteins with a cleavable signal sequence and single-spanning membrane proteins with an uncleaved N-terminal signal–anchor sequence. Because the HA/Suc2/His4C topology reporter is suitable only for proteins targeted to the secretory pathway, an additional 58 proteins annotated to be encoded on the mitochondrial chromosome or localized in the mitochondrial or peroxisomal membranes (9, 21, 22) were excluded. Finally, 161 ORFs that were <100 codons long, contained introns or unannotated stop codons, were defined as "spurious" in ref. 6, had already been analyzed by HA/Suc2/His4C fusions (18), or were altered in the *Saccharomyces* Genome Database (SGD) (22) during the course of our study were also excluded. From the resulting set of 629 target proteins, we successfully cloned and expressed 617 proteins as

Conflict of interest statement: No conflicts declared.

Abbreviations: Endo H, endoglycosidase H; ER, endoplasmic reticulum; GO, Gene Ontology; SGD, *Saccharomyces* Genome Database.

[†]H.K., K.M., and M.Ö. contributed equally to this work.

[§]To whom correspondence should be addressed. E-mail: gunnar@dbb.su.se.

www.pnas.org/cgi/doi/10.1073/pnas.0604075103

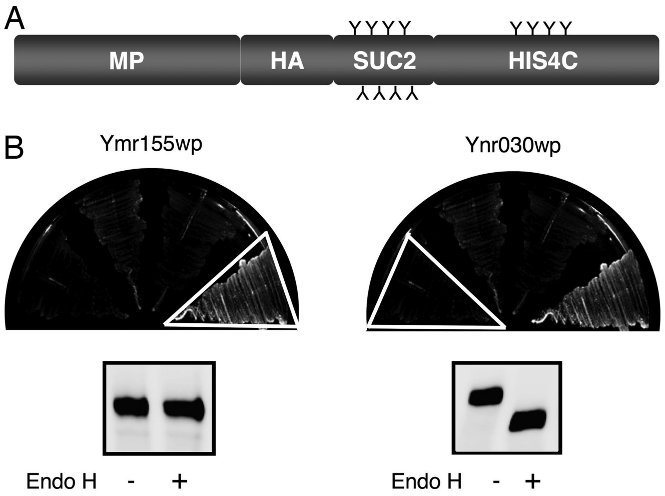

Fig. 1. Topology mapping using the HA/Suc2/His4C reporter. (*A*) Schematic diagram of the membrane protein (MP) reporter fusions. Consensus acceptor sites for N-linked glycosylation are indicated (Y). (*B*) Growth and N-glycosylation assays. Plasmid-transformed *his4*$^-$ cells were streaked on His-negative plates supplemented with 6 mM histidinol and incubated at 30°C for 2–3 days. N-glycosylation was assessed by Endo H treatment of whole-cell lysates, and identification of MP–reporter fusions was determined by Western blotting using an anti-HA antibody. The Ymr155wp fusion grows on His-negative plates, and the SUC2 domain is not glycosylated, demonstrating that the C terminus is cytosolic (C_{in}); the Ynr030wp fusion behaves in the opposite way and is, therefore, assigned as C_{out}.

assessed by Western blot analysis using an antibody directed against the HA epitope in the reporter (23). Typical results from the growth and Endo H assays are shown in Fig. 1*B*.

Among the 617 fusion proteins, 468 (76%) gave consistent results in the two assays (i.e., if the protein was not glycosylated, the *his4*$^-$ transformant expressing the fusion protein grew on histidinol or vice versa) and their C-terminal locations were assigned accordingly (Table 2, which is published as supporting information on the PNAS web site). Among the remaining 147 unassigned proteins, 76 gave multiple bands on the gel, 48 were neither glycosylated nor grew on histidinol, 16 were both glycosylated and grew on histidinol, 4 were too large to allow the reliable detection of a shift in molecular mass upon Endo H digestion, and 3 had an intermediate growth phenotype. Notably, one of two proteins identified in a previous study that were neither glycosylated nor grew on histidinol (18) was later found to be localized to mitochondria (24), suggesting a mitochondrial location for at least some of the 48 such proteins found here. Indeed, of these 48 proteins, 4 (Ugo1p, Aus1p, Fre5p, and Vmr1p) are annotated as mitochondrial in the latest update of the SGD (ftp://ftp.yeastgenome.org/yeast), and an additional three (Yor071cp, Yor192cp, and Yor291wp) are predicted by the program TARGETP (25) to have a mitochondrial targeting peptide. Four peroxisomal proteins (Pex28p, Pex29p, Pex31p, and Pex32p) included in the study all behaved as if they have a cytosolic (C_{in}) orientation, suggesting that the His4C-fusion approach may possibly be used to map the topology of peroxisomal membrane proteins.

Twenty of the proteins that we have annotated as having a lumenal (C_{out}) orientation contain >4 potential internal glycosylation sites, which, if they were all localized in the ER lumen and modified, could complicate the interpretation of the glycosylation status of the Suc2p reporter and, hence, the C_{out} assignment. However, should any such internally glycosylated (and, hence, properly ER-targeted) protein have a cytosolically localized HA/Suc2/His4C reporter, it is very unlikely that the His4Cp part would not confer growth on histidinol plates.

Table 1. Proteins with independently mapped C-terminal locations

ORF	Protein	This study	Previously reported	Reference
YNL275W	Bor1p	In	In	41
YBR201W	Der1p	Out	In	42
YDR411C	Dfm1p	In	In	42
YBR207W	Fth1p	In	In	43
YBR021W	Fur4p	In	In	44
YKR039W	Gap1p	In	In	45
YGL084C	Gup1p	In	In	46
YGL008C	Pma1p	In	In	47
YDL095W	Pmt1p	Out	Out	39
YMR129W	Pom152p	Out	Out	48
YKL212W	Sac1p	In	In	49
YLR378C	Sec61p	In	In	50
YOR254C	Sec63p	In	In	51
YDL212W	Shr3p	In	In	52
YDR410C	Ste14p	In	In	53
YMR117W	Ste24p	In	In	54
YGL022W	Stt3p	In	In	55, 56
YPL234C	Vma11p	Out	Out	57
YEL027W	Vma3p	Out	Out	57

Consistency of Assigned C-Terminal Locations. Although the Suc2/His4C reporter has proven reliable (17, 18, 26), we cannot rule out that it may affect the C-terminal orientation in some fraction of the proteins analyzed here. As a first test of the reliability of our assignments, we searched the SGD (22) and the literature for *S. cerevisiae* membrane proteins with independently determined C-terminal locations. Among 19 such proteins found, only one (Der1p) has a reported C-terminal location different from our assignment, Table 1.

As a further validation, we performed an all-against-all BLAST search (27) among the assigned proteins (including 37 assignments from an earlier study) (18). We retained all pair-wise hits with an E value $<10^{-5}$ and for which the BLAST alignment reached within 15 residues of the C termini of both the query and target sequences. With these restrictions, it is unlikely that there would be an additional transmembrane segment between the end of the alignment and the C terminus of either the query or the target sequence, and homologs found in this way can be assumed to have the same C-terminal orientation (16).

Among the 153 proteins that matched our search criteria, only two, Ynr002cp and Ygl263wp, matched homologs with an opposite C-terminal assignment. Ynr002cp belongs to a family of ATO (ammonia/ammonium transport outward) proteins (28) of which only three were represented in our data set and was not studied further. Ygl263wp is a member of the larger COS (conserved sequence) family (29). The other eight COS family members in our data set were all assigned with a cytosolic C terminus (C_{in}), whereas Ygl263wp was assigned with a lumenal C terminus (C_{out}). To confirm the C_{out} orientation of Ygl263wp, we took advantage of three potential N-glycan acceptor sites in the protein (N^{32} in the N-terminal tail, N^{206} in the second predicted loop, and N^{309} in the C-terminal tail) (Fig. 2*A*). If the protein has a C_{out} orientation, as suggested by the HA/Suc2/His4C fusion, all three sites are predicted to be exposed to the ER lumen and, hence, should be glycosylated. *YGL263W* was recloned with the HA/Suc2/His4C reporter replaced by an HA/His_8 tag at the C terminus, and the glycosylation status of the protein was assessed by Endo H digestion. Indeed, all three sites appeared to be glycosylated (Fig. 2*B*), confirming the C_{out} assignment. Three other members of the COS family (Ydl248wp, Yhl048wp, and Ynr075wp) have potential glycosylation acceptor

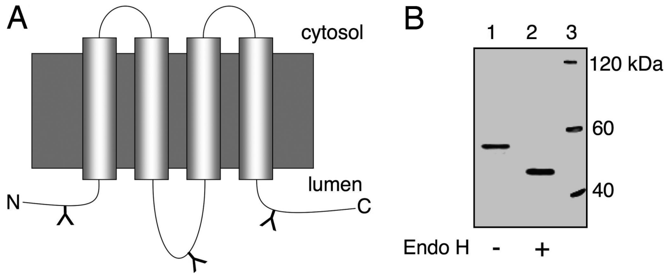

Fig. 2. The C terminus of Ygl263wp is on the extracytosolic side of the membrane. (*A*) Topology model for Ygl263wp based on a COS family consensus prediction with four transmembrane segments and the experimentally determined C_{out} orientation. The three potential N-linked glycosylation acceptor sites are indicated (Y). (*B*) The modification of the three internal N-glycosylation sites in Ygl263wp assessed by Endo H digestion of an HA/His$_8$-tagged full-length protein lacking the C-terminal topology reporter. Samples were analyzed by 10% SDS/PAGE, followed by Western blotting using an anti-HA antibody. Molecular mass standards are included in lane 3.

sites in the predicted loop between transmembrane helices 2 and 3, and Ynr075wp has an additional site in the C tail; none of these sites were glycosylated when the proteins were expressed with a C-terminal HA/His$_8$ tag (data not shown), consistent with their assigned C_{in} orientation. We conclude that Ygl263wp, although clearly homologous to the other COS family members, is oppositely oriented in the membrane. This is in accordance with the "positive inside" rule (30), because Ygl263wp has a higher number of positively charged residues in the short predicted loops between transmembrane segments 1/2 and 3/4 as compared with the other COS family proteins (a total of six Lys and Arg for Ygl263wp vs. two or three for the other COS proteins). Although a few homologous proteins with opposite topologies have been found in *E. coli* (15, 31, 32), we are not aware of another instance of this phenomenon in *S. cerevisiae*. Based on the validation tests, we further conclude that the error rate in our C-terminal assignments is, at most, a few percent.

We next tried to extend the assignments to the remaining unassigned proteins using the same BLAST-based approach as used for the internal consistency test described above (but excluding the 12 proteins in the ATO and COS families). In this way, the C-terminal location could be assigned for an additional 41 proteins, increasing the total number of assigned proteins (including the 37 previously published assignments) (18) to 546. For 69% of the 546 proteins, the initial TMHMM prediction of the C-terminal location agreed with the experimental result (for *E. coli*, the corresponding figure is 78%) (15). The inclusion of the C-terminal assignments thus leads to a major improvement in topology-prediction quality; earlier estimates suggest that the fraction of correctly predicted topologies for the *S. cerevisiae* membrane proteome should increase from ≈53% to ≈68% when they are constrained by known C-terminal locations (14).

Topology Models. Topology models for the 546 proteins were produced by using the C-terminal locations as constraints for TMHMM and for the more recent PRODIV-TMHMM predictor (33) that also takes sequence conservation into account, Table 2. Compared with the unconstrained predictions, the constrained predictions generally have higher TMHMM reliability scores (14), as expected (data not shown).

As seen from the TMHMM predictions in Fig. 3 [and from the corresponding PRODIV-TMHMM results (Fig. 5, which is published as supporting information on the PNAS web site)], proteins with a C_{in} orientation are four times more frequent than those with a C_{out} orientation (82% vs. 18%), and, for the C_{in} proteins, topologies with an even number of predicted transmembrane helices dominate (note that proteins with a single predicted transmembrane helix were not studied, as explained above). The main difference between the TMHMM and PRODIV-TMHMM predictions is that PRODIV-TMHMM predicts more N_{in}-12TM-C_{in} and fewer N_{in}-10TM-C_{in} and N_{out}-11TM-C_{in} topologies (Fig. 5). The topology distribution characterizing the *S. cerevisiae* membrane proteome is strikingly similar to the distribution obtained for the *E. coli* inner membrane proteome (15) (Figs. 3*B* and 5), with only a couple of obvious exceptions: there is a higher fraction of proteins with N_{in}-6TM-C_{in} topology (mainly transporters) in *E. coli* and a higher fraction of proteins with a N_{out}-7TM-C_{in} topology in *S. cerevisiae*. The latter is the classic G protein-coupled receptor (GPCR) topology (34), but, so far, only three bona fide GPCR proteins (Ste2p, Ste3p, and Gpr1p) are listed in the SGD.

Some general functional categories, as described by Gene Ontology (GO) terms (35), correlate strongly with the number

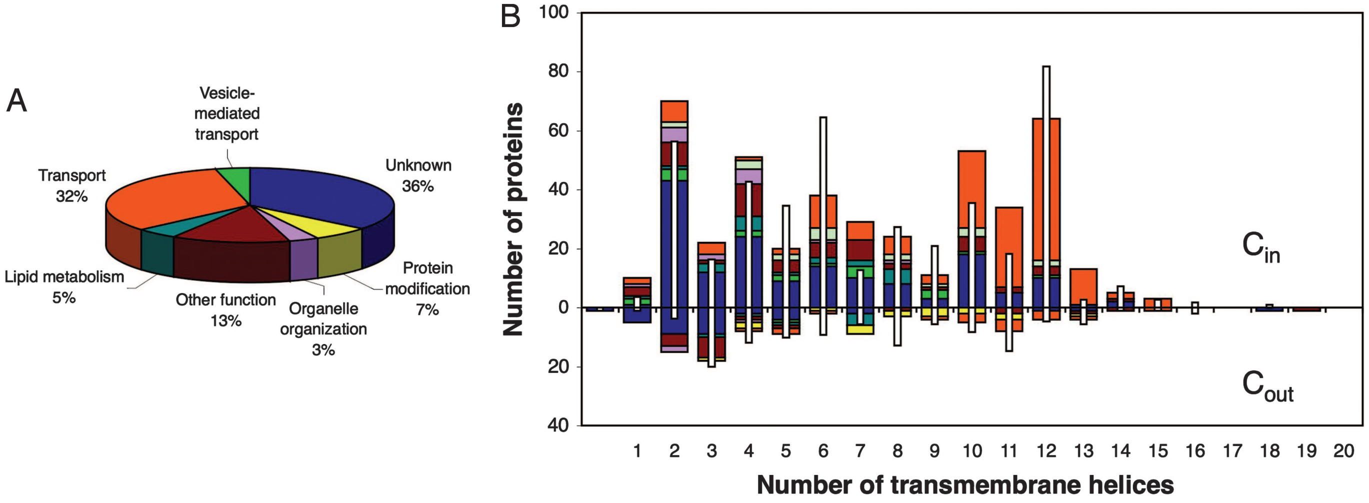

Fig. 3. The yeast membrane proteome. (*A*) GO annotations (GO slim terms) (35) for the 546 proteins with an assigned C-terminal location. (*B*) The different GO categories in *A* subdivided with respect to topology. C_{in} topologies are plotted upward and C_{out} downward. Although all predicted single-spanning membrane proteins were excluded from the study, a couple of proteins initially predicted to have two transmembrane helices were predicted with only one transmembrane helix in the C-terminally constrained models. The corresponding results for 601 *E. coli* inner membrane proteins (15) (scaled to match the 546 yeast proteins) are shown as white bars inside the bars representing the *S. cerevisiae* proteins.

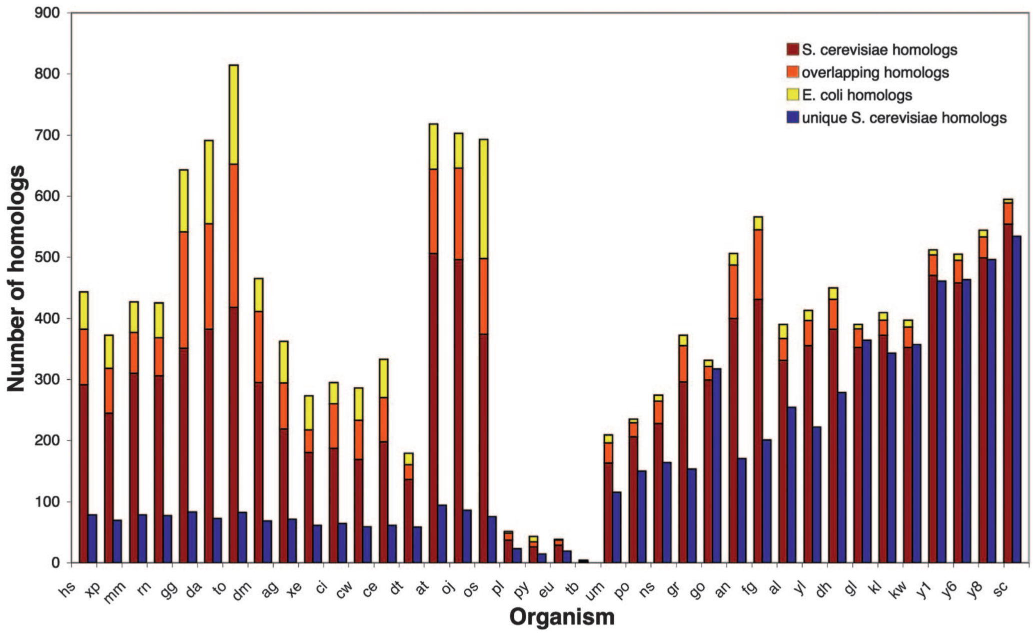

Fig. 4. Homologs in 38 fully sequenced eukaryotic genomes to 534 assigned *S. cerevisiae* proteins and to 612 previously analyzed *E. coli* proteins (15) for which the BLAST alignment extends to within 15 residues of the C terminus of both query and target proteins. Dark red bars represent the total number of primary and secondary homologs (see *Results*) that are assigned only by *S. cerevisiae* proteins in each organism; yellow bars represent primary and secondary homologs assigned only by *E. coli* proteins; orange bars represent primary and secondary homologs that are assigned by both *S. cerevisiae* and *E. coli* proteins. Dark blue bars show the number of unique *S. cerevisiae* proteins that have at least one homolog in the other genome. Abbreviations are from the Superfamily database (38): hs, *H. sapiens*; xp, *P. troglodytes*; mm, *M. musculus*; rn, *R. norvegicus*; gg, *G. gallus*; da, *D. rerio*; to, *F. rubripes*; dm, *D. melanogaster*; ag, *A. gambiae*; xe, *X. tropicalis*; ci, *C. intestinalis*; cw, *C. briggsae*; ce, *C. elegans*; dt, *D. discoideum*; at, *A. thaliana*; oj, *O. sativa* ssp. *japonica*; os, *O. sativa* ssp. *indica*; pl, *P. falciparum*; py, *P. yoelii* ssp. *yoelii*; eu, *E. cuniculi*; tb, *T. brucei*; um, *U. maydis*; po, *S. pombe*; ns, *N. crassa*; gr, *M. grisea*; go, *A. gossypii*; na, *A. nidulans*; fg, *F. graminearum*; al, *C. albicans*; yl, *Y. lipolytica*; dh, *D. hansenii*; gl, *C. glabrata*; kl, *K. lactis*; kw, *K. waltii*; y1, *S. bayanus*; y6, *S. mikatae*; y8, *S. paradoxus*; sc, *S. cerevisiae*.

of transmembrane helices. As seen in Fig. 3, proteins with 10 or more transmembrane helices are nearly always involved in solute transport, whereas the majority of proteins with <5 transmembrane helices have no annotated function. Unexpectedly, among the membrane proteins classified by GO as involved in protein modification (mostly ER-localized proteins), 50% have a C_{out} orientation, whereas a strong C_{in}/C_{out} bias is evident in all of the other major GO classes.

Homology-Based Annotation of Membrane Proteins in Other Organisms. As described in ref. 16, experimentally determined C-terminal locations can be used to assign C-terminal locations to membrane proteins in other organisms based on sequence homology. The basic assumption behind this approach is that pairs of homologous sequences for which a BLAST alignment extends to within 15 residues of both C termini have the same C-terminal location. Although homologous proteins with opposite C-terminal orientations are known, this is a very rare phenomenon: As noted above, we have so far identified only two possible cases in *S. cerevisiae*, and, in 225 bacterial genome sequences, we have identified only five protein families that appear to contain oppositely oriented proteins (32).

We used 534 of the assigned *S. cerevisiae* proteins (excluding the 12 proteins in the ATO and COS families, see above) as queries in BLAST searches against a database of 139,234 predicted membrane proteins from 38 fully sequenced eukaryotic genomes (see *Methods*). The first search was carried out with a very conservative E value cutoff of 10^{-6}; this search gave a total of 7,583 proteins ("primary homologs") for which we could transfer the C-terminal assignment from the yeast query to the target sequence. The BLAST search was then iterated once, by using the 8,117 (7,583 + 534) sequences as queries, yielding an additional 5,698 proteins ("secondary homologs") for which the orientation of the C termini could again be assigned. In this way, the C-terminal location of the original 534 *S. cerevisiae* proteins could be transferred to a total of 13,281 eukaryotic proteins from 38 organisms (Fig. 4).

In the same way, 612 *E. coli* inner membrane proteins with assigned C-terminal locations (15, 16) were used as queries in a BLAST search (one iteration) against the database of 139,234 predicted eukaryotic membrane proteins, yielding 4,051 eukaryotic homologs for which the C-terminal locations could be assigned based on the *E. coli* data set. Of these, 2,522 proteins were found also by using the yeast data set. In all cases but one (the ATO family protein Ynr002cp, see above), the C-terminal assignments based on the *S. cerevisiae* and *E. coli* data sets were the same, supporting the assumption that C-terminal assignments can be transferred between close homologs. Combining the results for the *S. cerevisiae* and *E. coli* homologs, the C-terminal location could be assigned for a total of 14,810 eukaryotic proteins (including 443 human proteins) (Fig. 4); the constrained TMHMM and PRODIV-TMHMM topology predictions are listed in Table 3, which is published as supporting information on the PNAS web site. Notably, proteins with homologs in both *S. cerevisiae* and *E. coli* are often plasma membrane transporters, whereas proteins with homologs in *S. cerevisiae* and mammalian organisms tend to be located in intracellular organelles along the secretory pathway (data not shown). In particular, ER proteins involved in protein translocation, N-glycosylation, glycosylphosphatidylinisotol anchoring, and lipid synthesis are overrepresented among the S. *cerevisiae*/human homologs.

Conclusions

Considering the central role of *S. cerevisiae* as a model eukaryotic organism, it is disturbing that so little is known about its membrane proteome. Based on an analysis of a collection of >600 membrane proteins fused to a C-terminal HA/Suc2/His4Cp reporter tag, we now present the first global view of the topology of the *S. cerevisiae* membrane proteins (excluding single-spanning, mitochondrial, and peroxisomal membrane proteins).

Using a combination of glycosylation and growth assays, we have assigned the in/out location of the C termini of 546 *S. cerevisiae* membrane proteins and have produced experimentally constrained topology models for these proteins. As seen for the *E. coli* inner membrane proteome (15), the large majority of the *S. cerevisiae* membrane proteins have their N and C termini localized in the cytosol and, hence, contain an even number of transmembrane helices, suggesting that pairs of closely spaced transmembrane helices connected by a short extracytosolic loop (“helical hairpins”) may be fundamental building blocks in helical membrane proteins and may be particularly easy to handle for the ER translocon.

The experimental C-terminal localization data sets now available for the *E. coli* (15) and *S. cerevisiae* membrane proteomes can be used to deduce the C-terminal localizations of both prokaryotic and eukaryotic membrane proteins, based on homology to the experimentally mapped proteins. Using a simple BLAST-based approach with conservative cutoffs, we can already annotate >50,000 prokaryotic (16) and nearly 15,000 eukaryotic membrane proteins in this way and, hence, make it possible to produce experimentally constrained topology models on a large scale by using prediction programs such as TMHMM, PRODIV-TMHMM, and HMMTOP (36). Beyond the obvious value of improving topology models for tens of thousands of proteins, the experimental C-terminal localization data may also find use in the development and bench-marking of novel topology-prediction methods.

Methods

Data Sets. All 6,355 predicted ORFs in the *S. cerevisiae* genome were downloaded from the SGD (22) at ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/orf_protein on June 23, 2003. Thirty-eight fully sequenced eukaryotic genomes (including the April 2, 2004, version of the *S. cerevisiae* genome), together containing 657,284 predicted protein sequences, were downloaded on March 8, 2005, from either the Ensembl database (37) (*Anopheles gambiae*, *Danio rerio*, *Fugu rubripes*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Pan troglodytes*, and *Rattus norvegicus*) or the Superfamily 1.67 database (38) (*Ashbya gossypii*, *Aspergillus nidulans*, *Arabidopsis thaliana*, *Caenorhabditis briggsae*, *Caenorhabditis elegans*, *Candida albicans*, *Candida glabrata*, *Ciona intestinalis*, *Debaromyces hansenii*, *Drosophila melanogaster*, *Dictyostelium discoideum*, *Encephalitozoon cuniculi*, *Fusarium graminearum*, *Kluyveromyces lactis*, *Kluyveromyces waltii*, *Magnaporthe grisea*, *Neurospora crassa*, *Oryza sativa* ssp. *indica*, *Oryza sativa* ssp. *japonica*, *Plasmodium falciparum*, *Plasmodium yoelii* ssp. *yoelii*, *Saccharomyces bayanus*, *Saccharomyces mikatae*, *Saccharomyces paradoxus*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Trypanosoma brucei*, *Ustilago maydis*, *Xenopus tropicalis*, and *Yarrowia lipolytica*). GO annotations were downloaded from the SGD database (22), version 2006-05.

Yeast Genetic Manipulations. Genomic DNA was purified from *S. cerevisiae* strain W303-1a (*MATa ade2 can1 his3 leu2 trp1 ura3*) by using the Wizard Genomic DNA purification kit (Promega) and was used as a template to amplify each gene by PCR. Primers were designed as described (18) and purchased from MWG Biotech (Ebersberg, Germany). All plasmids were constructed by homologous recombination with a linearized pJK90 vector (26) in strain STY50 (*MATa, his4-401, leu2-3, -112, trp1-1, ura3-52, hol1-1, SUC2*::*LEU2*) (39) as described in ref. 18. Positive clones were selected by colony PCR using one primer specific for the gene and the other complementing a vector sequence.

To construct plasmids carrying a His_8 tag followed by a stop codon between the HA-tag and the start of *SUC2* in pJK90, a His_8 fragment with a stop codon at the end of the His_8 sequence was amplified from an unpublished plasmid (a kind gift from Dr. N. Meindl-Beinker, Stockholm). The vector was used for subcloning of *YGL263W* and other members of the COS family. Yeast homologous recombination was carried out in strain STY50 (39) with XhoI linearized pJK90 (26) and the His_8 PCR product. Transformants were selected on Ura-negative plates, plasmid was isolated, and the correct sequence was confirmed by DNA sequencing.

Topology Determination. Growth of transformants on medium supplemented with histidinol was assessed as described (18). For Endo H treatment, cells were grown in 5 ml of Ura-negative medium to stationary phase, harvested, and washed once with 5 ml of distilled (d)H_2O. The cell pellet was incubated at −20°C for a minimum of 2 h, resuspended in 100 μl of sample buffer (6), and centrifuged at 13,000 × *g* for 5 min, and the supernatant fractions were incubated at 56°C for 15 min. Eighteen microliters of the supernatant was mixed with 18 μl of dH_2O and 4 μl of buffer (800 mM sodium acetate, pH 5.7), 1.5 μl of Endo H (5 units/ml; Roche) or dH_2O (for a mock sample) was added, and the sample was incubated at 37°C for 2 h. To stop the Endo H reaction, the sample was incubated at 56°C for 15 min before loading on a 6.25% SDS gel. Western blotting was performed by using an anti-HA antibody (Babco, Richmond, CA) and visualized by using the ECL plus kit (Amersham Pharmacia, Uppsala)

Bioinformatics Analysis. The membrane proteome of *S. cerevisiae* was defined by applying the hidden Markov model topology predictor TMHMM (20) to the 6,355 predicted ORFs. The prediction results were divided into three categories: 4,990 nonmembrane proteins, 517 single-spanning proteins, and 848 multispanning proteins. To reduce the risk of selecting secretory proteins with a cleavable N-terminal signal sequence, only the 848 ORFs with at least two predicted transmembrane helices were selected for the experimental study. Because the experimental setup was designed for proteins targeted to the secretory pathway, we excluded 58 ORFs annotated or known from literature to be located in mitochondria or peroxisomes (9, 22, 40). We further excluded 161 ORFs that were <100 residues, contained introns or unannotated stop codons, were defined as “spurious” in ref. 6, were previously analyzed in ref. 18, or had been altered in the SGD database during the study (YBR074W and YBR075W merged). The final set chosen for the topology-mapping experiments contained 629 integral membrane proteins.

TMHMM was also used to define the membrane proteomes of the 38 eukaryotic genomes listed above and identified a set of 139,234 proteins with at least one predicted transmembrane helix.

Finally, TMHMM and PRODIV-TMHMM (33) were used to generate topology models for the set of 546 yeast proteins as well as for the full set of 14,810 eukaryotic membrane protein homologs (13,281 homologs to 534 *S. cerevisiae* proteins studied here and 1,529 homologs to 612 annotated *E. coli* proteins) (15) by constraining the predictions with the assigned locations of the C termini (14).

This work was supported by grants from the Swedish Research Council, the Marianne and Marcus Wallenberg Foundation, the Swedish Foundation for Strategic Research, and the Swedish Cancer Foundation (to G.v.H.) and the Swedish Knowledge Foundation (to K.M.).

1. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., *et al.* (1996) *Science* **274,** 562–567.
2. Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. & Barkai, N. (2002) *Nat. Genet.* **31,** 370–377.
3. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., *et al.* (2002) *Nature* **418,** 387–391.
4. Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., *et al.* (1999) *Science* **285,** 901–906.
5. Warringer, J., Ericson, E., Fernandez, L., Nerman, O. & Blomberg, A. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 15724–15729.
6. Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K. & Weissman, J. S. (2003) *Nature* **425,** 737–741.
7. Sickmann, A., Reinders, J., Wagner, Y., Joppich, C., Zahedi, R., Meyer, H. E., Schonfisch, B., Perschil, I., Chacinska, A., Guiard, B., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 13207–13212.
8. Kumar, A., Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., *et al.* (2002) *Genes Dev.* **16,** 707–719.
9. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S. & O'Shea, E. K. (2003) *Nature* **425,** 686–691.
10. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002) *Nature* **417,** 399–403.
11. Miller, J. P., Lo, R. S., Ben-Hur, A., Desmarais, C., Stagljar, I., Noble, W. S. & Fields, S. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 12123–12128.
12. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) *Nucleic Acids Res.* **28,** 235–242.
13. Jayasinghe, S., Hristova, K. & White, S. H. (2001) *Protein Sci.* **10,** 455–458.
14. Melén, K., Krogh, A. & von Heijne, G. (2003) *J. Mol. Biol.* **327,** 735–744.
15. Daley, D. O., Rapp, M., Granseth, E., Melén, K., Drew, D. & von Heijne, G. (2005) *Science* **308,** 1321–1323.
16. Granseth, E., Daley, D. O., Rapp, M., Melén, K. & von Heijne, G. (2005) *J. Mol. Biol.* **352,** 489–494.
17. Deak, R. & Wolf, D. (2001) *J. Biol. Chem.* **276,** 10663–10669.
18. Kim, H., Melén, K. & von Heijne, G. (2003) *J. Biol. Chem.* **278,** 10208–10213.
19. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., *et al.* (2002) *Nature* **415,** 141–147.
20. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. (2001) *J. Mol. Biol.* **305,** 567–580.
21. Riley, M. L., Schmidt, T., Wagner, C., Mewes, H. W. & Frishman, D. (2005) *Nucleic Acids Res.* **33,** D308–D310.
22. Christie, K. R., Weng, S., Balakrishnan, R., Costanzo, M. C., Dolinski, K., Dwight, S. S., Engel, S. R., Feierbach, B., Fisk, D. G., Hirschman, J. E., *et al.* (2004) *Nucleic Acids Res.* **32,** D311–D314.
23. Österberg, M., Kim, H., Warringer, J., Melén, K., Blomberg, A. & von Heijne, G. (2006) *Proc. Natl. Acad. Sci. USA* **103,** 11148–11153.
24. van der Laan, M., Chacinska, A., Lind, M., Perschil, I., Sickmann, A., Meyer, H. E., Guiard, B., Meisinger, C., Pfanner, N. & Rehling, P. (2005) *Mol. Cell. Biol.* **25,** 7449–7458.
25. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. (2000) *J. Mol. Biol.* **300,** 1005–1016.
26. Kim, H., Yan, Q., von Heijne, G., Caputo, G. A. & Lennarz, W. J. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 7460–7464.
27. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
28. Palkova, Z., Devaux, F., Icicova, M., Minarikova, L., Le Crom, S. & Jacq, C. (2002) *Mol. Biol. Cell* **13,** 3901–3914.
29. Spode, I., Maiwald, D., Hollenberg, C. P. & Suckow, M. (2002) *J. Mol. Biol.* **319,** 407–420.
30. von Heijne, G. (1986) *EMBO J.* **5,** 3021–3027.
31. Sääf, A., Johansson, M., Wallin, E. & von Heijne, G. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 8540–8544.
32. Rapp, M., Seppälä, S., Granseth, E. & von Heijne, G. (2006) *Nat. Struct. Mol. Biol.* **13,** 112–116.
33. Viklund, H. & Elofsson, A. (2004) *Protein Sci.* **13,** 1908–1917.
34. Wistrand, M., Käll, L. & Sonnhammer, E. L. (2006) *Protein Sci.* **15,** 509–521.
35. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000) *Nat. Genet.* **25,** 25–29.
36. Tusnady, G. E. & Simon, I. (2001) *Bioinformatics* **17,** 849–850.
37. Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., *et al.* (2005) *Nucleic Acids Res.* **33,** D447–D453.
38. Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001) *J. Mol. Biol.* **313,** 903–919.
39. Strahl-Bolsinger, S. & Scheinost, A. (1999) *J. Biol. Chem.* **274,** 9068–9075.
40. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. & Weil, B. (2002) *Nucleic Acids Res.* **30,** 31–34.
41. Zhao, R. & Reithmeier, R. A. (2001) *Am. J. Physiol. Cell Physiol.* **281,** C33–C45.
42. Hitt, R. & Wolf, D. H. (2004) *FEMS Yeast Res.* **4,** 721–729.
43. Urbanowski, J. L. & Piper, R. C. (1999) *J. Biol. Chem.* **274,** 38061–38070.
44. Garnier, C., Blondel, M. O. & Haguenauer-Tsapis, R. (1996) *Mol. Microbiol.* **21,** 1061–1073.
45. Gilstring, C. F. & Ljungdahl, P. O. (2000) *J. Biol. Chem.* **275,** 31488–31495.
46. Bleve, G., Zacheo, G., Cappello, M. S., Dellaglio, F. & Grieco, F. (2005) *Biochem. J.* **390,** 145–155.
47. Monk, B. C., Montesinos, C., Ferguson, C., Leonard, K. & Serrano, R. (1991) *J. Biol. Chem.* **266,** 18097–18103.
48. Tcheperegine, S. E., Marelli, M. & Wozniak, R. W. (1999) *J. Biol. Chem.* **274,** 5252–5258.
49. Konrad, G., Schlecker, T., Faulhammer, F. & Mayinger, P. (2002) *J. Biol. Chem.* **277,** 10547–10554.
50. Wilkinson, B. M., Critchley, A. J. & Stirling, C. J. (1996) *J. Biol. Chem.* **271,** 25590–25597.
51. Feldheim, D., Rothblatt, J. & Schekman, R. (1992) *Mol. Cell. Biol.* **12,** 3288–3296.
52. Ljungdahl, P. O., Gimeno, C. J., Styles, C. A. & Fink, G. R. (1992) *Cell* **71,** 463–478.
53. Romano, J. D. & Michaelis, S. (2001) *Mol. Biol. Cell* **12,** 1957–1971.
54. Tam, A., Schmidt, W. K. & Michaelis, S. (2001) *J. Biol. Chem.* **276,** 46798–46806.
55. Spirig, U., Glavas, M., Bodmer, D., Reiss, G., Burda, P., Lippuner, V., te Heesen, S. & Aebi, M. (1997) *Mol. Gen. Genet.* **256,** 628–637.
56. Kim, H., von Heijne, G. & Nilsson, I. (2005) *J. Biol. Chem.* **280,** 20261–20267.
57. Flannery, A. R., Graham, L. A. & Stevens, T. H. (2004) *J. Biol. Chem.* **279,** 39856–39862.