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

Hyun Kim, Karin Melén, Marie Österberg, and Gunnar von Heijne

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.

---

**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 terminus of the yeast membrane proteins, we chose the previously characterized HA/Suc2/His4C chimeric protein (17, 18) as a topology reporter, Fig. L4. 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.

1H.K., K.M., and M.O. contributed equally to this work.

2To whom correspondence should be addressed. E-mail: gunnar@dibb.su.se.

© 2006 by The National Academy of Sciences of the USA
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. 1B.

Among the 617 fusion proteins, 468 (76%) gave consistent results in the two assays (i.e., if the protein was not glycosylated, the His4C transformant expressing the fusion protein grew on histidinol plates 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 C-terminal assignment. Ynr002cp belongs to a family of ATO (ammonia/ammonium transport outward) proteins (28) of 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 eight other COS family members in our data set were all assigned with a cytosolic C-terminal orientation (16).

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 eight other COS family members in our data set were all assigned with a cytosolic C-terminal orientation (16).

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 eight other COS family members in our data set were all assigned with a cytosolic C-terminal orientation (16).

Table 1. Proteins with independently mapped C-terminal locations

<table>
<thead>
<tr>
<th>ORF</th>
<th>Protein</th>
<th>This study</th>
<th>Previously reported</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>YNL275W</td>
<td>Bor1p</td>
<td>In</td>
<td>In</td>
<td>41</td>
</tr>
<tr>
<td>YBR201W</td>
<td>Der1p</td>
<td>Out</td>
<td>In</td>
<td>42</td>
</tr>
<tr>
<td>YDR411C</td>
<td>Dfm1p</td>
<td>In</td>
<td>In</td>
<td>42</td>
</tr>
<tr>
<td>YBR207W</td>
<td>Fth1p</td>
<td>In</td>
<td>In</td>
<td>43</td>
</tr>
<tr>
<td>YBR021W</td>
<td>Fur4p</td>
<td>In</td>
<td>In</td>
<td>44</td>
</tr>
<tr>
<td>YKR039W</td>
<td>Gap1p</td>
<td>In</td>
<td>In</td>
<td>45</td>
</tr>
<tr>
<td>YGL084C</td>
<td>Gup1p</td>
<td>In</td>
<td>In</td>
<td>46</td>
</tr>
<tr>
<td>YGL008C</td>
<td>Pma1p</td>
<td>In</td>
<td>In</td>
<td>47</td>
</tr>
<tr>
<td>YDL095W</td>
<td>Pmt1p</td>
<td>Out</td>
<td>Out</td>
<td>39</td>
</tr>
<tr>
<td>YMR129W</td>
<td>Pom152p</td>
<td>Out</td>
<td>Out</td>
<td>48</td>
</tr>
<tr>
<td>YKL121W</td>
<td>Sac1p</td>
<td>In</td>
<td>In</td>
<td>49</td>
</tr>
<tr>
<td>YLR378C</td>
<td>Sec61p</td>
<td>In</td>
<td>In</td>
<td>50</td>
</tr>
<tr>
<td>YOR254C</td>
<td>Sec63p</td>
<td>In</td>
<td>In</td>
<td>51</td>
</tr>
<tr>
<td>YDL212W</td>
<td>Shp3p</td>
<td>In</td>
<td>In</td>
<td>52</td>
</tr>
<tr>
<td>YDR410C</td>
<td>Ste14p</td>
<td>In</td>
<td>In</td>
<td>53</td>
</tr>
<tr>
<td>YMR117W</td>
<td>Ste24p</td>
<td>In</td>
<td>In</td>
<td>54</td>
</tr>
<tr>
<td>YGL022W</td>
<td>Ste3p</td>
<td>In</td>
<td>In</td>
<td>55</td>
</tr>
<tr>
<td>YPL234C</td>
<td>Vma11p</td>
<td>Out</td>
<td>Out</td>
<td>56</td>
</tr>
<tr>
<td>YEL027W</td>
<td>Vma3p</td>
<td>Out</td>
<td>Out</td>
<td>57</td>
</tr>
</tbody>
</table>

Kim et al.
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/His8 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 loop between transmembrane helices 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_out orientation are four times more frequent than those with a C_in 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_out-10TM-C_in and N_out-11TM-C_out 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. 3B 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...
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, glycosylphosphatidylinositol 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
mology to the experimentally mapped proteins. Using a simple can be used to deduce the C-terminal localizations of both produce experimentally constrained topology models on a large Methods

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, Dictostelium discoideum, Eねphelozoon coeniculati, Fusarium graminearum, Kluyveromyces lactis, Kluyveromyces waltii, Magnaportha grisea, Neurospora crassa, Oryza sativa ssp. indica, Oryza sativa ssp. japonica, Plasmodium falciparum, Plasmodium yoelii ssp. yoelii, Saccharomyces bayanus, 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 (MATα 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 (MATα, 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 His8 tag followed by a stop codon between the HA-tag and the start of SUC2 in pJK90, a His8 fragment with a stop codon at the end of the His8 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 Xhol linearized pJK90 (26) and the His8 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)H2O. 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 dH2O and 4 μl of buffer (800 mM sodium acetate, pH 5.7), 1.5 μl of Endo H (5 units/ml; Roche) or dH2O (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 (800 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 (Gv.H.) and the Swedish Knowledge Foundation (to K.M.).
Kim et al.


