CMLS, Cell. Mol. Life Sci. 61 (2004) 1341–1353
1420-682X/04/111341-13
DOI 10.1007/s00018-004-4005-3

CMLS Cellular and Molecular Life Sciences

Research Article

Annotating proteins from endoplasmic reticulum and Golgi apparatus in eukaryotic proteomes

K. O. Wrzeszczynski [a–c] and B. Rost [a,c,d,*]

[a] CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, New York 10032 (USA), Fax: +1 212 305 7932, e-mail: rost@cubic.bioc.columbia.edu
[b] Integrated Program in Cellular, Molecular and Biophysical Studies, Columbia University, 630 West 168th Street, New York, New York 10032 (USA)
[c] Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St. Nicholas Avenue, New York, New York 10032 (USA)
[d] North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street, BB217, New York, New York 10032 (USA)

Received 6 January 2004; received after revision 10 March 2004; accepted 29 March 2004

Abstract. The sub-cellular localization of a native protein constitutes one coarse-grained aspect of its function. Transport between compartments is often regulated through short sequence motifs. Here, we analyzed experimentally characterized endoplasmic reticulum (ER)/Golgi retrieval motifs and investigated the accuracy of homology-transfer. Only the C-terminal ER retrieval motifs KDEL, HDEL and AIAKE were sufficiently specific. However, even unspecific motifs may help, provided we know the probability for localization given the motif. We provided such estimates. We also rigorously estimated the accuracy and coverage for inferring ER and Golgi localization through homology-transfer by sequence similarity. In entire proteomes, we could thereby annotate 3304 ER (3182 membrane) and 1853 Golgi (759 membrane) proteins. We identified another putative 5157 globular and 3941 membrane ER or Golgi proteins. Each experimental annotation yielded, on average, one to three high-accuracy and five to six low-accuracy homology-transfers in the six proteomes. These numbers will increase with each new experimental annotation.

Key words. Endoplasmic reticulum; Golgi apparatus; genome sequence analysis; sub-cellular localization; protein sequence motifs.

The major constituents of eukaryotic cells are extra-cellular space, cytoplasm, nucleus, mitochondria, Golgi apparatus, endoplasmic reticulum (ER), peroxisome, and lysosomes. The native sub-cellular localization of a protein is assumed to be determined largely by a trafficking system that is reasonably well captured experimentally for some of the organelles [1–7]. The system has two main branches [8]. On one branch, proteins are synthesized on cytoplasmic ribosomes, and from there can go to the nucleus, mitochondria, or peroxisomes. The second branch leads from the ribosomes attached to the ER to the Golgi apparatus, then to lysosomes, or secretory vesicles, and on to the extra-cellular space. At each branch point in the trafficking system, a 'decision is made': either retain the protein in the current compartment or transport it onward to the next. For many examples, we have experimental evidence that membrane transport complexes 'make these decisions' by recognizing motifs on the shuttled proteins. The most comprehensively characterized branch point is the second one leading to secretion

* Corresponding author.

[9–13]. Most proteins destined for this branch are assumed to have an N-terminal signal peptide that causes their transfer into the ER as they are being synthesized; most proteins lacking this signal are synthesized in the cytoplasm and follow the former mentioned branch of protein trafficking. Some proteins appear to be secreted through a different pathway and clearly lack signal peptides.

Protein sorting through the secretory pathway

The secretory pathway involves a complex protein transport system between its organelles while maintaining no significant loss in organelle resident proteins. This highly selective process permits the post-translational modification and maturation of newly synthesized proteins passing through the ER and Golgi apparatus while strictly sorting and retaining residential soluble and membrane-bound proteins [14, 15]. A small fraction of proteins undergo ER/Golgi-independent protein secretion. This process is performed through at least four distinct pathways under varying cellular conditions [16]. The common assumption is that proteins are kept within the ER and to a much lesser extent within the Golgi through specific short peptides that act as signals important for a retention and retrieval/recycle mediated sorting mechanisms [17–19].

Technically, groups of a few, specific residues are referred to as sequence motifs. Secreted proteins usually contain an N-terminal signal peptide of 10–30 residues that is cleaved upon successful transport through the extra-cellular membranes in the ER [17]; these signal peptides have distinct sequence features and can therefore be predicted accurately for proteins of unknown localization [13, 20]. Conceptually, we can distinguish between three types of motif that are recognized by transport proteins: (i) generic, sequence-consecutive motifs with common features such as cleaved signal peptides, (ii) specific, sequence-consecutive motifs like nuclear localization signals, and (iii) non-sequence consecutive motifs recognizable only after the protein has folded. While binding motifs that require details of the three-dimensional, folded protein structure are common for all kinds of protein function such as small-molecule or enzyme binding, surprisingly few cases have been implicated in the regulation of protein trafficking. One prominent exception are the mannose-6-phosphate (M6P) receptors that bind M6P-containing soluble acid hydrolases in the Golgi and transport them on to the endosomal-lysosmal system [21, 22]. Only the first type of motif for regulation – generic and sequence-consecutive – can currently be predicted to identify proteins with signal peptides [13, 23], chloroplast transit peptides [23, 24], as well as peroxisomal [25] and mitochondrial targeting signals [23, 26]. For the second type of specific, sequence-consecutive motifs, all computational biology can do so far is to archive these in databases that can be queried to find unknown nuclear proteins [27–30].

Computational analysis of ER and Golgi retrieval signals has been limited to PROSITE [31] and PSORT [32] both of which only rely specifically on the classical ER and Golgi retrieval motifs (or conservative derivations of these classical motifs); few attempts have analyzed Golgi and ER proteins at the level of entirely sequenced proteomes [33]. Large-scale genomic consortiums rely in part on valid function and sub-cellular localization information for their target selection decisions. Large-scale experimental efforts can adequately account for a proportion of a specific proteome [34, 35] but are often limited by the experimental design. Therefore, computational efforts are often needed to complete an entire proteomic analysis. Our laboratory has recently reported that sequence conservation established using the HSSP-distance value correlated well with sub-cellular localization [36]. We have further applied this technique specifically to the ER and Golgi organelles. Here, we analyzed the extent to which ER and Golgi proteins can be identified through such short sequence motifs and/or through sequence similarity to proteins known to reside in these two compartments. This analysis required three steps: (i) collect known signals from literature and databases, (ii) build unbiased, trusted data sets of proteins experimentally known to reside in ER and Golgi, and (iii) test specificity and accuracy of the signals found. (Note: we failed to uncover novel motifs through motif-finding algorithms.) The first part of our work explored the limits to predicting ER and Golgi proteins from experimentally known and theoretically refined signals. Next, we established thresholds for significant sequence similarity, i.e., for accurate inference of ER and Golgi location through homology. Finally, we applied our results to annotate ER and Golgi proteins in the proteomes of *Saccharomyces cerevisiae* (yeast), *Drosophila melanogaster* (fruit-fly), *Caenorhabditis elegans* (worm), *Arabidopsis thaliana* (weed), *Homo sapiens* (human), and *Mus musculus* (mouse).

Methods

Trusted data sets of proteins with known localization

We retrieved all proteins from Swiss-Prot [37] that had experimental annotations about sub-cellular localization, removing all with 'putatively known' localization that contained either 'PROBABLE,' 'PUTATIVE,' or 'BY SIMILARITY' as additional qualifiers in Swiss-Prot. We split these proteins into 'trusted ER/Golgi' and 'trusted non-ER/non-Golgi'. As another control data set, we also retrieved all non-eukaryotic Swiss-Prot proteins. The resulting trusted data sets contained 676 ER proteins, 131 lumenal ER proteins, 102 proteins with a [KH]DEL C-terminal motif, 545 ER membrane proteins, 312 trusted

Table S1. Numbers of proteins in trusted data set.

Data set	Trusted all	Trusted unique
ER all	676	114
ER membrane	545	86
ER lumenal	131	21
ER retention motif (KHDEL)	102	14
ER all + yeast-GFP ER	784	144
Golgi all	312	73
Golgi membrane	194	47
Golgi all + yeast-GFP Golgi	351	85

Golgi proteins, and 194 Golgi membrane proteins (numbers summarised in table S1). A non-ER/non-Golgi set of 8417 localization annotated eukaryotic proteins was used to identify false positives. Additional experimentally annotated yeast proteins were collected from the Yeast GFP Fusion Localization Database (http://yeastgfp.ucsf.edu). However, we considered only open reading frames (ORFs) without the keyword 'Hypothetical Protein.' This increased the total number of ER proteins to 784 and the Golgi trusted set to 351. The trusted data sets can be obtained from 'ER-GolgiDB' at http://cubic.bioc.columbia.edu/db/ERGolgiDB.

Data sets for entire proteomes

The human proteins constituted the set currently available in the latest versions of Swiss-Prot (release 40) and TrEMBL (release 22) [37]; *D. melanogaster* was obtained from http://www.fruitfly.org/ (release 2), *M. musculus* was obtained from http://www.ensembl.org/Mus_musculus/ and *C. elegans* from http://www.sanger.ac.uk/Projects/C_elegans/wormpep/ (wormpep 65). All remaining proteomes (plant and yeast) were downloaded from ftp://ncbi.nih.gov/genbank/genomes/.

Aligning proteins

We aligned the trusted ER and Golgi proteins against all proteins of known localization using pairwise BLAST [38]. Next, we built PSI-BLAST profiles for all data sets using a filtered version of all currently known sequences with three iterations [39]. These profiles were then aligned against all proteins of known localization. After compiling the results for the sequence conservation (fig. 1), we changed these profiles such that we only included homologues with HVAL ≥ 40 for ER and ≥ 20 for Golgi proteins. We based our identification of ER/Golgi proteins in entire proteomes on these three data sets: 'trusted,' 'trusted families,' and 'unique subset of trusted families.'

Scores for measuring sequence similarity

The simplest way to measure sequence similarity is percentage pairwise sequence identity (PIDE), i.e., the percentage of residues identical between two proteins (not counting gaps). Another measure is the statistical expectation values as reported by BLAST (EVAL; note: we typically report the logarithm of this value in our figures). As a third measure we used the HSSP-value (HVAL) [40, 41]:

$$\mathrm{HVAL} = \mathrm{PIDE} - \begin{cases} 100 & \text{for } L \leq 11 \\ 480L^{-0.32\cdot(1+\exp^{-L/1000})} & \text{for } L \leq 450 \\ 19.5 & \text{for } L > 450 \end{cases} \quad \text{(Eq.1)}$$

where L was the number of residues aligned between two proteins and PIDE the percentage of pairwise identical residues. The HSSP-value reflects whether an alignment is above the HSSP-curve [40, 41] (HVAL > 0) or below (HVAL < 0). For the first case (> 0), the HSSP-value can be seen as a degree of sequence proximity or similarity (the higher the value the more similar the two proteins), whereas the latter (HVAL < 0) estimates the distance, or level of divergence between two proteins (the more negative the value, the less similar the two proteins). An HSSP-value of 0 defines the line above which (almost) no two naturally evolved proteins differ grossly in their three-dimensional structures. To illustrate the curve: for alignment lengths around 100 residues, 33% pairwise sequence identity suffices to infer structure, above 250 residues 21% is significant, and below 11 residues even 100% identity is not enough to infer structural (or functional) similarity. Although the HSSP-curve was derived to describe structural similarity, we noted that it also constitutes a sensitive approach when distinguishing between proteins of similar and dissimilar enzymatic activity [42], between the largest four compartments (nucleus, extracellular space, cytoplasm, and mitochondria) [36], and between proteins involved in cell cycle control [43].

Sequence-unique subsets

We built sequence-unique subsets for all types of protein under consideration to avoid bias that is likely to skew estimates for accuracy and coverage [42]. 'Sequence-unique' was defined as no pair of proteins in the set having HVAL > 0 (Eq. 1). Given an all-against-all pairwise alignment for the biased set, we simply used a greedy search to find the largest subset that fulfilled the above condition. (Note: a tool performing this type of reduction is available through the web [44].)

Measuring accuracy and coverage

We used the following definition to measure accuracy/specificity:

$$\mathrm{Accuracy} = 100 \times \frac{\text{number of true pairs found above threshold}}{\text{number of all pairs above threshold}} \quad \text{(Eq. 2)}$$

with the thresholds given by either (i) percentage pairwise sequence identity (PIDE), (ii) BLAST expectations values (EVAL), or (iii) the HSSP-value (HVAL). We con-

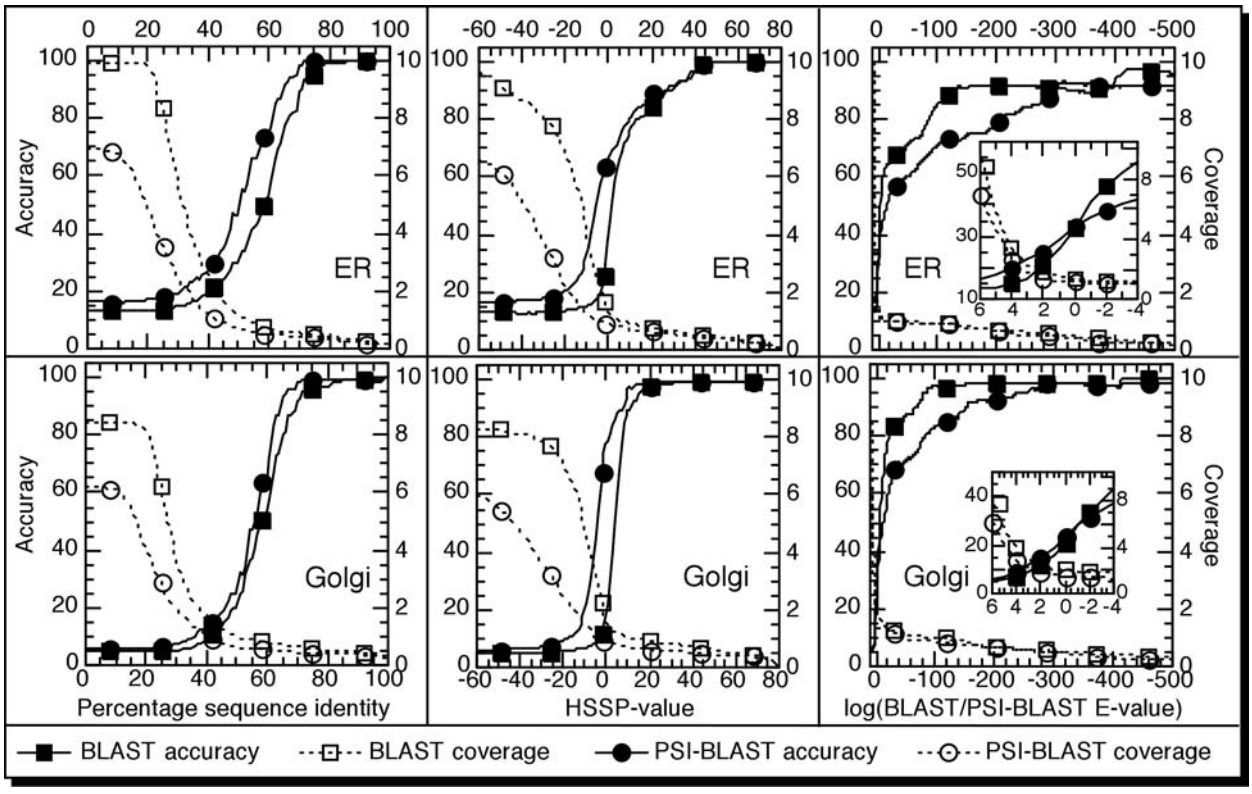

Figure 1. Sequence conservation for ER and Golgi apparatus. We aligned all experimentally annotated, sequence-unique ER and Golgi proteins (ER, upper graphs, Golgi, lower graphs) against all true negatives using BLAST (squares) and PSI-BLAST (circles). Solid lines with filled symbols describe cumulative accuracy/specificity (percentage of correctly identified localized proteins at given threshold, Eq. 2); dotted lines with open symbols describe cumulative coverage/selectivity (true positive proteins found at threshold/all true positive proteins, Eq. 3). We measured sequence similarity in three different ways: (i) by the percentage pairwise sequence identity (left graphs), (ii) by the HSSP-value (Eq. 1, central graphs), and (iii) by the logarithm of the BLAST/PSI-BLAST expectation values (right graphs; note: $\log_{10}$). The inserts in the right graphs focus on accuracy and coverage for less reliable expectation values that capture the region with most of the data.

sidered all pairs as 'true' that were experimentally found in the same sub-cellular compartment. By analogy, we used the following definitions for coverage/sensitivity:

$$\text{Coverage} = 100 \times \frac{\text{number of true pairs found above threshold}}{\text{number of all true pairs}} \quad \text{(Eq. 3)}$$

Results and discussion

Retention and recycle signals can predict unknown ER and Golgi proteins

Collecting retention and retrieval/recycle motifs from literature and databases

We retrieved experimentally annotated retention and recycle signals for ER and Golgi from the literature, Swiss-Prot [37] and PROSITE. The resulting list was tiny in comparison to that obtained previously for nuclear localization signals [28, 30]. The reason supposedly is that many soluble and membrane proteins in the ER have rather specific retention signals (table 1). Predominantly, the C-terminal motifs KDEL, HDEL, and closely related derivatives have been experimentally related to the retrieval mechanism [14]. Other motifs implicated in ER and Golgi targeting include [15]: (i) the C-terminal motif HDEF in the Ca^{2+}-binding protein calumenin [52], (ii) the di-lysine motif KK [46], the di-arginine motif RR [48] or RKR (RxR) [64], (iii) the tyrosine-based tetra-peptide motif Yxxh (where x can be any amino acid and h signifies a hydrophobic residue), predominately associated in vesicular traffic sorting mechanisms [65], which has also been shown as a localization motif as evident in YQRL of TGN38 [58] and for the retrieval of UCE [57], (iv) the di-acidic ER-export motifs [DE]x[DE] often associated with the Yxxh motif [61], (v) the cytoplasmic tail FxFxD motif in DPAP-A necessary for retrieval back to the Golgi [59], and (vi) the targeting domain GRIP, found in peripheral Golgi membrane proteins [62, 63]. The only motifs that were previously available to automatic proteome searches were KDEL and HDEL, as well as some derivatives of these deposited in PROSITE [45] and PSORT

Table 1. Analyzing ER and Golgi retention and retrieval signals.

Sequence motif[a]	Total	Eukaryotes		Non-eukaryotes		ER/Golgi		Non-ER/non-Golgi		Non-annotated	
	N	N	%	N	%	N	%	N	%	N	%
Endoplasmic reticulum[b]											
KDEL-C-term	67	60	90	7	10	60	90	0	0	0	0
KDEL	1201	636	53	565	47	76	6	560	47	230	41
HDEL-C-term	64	64	100	0	0	62	97	2	3	2	100
HDEL	498	261	52	237	48	68	14	193	38	121	63
HDEF-C-term	4	3	75	1	25	2	50	1	25	0	0
HDEF	91	50	55	41	45	2	2	48	53	28	58
KKxx-C-term	907	492	52	415	46	55	6	437	48	211	48
KKxx-C-term (membrane protein subset)	254	183	72	71	28	55	22	128	50	21	16
KKxx	57848	32493	56	25355	44	810	1	31683	55	15171	48
KxKxx-C-term	810	420	52	390	48	42	5	378	47	177	47
KxKxx-C-term (membrane protein subset)	230	139	60	91	40	42	18	97	42	25	26
xxRR	83869	39769	47	44100	53	1062	1	38707	46	16050	41
KKFF-C-term	8	5	63	3	37	3	38	3	25	2	67
KKFF	416	234	56	93	22	7	2	316	76	118	37
KKAA-C-term	29	7	24	22	76	5	17	2	7	0	0
KKAA	1639	824	50	815	50	40	2	784	48	267	34
AIAKE-C-term	10	10	100	0	0	10	100	0	0	0	0
AIAKE	161	55	34	106	66	11	7	44	27	11	25
CRAR	199	127	64	72	36	0	0	127	64	42	33
Golgi apparatus[c]											
YQRL	442	212	48	230	52	10	2	202	46	83	41
YKGL	632	304	48	328	52	5	1	299	47	143	48
YHPL	150	70	47	80	53	7	5	65	43	29	45
Yxxh	135637	62800	46	72837	54	859	1	62941	45	27729	44
NPFKD	17	13	76	4	24	0	0	13	76	8	62
FxFxD	4971	2513	51	2458	49	67	1	2446	49	1101	45
FQFND	7	4	57	3	43	3	43	1	14	1	100
PxPxP	8856	2766	31	4023	45	139	2	4694	53	3088	66
[DE]x[DE]	131139	59784	46	71355	54	834	1	58941	45	25843	44
GRIP-motif[d]	11	11	100	0	0	10	90	1	10	1	100
GRIP-motif (shortened)[e]	58	32	55	24	41	10	18	24	41	11	46
C-term variations[f]											
PROSITE Pattern[g]	232	197	85	35	15	167	72	30	13	13	43
[KH]DEL	131	124	95	7	5	122	93	2	2	2	100
[KHR][DENQ]EL	203	174	86	29	14	157	77	17	8	9	52
[KHR][DENQ][DE]L	230	187	81	43	19	159	72	28	1	13	46
[KHRDENQAS] [DENQIYCV] [DENQ]L	696	428	61	268	39	193	28	235	33	107	45
[KRDEAVYF] [KRDEVYFMQ] [KHED] [DK]EL	80	59	74	21	26	50	63	9	11	5	55

Total: number of proteins found in Swiss-Prot that have the respective motif. N: number of proteins found in subset. %: percentage of proteins in subset (column 'Total' gives 100%). Eukaryotes: all eukaryotic proteins. Non-eukaryotes: since only eukaryotes have ER and Golgi, this column estimates a lower bound for the false positives. ER/Golgi: subset of eukaryotic proteins that have the respective motif and are experimentally known to be in either ER (for ER motifs) or Golgi (for Golgi motifs); this column gives a lower bound for the true positives (percentage is based on total number). Non-ER/non-Golgi: subset of eukaryotic proteins that have the respective motif and are experimentally known to be neither in ER nor in Golgi (percentage is based on total number) or do not contain any sub-cellular localization information in the Swiss-Prot database. Non-annotated: subset of non-ER/non-Golgi which does not contain any localization information in Swiss-Prot. The Non-eukaryotes and Non-ER/non-Golgi columns provide the total FP percentage. The total numbers were: ER = 1060 of which 324 (30%) were annotated as 'Probable, Putative or By Similarity' and 72 (7%) Viral/Prokaryotic/Archaea; Golgi subcellular localization total = 495 of which 163 (33%) were annotated as 'Probable, Putative or By Similarity' and 9 (2%) Viral/Prokaryotic/Archaea.

[a] 'C-term' indicates the carboxy-terminal (last) residue of the protein; motifs are given by the one-letter code of the respective amino acids with the following conventions: [AG] means either A or G, 'x' stands for 'any' amino acid, 'h' stands for any hydrophobic amino acid.

[b] Source of ER motifs: KDEL [14], HDEL [14], KKxx [46] [47], xxRR [48], KKFF [49], KKAA [50, 51], HDEF [52], AIAKE [53], CRAR [54].

[c] Source of Golgi motifs: YQRL [55], YKGL [56], YHPL [57], Yxxh [58], NPFKD [57], FxFxD [59], FQFND [59], PxPxP [60], [DE]x[DE] [61], GRIP-motif [62, 63].

[d] The consensus pattern of the GRIP-motif is described by: [DEA]Y[LIT][KR][KHN][VI][VILF]XX[YF][MIL].

[e] Shortened derivative of GRIP-motif: [DEA]Y[LIT][KR][KHN][VI][VILF]

[f] C-term variations: most of these motifs were compiled for this work.

[g] ER retrieval motif found in PROSITE: [KHRQSA][DENQ]EL [31].

[32]. For the Golgi apparatus, other than C-terminal YQRL also used by PSORT, there is currently no other specific sequence motif available for automatic database searches [19].

Validating motifs against databases

For each motif found (table 1), we retrieved all proteins with this motif deposited in Swiss-Prot [37], TrEMBL [37], and PDB [66]. Next, we extracted a subset of proteins annotated in Swiss-Prot by their experimentally known sub-cellular localization. This subset along with a grouping of all Swiss-Prot species into eukaryotes and non-eukaryotes provided two means of assessing the specificity/accuracy of a given motif. The most specific ER motifs were KDEL and HDEL when restricted to the carboxy terminus (table 1). These two retrieved 131 proteins from Swiss-Prot, most of which have already been experimentally characterized as 'retained in the ER' (data not shown). While the KDEL motif was also present in a few non-eukaryotic proteins, the HDEL motif was found in only two eukaryotic non-ER proteins which were orthologues for the protein 'Protein Kinase C Substrate' in bovine and human (g19p_human and g19p_bovin). While this identification of ER and Golgi localization from such motifs clearly seems very reliable, this finding illustrated the other problem of these two motifs: they occur frequently in non-ER proteins at positions other than the C termini. In other words, in order to rely on KDEL/HDEL to infer localization, we must know the C terminus of the full-length protein. All other ER motifs published were either very unspecific (found in many non-ER proteins) or far too specific (found in very few ER protein families), or both. For example, the di-lysine (KKxx and KxKxx) motif retrieved all known ER proteins when located at the C-terminal position of membrane proteins; however, this included a set of 128 proteins (KKxx) and 97 proteins (KxKxx), several of which could not be classified as ER proteins (table 1). When including this motif [and the more difficult to distinguish di-arginine (xxRR) N-terminal motif] among a non-membrane subset of proteins and more significantly when not limiting the motif to the terminal ends, this high sensitivity is greatly compromised at the cost of an extremely low specificity: both motifs were found in most non-ER proteins. In fact, over 80% of the matches were wrong. Overall, the information contained in the published Golgi motifs was even less promising. For example, the most sensitive GRIP-motif [62, 63] was found in 11 proteins which were mainly orthologues of each other. A generalized GRIP-motif matched in slightly more proteins, none from the Golgi, and many from non-eukaryotic proteins. Similarly, Yxxh, matched in most known Golgi proteins, also matched almost the entire Swiss-Prot database. Obviously, only C-terminal motifs KDEL, HDEL, and AIAKE suffice to accurately annotate ER proteins. All other experimentally characterized retention and recycle motifs for ER and Golgi need to be combined with other means of annotation.

ER and Golgi localization conserved at high levels of sequence similarity

We explored the power of using sequence similarity for the entire proteins to identify ER and Golgi proteins. Toward this end we had (i) to establish thresholds for sequence similarity that enable accurate inference by homology and (ii) to build family profiles of known ER/Golgi proteins. The final 'prediction step' requires searching with a query protein of unknown localization against these family profiles. We could have simplified this final step by aligning all query proteins against the known ER/Golgi proteins. However, sequence-profile alignments are more sensitive and more specific than sequence-sequence alignments. Note that we looked for similarities over the entire proteins, rather than for similarities between short signal peptides [67].

ER and Golgi proteins correctly detected by homology at high levels of similarity

We aligned all experimentally known ER and Golgi (true positives) and all known non-ER and non-Golgi proteins (true negatives) by pairwise BLAST [38, 68] and by the more powerful PSI-BLAST [69]; alignments were ranked by expectation values [68] (EVAL), percentage pairwise sequence identity (PIDE), and the HSSP-value (HVAL; Eq. 1). At HVAL = 0, the accuracy for homology inference was 65% (fig. 1, top); it increased to 98% at HVAL > 40. The majority of false positives (non-ER proteins) at high HSSP-values were of two specific types: heat shock protein 70 and elongation factor alpha. These two large families are not exclusive to the ER but are also abundant in other cellular compartments. They caused the transition between the regions of 'mostly incorrect inference' (HVAL < 20) and 'mostly correct inference' (HVAL > 40) to be more gradual for ER than for Golgi proteins. The accuracy for Golgi proteins (fig. 1, bottom) was slightly higher than that for the ER proteins: 98% accuracy was reached at HVAL > 20. We also investigated the effect of database bias [42], confirming that biased data sets – incorrectly – suggested much higher levels of accuracy at all thresholds (data not shown). At high levels of accuracy, the coverage versus accuracy curve was slightly higher for HSSP-values than for expectation values (data not shown). Thus, we relied on the HSSP-value for the annotations of entire proteomes.

Detailed distinction of ER and Golgi proteins

We also collected data sets for more specific subsets of ER proteins: (i) lumenal, (ii) proteins containing only the

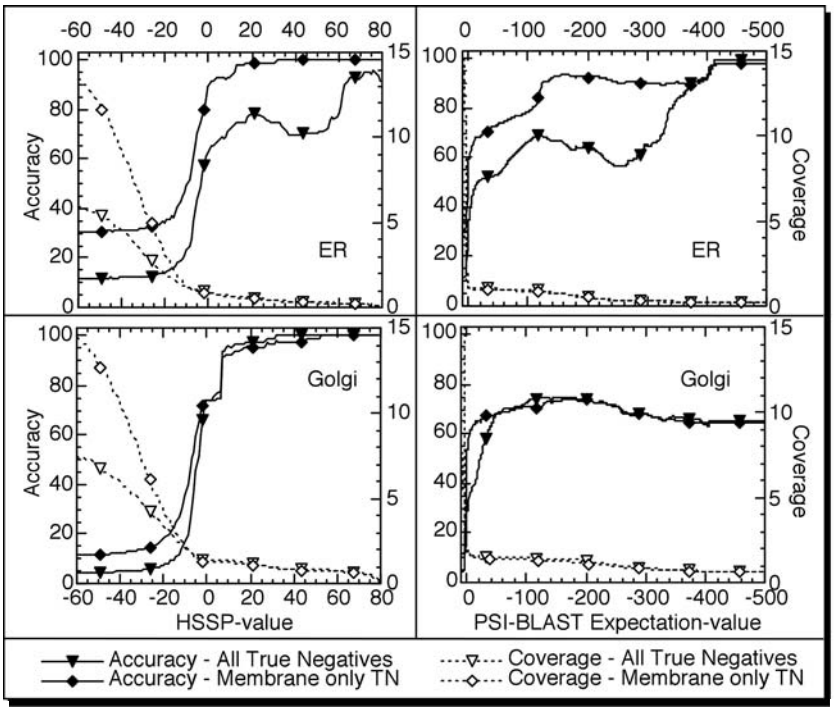

Figure 2. Sequence conservation for ER and Golgi membrane proteins. We aligned all sequence-unique ER (top graphs) and Golgi (lower graphs) membrane proteins against all non-ER/non-Golgi proteins (triangles) and against all non-ER/non-Golgi membrane proteins (diamonds). Solid lines with filled symbols describe cumulative accuracy/specificity (Eq. 2); dotted lines with open symbols describe cumulative coverage/selectivity (Eq. 3). We measured sequence similarity by the HSSP-value (Eq. 1, left graphs) and by the logarithm of the BLAST E-values (right graphs).

[KH]DEL motif (see below), (iii) proteins containing the Swiss-Prot annotation 'PREVENT SECRETION FROM ER,' and (iv) membrane proteins. For the first three subsets of ER proteins, each sequence-unique set contained very few proteins: 21 of 131 total for the lumenal set, 14 of 102 total for those with a [KH]DEL motif, and 30 of 212 total for proteins with a Swiss-Prot ER retention annotation. While these sets were too specific and too small to establish reliable conservation thresholds, the detailed distinction of ER/Golgi subtypes could be used to annotate proteomes. Although the sequence-unique sets of ER and Golgi membrane proteins were also rather small, we could analyze the sequence conservation for these subsets. We compared two different sets of true negatives: (i) all non-ER/non-Golgi proteins and (ii) only non-ER/non-Golgi membrane proteins. Not surprisingly, inference by homology was more accurate when using the additional constraint that the protein had to be in the membrane (fig. 2, diamonds vs triangles). Due to the small number of proteins in the set, the higher levels of accuracy may not hold in general. However, the data certainly supported the assumption that homology inference for ER/Golgi membrane proteins is at least as accurate as that for all other ER/Golgi proteins. We applied this result to searching ER/Golgi membrane proteins in proteomes.

Annotating ER and Golgi proteins in six eukaryotic proteomes

Identifying ER and Golgi proteins in six proteomes

We aimed at annotating as many ER/Golgi proteins as possible through homology and retention and recycle signals in six entirely sequenced eukaryotes (yeast, plant, worm, fly, mouse, and human). Swiss-Prot currently annotates 257 ER and 204 Golgi proteins in these six proteomes (table 2, column labelled 'ER(Golgi)-trusted'). Alignments using the trusted data sets added 718 potential ER and 800 potential Golgi proteins (table 2, rows labeled by an HVAL corresponding to 98% accuracy). Forty-one of these proteins were previously annotated as 'Hypothetical protein.' Swiss-Prot also contains annotations for localization based on sequence similarity to proteins of experimentally known localization. To establish how many of the proteins identified by our homology inference were also annotated by Swiss-Prot, we identified the closest Swiss-Prot homologue for each protein in any of the six proteomes from the PEP database (Predictions for Entire Proteomes) [70]. This revealed that most of our annotations corresponding to 98% were also annotated by Swiss-Prot as either 'probable,' 'putative,' or 'by similarity.' In contrast, most putative annotations according to

Table 2. ER and Golgi proteins in eukaryotic proteomes.

Proteome	HVAL (%)	Total	ER(Golgi)-trusted	Annotated-ER(Golgi)/ annotated-other		Hypothetical
A. Endoplasmic reticulum						
Saccharomyces cerevisiae (yeast)	45 (98)	53	51	51	2	0
	5 (75)	149		64	85	14
Arabidopsis thaliana (plant)	45 (98)	38	9	22	16	0
	5 (75)	570		126	444	9
Caenorhabditis elegans (worm)	45 (98)	12	5	9	3	1
	5 (75)	394		96	298	138
Drosophila melanogaster (fruit-fly)	45 (98)	17	8	14	3	0
	5 (75)	367		169	198	2
Mus musculus (mouse)	45 (98)	289	82	269	20	0
	5 (75)	860		412	448	5
Homo sapiens (human)	45 (98)	309	102	274	35	0
	5 (75)	964		426	538	8
All 6 proteomes	45 (98)	718	257	639	79	1
	38 (95)	830		686	144	3
	27 (90)	1098		795	303	14
	17 (85)	1528		930	598	44
	10 (80)	2328		1151	1177	123
	5 (75)	3304		1293	2011	176
B. Golgi apparatus						
Saccharomyces cerevisiae (yeast)	23 (98)	70	52	53	17	8
	7 (75)	119		55	64	10
Arabidopsis thaliana (plant)	23 (98)	70	10	31	39	5
	7 (75)	260		80	180	15
Caenorhabditis elegans (worm)	23 (98)	57	7	23	34	27
	7 (75)	185		52	133	96
Drosophila melanogaster (fruit-fly)	23 (98)	61	9	40	21	0
	7 (75)	185		76	109	3
Mus musculus (mouse)	23 (98)	195	47	145	50	0
	7 (75)	425		206	219	1
Homo sapiens (human)	23 (98)	347	79	273	74	0
	7 (75)	679		357	322	9
All 6 proteomes	23 (98)	800	204	565	235	40
	16 (95)	1110		675	435	66
	12 (90)	1358		728	630	99
	8 (85)	1726		812	914	125
	7 (78)	1853		826	1027	134

The first column identifies each eukaryotic proteome examined for ER and Golgi proteins. The second column HVAL, i. e., the HSSP-value threshold (Eq. 1), marks the threshold for sequence similarity; the corresponding estimated percentage accuracy for inference by homology at this threshold (fig. 1) is given in parentheses; note that the 98% thresholds differ between ER (HVAL > 45) and Golgi proteins (HVAL > 23). All following columns give the number of proteins identified at the given thresholds for similarity. Total gives the total number of proteins from each proteome found. ER(Golgi)-trusted shows the number of proteins with experimental annotations in our trusted data sets of ER and Golgi. Annotated-ER(Golgi) gives the number of proteins for which the closest Swiss-Prot homologue (taken from the PEP database [70]) has ER(Golgi) annotations marked as PROBABLE, PUTATIVE or BY SIMILARITY. Annotated-other gives the number of proteins for which the closest Swiss-Prot homologue – irrespectiv of the similarity threshold – is not annotated as either ER or Golgi. Hypothetical lists the number of proteins for which the only previous annotation was 'Hypothetical protein.' The entire set of results is publicly available at http://cubic.bioc.columbia.edu/ERGolgiDB/.

sequence similarity thresholds that correspond to 75% accuracy are not annotated as ER/Golgi by Swiss-Prot. At this threshold, we could propose another 3304 possible ER and 1853 possible Golgi proteins (table 2, rows labeled by '75%' and '78%' for ER and Golgi, respectively). While we expect that a majority of these annotations are likely to be false, these subsets constitute a good 'hunting ground' for discovery of uncharacterized ER and Golgi proteins. Overall, each experimental annotation in our trusted set yielded about one to three (lower value for ER, higher for Golgi) homology transfers as high accuracy (98%) and about five to six at low accuracy (> 75%). The entire set of results is publicly available at http://cubic.bioc.columbia.edu.

Identifying ER and Golgi membrane proteins

We found a total of 3941 putative ER and Golgi membrane proteins in the six proteomes at a threshold corresponding to 75% accuracy. In most proteomes we could expand reliable annotations (98% accuracy threshold) for ER membrane proteins between 2.5-fold (human) and 8-fold (plant). At the same accuracy threshold, we also identified ER membrane proteins in worm for which our initial trusted set contained no ER membrane proteins (fig. 3). Homology inference allowed annotating between 190 (98 % accuracy) and 759 (75 % accuracy) Golgi membrane proteins (fig. 3). We also identified 155 possible lumenal ER proteins at 75 % accuracy (data not shown) based solely on using the much smaller but less reliable motif-only data sets. Identifying lumenal ER proteins is particularly relevant as 82 % of the current 257 experimentally annotated ER proteins within our dataset are membrane associated.

Closer inspection of a few examples

Many close homologues of ER and Golgi proteins also contained retention motifs, thereby increasing the reliability of the inference (table 3). Specifically, the ER-lumenal protein cyclophilin-B (Swiss-Prot identifier cypb_bovin, table 3) from the cyclophilin-type peptidyl prolyl

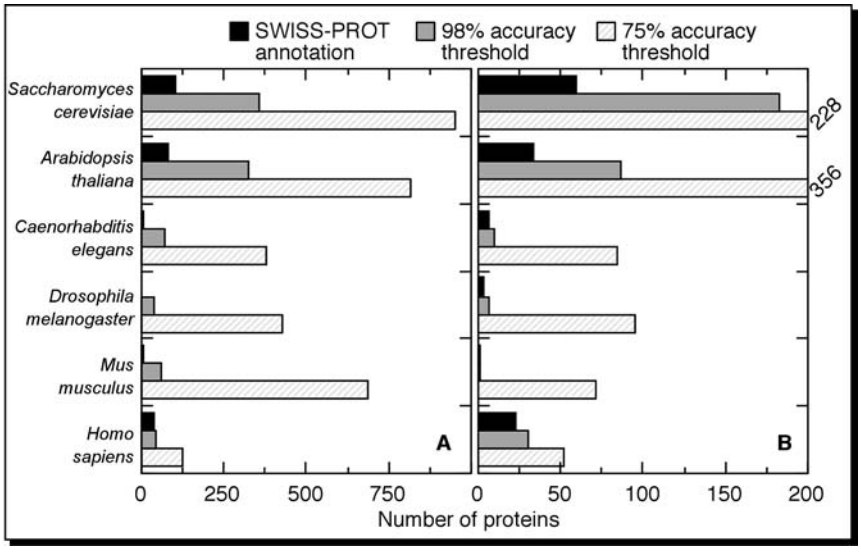

Figure 3. Identification of ER (A) and Golgi (B) membrane proteins in six eukaryotic proteomes**.** Gray bars represent total counts of proteins within each set of proteomes. The first bar (darkest of each color) represents the total number of annotated membrane proteins found in Swiss-Prot for each proteome. The next two bars (lighter in color) represent the number of additional membrane proteins found within threshold levels of 98% accuracy and 75% accuracy, respectively.

Table 3. ER and Golgi homologues with differently annotated localization.

True positive	Homologue	PIDE	HVAL	EVAL	Localization
ER					
hs7c caeel	*hs7f caeel*	52	25	–147	mitochondria
aca2 arath	*aca1 arath*	75	52	0	chloroplast
cypb bovin	*cyp4 bovin*	48	28	–47	cytoplasm
dha4 rat	*dhac rat*	26	5	–33	cytoplasm
scj1 yeast	*sis1 yeast*	44	3	–17	nucleus
calu mouse	*cb45 mouse*	27	3	–32	golgi
bip5 tobac	*gr75 mouse*	52	28	–157	mitochondria
fd61 soybn	*fd6c soybn*	25	–1	–23	chloroplast
Golgi					
maba rat	*pspd rat*	35	8	–32	extracellular
rb6b human	*ypt7 yeast*	35	7	–31	vacuole
ynd1 yeast	*ntpa pea*	26	3	–28	nucleus
arse human	*ids human*	35	0	–17	lysosome
cb45 mouse	*rcn1 mouse*	26	3	–30	ER

* ER/Golgi and their close homologues are presented using the Swiss-Prot identifier; bold identifiers mark proteins containing ER retrieval signals. *PIDE* is residue percent identity, *HVAL* is the HSSP-value (Eq. 1), and *EVAL* is the PSI-BLAST expectation value rounded to the closest exponent as measured between the two pairs of proteins.

cis-trans isomerase family is sequence similar to proteins of that family that are non-ER. However, since it also contains the ER recycle motif, we can correctly pick the annotation of the closest ER family member. Similarly, the heat shock proteins hs7c_caeel and hs7f_caeel differ mainly at their ends; hs7c_caeel contains a C-terminal KDEL motif and is ER, hs7f_caeel lacks the motif and is mitochondrial. Nevertheless, KDEL-like motifs are accurate for finding more distant homologues in a protein family, for example, the evolutionarily conserved family of boca chaperones [71, 72]. Boca proteins have very little similarity to other known chaperones that reside in the ER (closest homologue from our ER trusted list is protein disulfide isomerase in mouse (Swiss-Prot identifier: pdi_mouse: HVAL to boca = – 14 with only 13% identical residues). This lack of sequence similarity to other chaperones may be crucial for boca to specifically play a role in the quality control and trafficking of only one particular set of proteins, namely the LDL receptors. However, the six proteins resembling boca have variations of the generic and accurate [KH]DEL C-terminal ER recycle motifs that are generally far less accurate, in particular, the C-terminal motifs KKEL, RDEL, and REDL (with an estimated accuracy of 61%, data not shown). REDL and KKEL were not present in any trusted ER protein, with KKEL, in particular, being found in many non-ER proteins, including a ribosomal protein from mitochondrial DNA (Swiss-Prot identifier rt13_acaca). Why would some boca homologues diverge to less specific motifs? The answer remains unclear. Other regulatory mechanisms might prevent mislocalization for these proteins. All these examples illustrate how we can increase the accuracy in annotation by combining sequence similarity and motifs.

Recent large-scale experiment verified our estimates

Obviously, our homology-transfer-based assignments become stronger with increasing experimental data. However, here, we also utilized large-scale experimental data for cross-examination of our estimates. The TRIPLES database annotates results from large-scale experimental annotations of localization for yeast [35, 73]. While TRIPLES distinguishes ER proteins, it does not classify Golgi proteins separately. Overall, TRIPLES annotates 74 ER proteins (and another 60 annotated as ER and something else). Of the proteins that we annotated at high accuracy and that were classified by TRIPLES, only one, ydl093w (Swiss-Prot identifier pmt5_yeast; HVAL = 27 to pmt1_yeast), is not found to be ER by TRIPLES, rather it is annotated as cytoplasmic. Given that ydl093w appears to have globular, non-membrane-associated, regions in the cytoplasm, the TRIPLES result was actually compatible with our annotations. Overall, we identified 46 proteins as ER (at 75% accuracy/HVAL 5) for which TRIPLES had non-ER annotations. Similarly, we examined the yeast GFP fusion localization database (YEAST-GFP) [34] that, in contrast to TRIPLES, also annotates Golgi proteins. Overall, YEAST-GFP classified 295 ER and 144 Golgi proteins (including proteins with experimental annotations for more than one compartment). GFP proteins are usually fused at the C terminus; as many ER-recycle signals are C terminal, GFP fusions may be particularly difficult for ER proteins [34]. Nevertheless, the YEAST-GFP results agree very well with our findings at high accuracy (98%). At that level, we identified 31 ER proteins, and 30 of these were classified by YEAST-GFP, the only exception being yal026c, the membrane calcium-transporting ATPase DRS2 which YEAST-GFP classifies as Golgi (members of this protein family have been identified in both the ER and trans-Golgi network [74, 75]). (Note: yal026c is classified as cytoplasmic by TRIPLES.) At the same level of accuracy (98%), we annotated an additional 23 ER proteins that were not classified by YEAST-GFP. At 98% accuracy, we annotated 19 proteins as ER that overlapped with YEAST-GFP; 13 of these were also classified as Golgi by YEAST-GFP, while the other 6 were classified as ER. Looking in more detail, we found annotations that may indicate a dual localization for five of the six (ydr498c-sc20_yeast, ynr026c-sc 12_yeast, ycr067c-sed4_yeast, ygl054c-erv4_yeast, yg l145w-tp20_yeast; note: given are the genome identifiers and the identifiers of the corresponding annotated Swiss-Prot files). At 98% accuracy, we annotated an additional 54 Golgi proteins that were not classified by YEAST-GFP. Thus, the comparison to the latest large-scale experimental data suggested that overall our estimates were fairly accurate, and that even almost complete experimental coverage still requires homology-transfer for a considerable number of proteins.

Latest data surprisingly specific to yeast

The large-scale experimental results in TRIPLES and YEAST-GFP obviously make our efforts to annotate through homology-transfer much more powerful since they increase the number of proteins in the data sets of experimentally trusted proteins. In particular, the recent YEAST-GFP data increased our sets by a total of 108 ER proteins (30 of which corresponded to sequence-unique additions) and 39 (12 unique) Golgi proteins. Surprisingly, these additional annotations did not yield many highly reliable annotations for other proteomes: the 108 ER proteins identified only 12 non-yeast homologues at 98% accuracy (raising the total from 718 to 730) and 554 at 75% (total from 3304 to 3638). The 49 Golgi proteins yielded another 7 non-yeast proteins at 98% (total to 856) and another 187 at 78% (total to 2040). The increases from the YEAST-GFP data at lower accuracy were similar to the yield of our original trusted data, e.g., each trusted ER protein from our original set (676) yielded about four annotations in the proteomes (2949 when sub-

tracting yeast). In contrast, the original data yielded about 2/3 annotations at high accuracy (676 yielded 459 annotations), while the new data gave only 1/10 new annotations (108 yielded 12 annotations). Thus, the new data turned out to be fairly specific to yeast.

Proteins with KDEL and HDEL diverged less than expected

Only 114 proteins in all six eukaryotic proteomes contained the C-terminal ER recycle motifs KDEL or HDEL. Of these 114 proteins, 39 have a close homologue in Swiss-Prot that is experimentally annotated as ER. Assuming that all proteins with C-terminal KDEL or HDEL motifs identified in the eukaryotic proteomes are retained in the ER, we can analyze the extent to which these proteins could have been identified based on homology alone. Somewhat surprisingly, almost half of all these proteins mapped to proteins in our trusted profile families above thresholds corresponding to > 80% accuracy (fig. 4), and about 60% were found in proteins that shared a similar fold with one of the proteins in the trusted families (HVAL = 0, gray line in fig. 4). All of the [KH]DEL proteins mapped to trusted families at HVAL < –10. Although this value was too low to infer similarities, it was significantly higher than the comparable values for all ER proteins (fig. 1). This observation suggested a rather puzzling conclusion: on the one hand there is no good reason to expect that two proteins containing C-terminal [KH]DEL motifs are evolutionarily related. On the other hand, the majority of the proteins with this motif have not diverged very much from the ER proteins we know. In other words, proteins with this motif have diverged, on average, less than proteins with the same nuclear localization signals [28, 30].

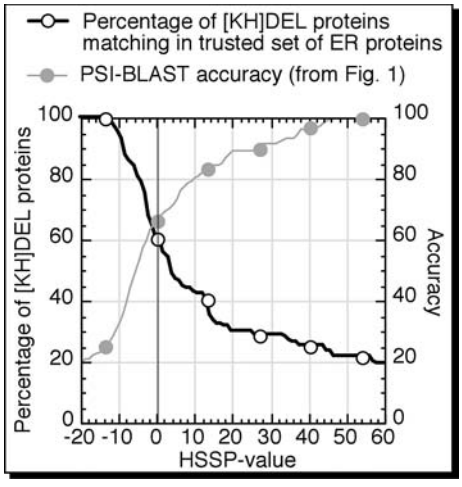

Figure 4. Mapping [KH]DEL proteins onto the trusted set. We retrieved all proteins in six eukaryotic proteomes that contain the motifs KDEL or HDEL and aligned these to all ER proteins in our trusted families. The y-axis gives the percentage of [KH]DEL proteins found above a certain HSSP-value (Eq. 1; for comparison the estimates for accuracy are copied from fig. 1 to the right y-axis). The grey line at an HSSP-value of 0 marks the point above which protein pairs share a similar three-dimensional fold.

Conclusions

Annotating sub-cellular localization is an essential aspect of large-scale experimental endeavors. Most methods that predict localization in the absence of experimental annotations either leave out or predict with limited results ER and Golgi due to lack of experimental data [20, 23, 30, 32, 36, 76–88]. We introduced estimates for the accuracy in inferring ER and Golgi localization based on experimentally characterized short sequence motifs and based on sequence similarity to experimentally characterized proteins (homology-transfer). Our results suggested that most of the few currently known retention and recycle motifs for ER and Golgi proteins are too specific and/or too inaccurate to be used in isolation to automatically annotate entire proteomes. Nevertheless, these motifs might be crucial indicators provided we know their reliability (table 1). Providing such estimates was one of the major tasks that we addressed here. Similarly, we found that even very high levels of sequence similarity might not suffice to infer ER and Golgi localization without errors (table 3). In fact, we confirmed our previous findings [89] that only extremely high levels of pairwise sequence identity (> 80%; fig. 1, left panel) and very low PSI-BLAST E-values (< 10^{-100}; fig. 1, right panel) enabled accurate homology-transfer at low coverage (few true positives identified at threshold). When we also considered alignment length (HVAL, Eq. 1), we could find thresholds for reliable homology-transfer at significantly higher levels of coverage (fig. 1, central panels). Similar observations have been observed for other cellular organelles [36]. We therefore applied these safe HVAL thresholds to the six entirely sequenced eukaryotes (human, mouse, fly, worm, plant, yeast). Knowing what to expect on average at a given threshold for sequence similarity was essential for such large-scale homology-transfer. Combining motifs and sequence similarity yielded the most reliable annotations (fig. 4). Another way to increase the reliability of inferring ER and Golgi proteins was by separating lumenal and membrane proteins (fig. 2). This might be explored in the context of considering other targeting mechanisms for membrane proteins [90, 91]. Our exploration of ER and Golgi resident proteins can also assist large-scale proteomic endeavors which often do not fully encompass the proteome, may be limited to specific organelles, or are unable to distinguish between lumenal and membrane proteins (table 2). Finally, more specific methods based on intrinsic protein structural features such as membrane boundaries [85], post-translational modifications or functional domains [92] may improve ER and Golgi localization prediction techniques to fully encompass all proteins localized within these cellular compartments.

Acknowledgements. Thanks to J. Liu (Columbia) for computer assistance and to R. Nair (Columbia) and D. Przybylski (Columbia) for providing preliminary information, programs and knowledgeable discussions. Thanks to R. Mann (Columbia) for bringing boca to our attention and for helpful discussions on this family. This work was supported by the grant DBI-0131168 from the National Science Foundation (NSF) and by a grant to the Northeast Structural Genomics Consortium from the Protein Structure Initiative of National Institutes of Health (P50 GM62413). Last, but not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

1 Pfeffer S. R. and Rothman J. E. (1987) Biosynthetic protein transport and sorting by the endoplasmic reticulum and golgi. Annu. Rev. Biochem. **56:** 829–852
2 Pemberton L. F., Blobel G. and Rosenblum J. S. (1998) Transport routes through the nuclear pore complex. Curr. Opin. Cell Biol. **10:** 392–399
3 Adam S. A. (1999) Transport pathways of macromolecules between the nucleus and the cytoplasm. Curr. Opin. Cell Biol. **11:** 402–406
4 Chen X. and Schnell D. J. (1999) Protein import into chloroplasts. Trends Cell Biol. **9:** 222–227
5 Hettema E. H., Distel B. and Tabak H. F. (1999) Import of proteins into peroxisomes. Biochim. Biophys. Acta **1451:** 17–34
6 Hood J. K. and Silver P. A. (1999) In or out? Regulating nuclear transport. Curr. Opin. Cell Biol. **11:** 241–247
7 Koehler C. M., Merchant S. and Schatz G. (1999) How membrane proteins travel across the mitochondrial intermembrane space. Trends Biochem. Sci. **24:** 428–432
8 Darnell J., Lodish H. and Baltimore D. (1990) Molecular Cell Biology, 2nd edn., Freeman, New York
9 Von Heijne G. (1985) Signal sequences: the limits of variation. J. Mol. Biol. **184:** 99–105
10 Verner K. and Schatz G. (1988) Protein translocation across membranes. Science **241:** 1307–1313
11 Briggs M. S. and Gierasch L. M. (1986) Molecular mechanisms of protein secretion: the role of the signal sequence. Adv. Protein Chem. **38:** 109–180
12 Sjöström M., Wold S., Wieslander Å. and Rilfors L. (1987) Signal peptide amino acid sequences in *Escherichia coli* contain information related to final protein localization. EMBO J. **6:** 823–831
13 Nielsen H., Engelbrecht J., Brunak S. and Von Heijne G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. **10:** 1–6
14 Pelham H. R. (1990) The retention signal for soluble proteins of the endoplasmic reticulum. Trends Biochem. Sci. **15:** 483–486
15 Gleeson P. A. (1998) Targeting of proteins to the golgi apparatus. Histochem. Cell Biol. **109:** 517–532
16 Nickel W. (2003) The mystery of nonclassical protein secretion: a current view on cargo proteins and potential export routes. Eur. J. Biochem. **270:** 2109–2119
17 Teasdale R. D. and Jackson M. R. (1996) Signal-mediated sorting of membrane proteins between the endoplasmic reticulum and the golgi apparatus. Annu. Rev. Cell Dev. Biol. **12:** 27–54
18 Mellman I. and Warren G. (2000) The road taken: past and future foundations of membrane traffic. Cell **100:** 99–112
19 Saint-Jore-Dupas C., Gomord V. and Paris N. (2004) Protein localization in the plant golgi apparatus and the trans-golgi network. Cell. Mol. Life Sci. **61:** 159–171
20 Nielsen H., Brunak S. and Von Heijne G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng. **12:** 3–9
21 Ghosh P., Dahms N. M. and Kornfeld S. (2003) Mannose 6-phosphate receptors: new twists in the tale. Nat. Rev. Mol. Cell Biol. **4:** 202–212
22 Traub L. M. and Kornfeld S. (1997) The trans-golgi network: a late secretory sorting station. Curr. Opin. Cell Biol. **9:** 527–533
23 Emanuelsson O., Nielsen H., Brunak S. and Von Heijne G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. **300:** 1005–1016
24 Emanuelsson O., Nielsen H. and Von Heijne G. (1999) Chlorop, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. **8:** 978– 984
25 Emanuelsson O., Elofsson A., Von Heijne G. and Cristobal S. (2003) In silico prediction of the peroxisomal proteome in fungi, plants and animals. J. Mol. Biol. **330:** 443–456
26 Claros M. G., Brunak S. and Von Heijne G. (1997) Prediction of n-terminal protein sorting signals. Curr. Opin. Struct. Biol. **7:** 394–398
27 Nair R., Carter P. and Rost B. (2003) NLSDB: database of nuclear localization signals. Nucleic Acids Res. **31:** 397–399
28 Nair R. and Rost B. (2003) Loc3d: annotate sub-cellular localization for protein structures. Nucleic Acids Res. **31:** 3337–3340
29 La Cour T., Gupta R., Rapacki K., Skriver K., Poulsen F. M. and Brunak S. (2003) Nesbase version 1.0: a database of nuclear export signals. Nucleic Acids Res. **31:** 393–396
30 Cokol M., Nair R. and Rost B. (2000) Finding nuclear localization signals. EMBO Rep. **1:** 411–415
31 Bucher P. and Bairoch A. (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. Proc. Int. Conf. Intell. Syst. Mol. Biol. **2:** 53–61
32 Nakai K. and Horton P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem. Sci. **24:** 34–36
33 Kimata Y., Ooboki K., Nomura-Furuwatari C., Hosoda A., Tsuru A. and Kohno K. (2000) Identification of a novel mammalian endoplasmic reticulum-resident KDEL protein using an EST database motif search. Gene **261:** 321–327
34 Huh W. K., Falvo J. V., Gerke L. C., Carroll A. S., Howson R. W., Weissman J. S. et al. (2003) Global analysis of protein localization in budding yeast. Nature **425:** 686–691
35 Kumar A., Cheung K. H., Tosches N., Masiar P., Liu Y., Miller P. et al. (2002) The triples database: a community resource for yeast molecular biology. Nucleic Acids Res. **30:** 73–75
36 Nair R. and Rost B. (2002) Sequence conserved for subcellular localization. Protein Sci. **11:** 2836–2847
37 Bairoch A. and Apweiler R. (2000) The swiss-prot protein sequence database and its supplement TREMBL in 2000. Nucleic Acids Res. **28:** 45–48
38 Altschul S. F. and Gish W. (1996) Local alignment statistics. Methods Enzymol. **266:** 460–480
39 Przybylski D. and Rost B. (2002) Alignments grow, secondary structure prediction improves. Proteins **46:** 195–205
40 Sander C. and Schneider R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. Proteins **9:** 56–68
41 Rost B. (1999) Twilight zone of protein sequence alignments. Protein Eng. **12:** 85–94
42 Rost B. (2002) Enzyme function less conserved than anticipated. J. Mol. Biol. **318:** 595–608
43 Wrzeszczynski K. O. and Rost B. (2004) Cataloging proteins in cell cycle control. Methods Mol. Biol. **241:** 219–233
44 Mika S. and Rost B. (2003) Uniqueprot: Creating representative protein sequence sets. Nucl. Acids Res. **31:** 3789–3791
45 Hofmann K., Bucher P., Falquet L. and Bairoch A. (1999) The PROSITE database, its status in 1999. Nucleic Acids Res. **27:** 215–219
46 Jackson M. R., Nilsson T. and Peterson P. A. (1990) Identification of a consensus motif for retention of transmembrane proteins in the endoplasmic reticulum. EMBO J. **9:** 3153–3162
47 Itin C., Kappeler F., Linstedt A. D. and Hauri H. P. (1995) A novel endocytosis signal related to the KKXX ER-retrieval signal. EMBO J. **14:** 2250–2256
48 Schutze M. P., Peterson P. A. and Jackson M. R. (1994) An N-terminal double-arginine motif maintains type II membrane

proteins in the endoplasmic reticulum. EMBO J. **13:** 1696–1705
49 Itin C., Schindler R. and Hauri H. P. (1995) Targeting of protein ergic-53 to the ER/ergic/cis-golgi recycling pathway. J. Cell Biol. **131:** 57–67
50 Andersson H., Kappeler F. and Hauri H. P. (1999) Protein targeting to endoplasmic reticulum by dilysine signals involves direct retention in addition to retrieval. J. Biol. Chem. **274:** 15080–15084
51 Dogic D., Dubois A., De Chassey B., Lefkir Y. and Letourneur F. (2001) Ergic-53 KKAA signal mediates endoplasmic reticulum retrieval in yeast. Eur. J. Cell Biol. **80:** 151–155
52 Yabe D., Nakamura T., Kanazawa N., Tashiro K. and Honjo T. (1997) Calumenin, a Ca^{2+}-binding protein retained in the endoplasmic reticulum with a novel carboxyl-terminal sequence, HDEF. J. Biol. Chem. **272:** 18232–18239
53 Arber S., Krause K. H. and Caroni P. (1992) S-cyclophilin is retained intracellularly via a unique COOH-terminal sequence and colocalizes with the calcium storage protein calreticulin. J. Cell Biol. **116:** 113–125
54 Boyd G. W., Doward A. I., Kirkness E. F., Millar N. S. and Connolly C. N. (2003) Cell surface expression of 5-ht3 receptors is controlled by an endoplasmic reticulum retention signal. J. Biol. Chem. **278:** 27681–27687
55 Humphrey J. S., Peters P. J., Yuan L. C. and Bonifacino J. S. (1993) Localization of TGN38 to the trans-golgi network: Involvement of a cytoplasmic tyrosine-containing sequence. J. Cell Biol. **120:** 1123–1135
56 Voorhees P., Deignan E., Van Donselaar E., Humphrey J., Marks M. S., Peters P. J. et al. (1995) An acidic sequence within the cytoplasmic domain of furin functions as a determinant of trans-golgi network localization and internalization from the cell surface. EMBO J. **14:** 4961–4975
57 Rohrer J. and Kornfeld R. (2001) Lysosomal hydrolase mannose 6-phosphate uncovering enzyme resides in the trans-golgi network. Mol. Biol. Cell **12:** 1623–1631
58 Bos K., Wraight C. and Stanley K. K. (1993) TGN38 is maintained in the trans-golgi network by a tyrosine-containing motif in the cytoplasmic domain. EMBO J. **12:** 2219–2228
59 Nothwehr S. F., Roberts C. J. and Stevens T. H. (1993) Membrane protein retention in the yeast golgi apparatus: dipeptidyl aminopeptidase A is retained by a cytoplasmic signal containing aromatic residues. J. Cell Biol. **121:** 1197–1209
60 Ugur O. and Jones T. L. (2000) A proline-rich region and nearby cysteine residues target xlalphas to the golgi complex region. Mol. Biol. Cell **11:** 1421–1432
61 Bannykh S. I., Nishimura N. and Balch W. E. (1998) Getting into the golgi. Trends Cell Biol. **8:** 21–25
62 Kjer-Nielsen L., Teasdale R. D., Van Vliet C. and Gleeson P. A. (1999) A novel golgi-localisation domain shared by a class of coiled-coil peripheral membrane proteins. Curr. Biol. **9:** 385–388
63 Munro S. and Nichols B. J. (1999) The grip domain – a novel golgi-targeting domain found in several coiled-coil proteins. Curr. Biol. **9:** 377–380
64 Zerangue N., Schwappach B., Jan Y. N. and Jan L. Y. (1999) A new ER trafficking signal regulates the subunit stoichiomtry of plasma membrane K(ATP) channels. Neuron **22:** 537–548
65 Kirchhausen T. (2002) Clathrin adaptors really adapt. Cell **109:** 413–416
66 Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H. et al. (2000) The protein data bank. Nucleic Acids Res. **28:** 235–242
67 Nielsen H., Engelbrecht J., Von Heijne G. and Brunak S. (1996) Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. Proteins **24:** 165–177
68 Altschul S. F. (1993) A protein alignment scoring system sensitive at all evolutionary distances. J. Mol. Evol. **36:** 290–300
69 Altschul S., Madden T., Shaffer A., Zhang J., Zhang Z., Miller W. et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402
70 Carter P., Liu J. and Rost B. (2003) Pep: predictions for entire proteomes. Nucleic Acids Res. **31:** 410–413
71 Culi J. and Mann R. S. (2003) Boca, an endoplasmic reticulum protein required for wingless signaling and trafficking of IDL receptor family members in *Drosophila*. Cell **112:** 343–354
72 Herz J. and Marschang P. (2003) Coaxing the IDL receptor family into the fold. Cell **112:** 289–292
73 Ross-Macdonald P., Coelho P. S., Roemer T., Agarwal S., Kumar A., Jansen R. et al. (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature **402:** 413–418
74 Hua Z. and Graham T. R. (2003) Requirement for neo1p in retrograde transport from the golgi complex to the endoplasmic reticulum. Mol. Biol. Cell **14:** 4971–4983
75 Hua Z., Fatheddin P. and Graham T. R. (2002) An essential subfamily of drs2p-related p-type ATPases is required for protein trafficking between golgi complex and endosomal/vacuolar system. Mol. Biol. Cell **13:** 3162–3177
76 Andrade M. A., O'donoghue S. I. and Rost B. (1998) Adaptation of protein surfaces to subcellular location. J. Mol. Biol. **276:** 517–525
77 Chou K. C. and Cai Y. D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem. **277:** 45765–45769
78 Chou K. C. and Elrod D. W. (1999) Protein subcellular location prediction. Protein Eng. **12:** 107–118
79 Emanuelsson O. (2002) Predicting protein subcellular localisation from amino acid sequence information. Brief Bioinform. **3:** 361–376
80 Fujiwara Y. and Asogawa M. (2001) Prediction of subcellular localizations using amino acid composition and order. Genome Inform. Ser. Workshop Genome Inform. **12:** 103–112
81 Nair R. and Rost B. (2002) Inferring sub-cellular localization through automated lexical analysis. Bioinformatics **18 (suppl 1):** S78–S86
82 Nakai K. (2000) Protein sorting signals and prediction of subcellular localization. Adv. Protein Chem. **54:** 277–344
83 Nakashima H. and Nishikawa K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol. **238:** 54–61
84 Zhou G. P. and Doctor K. (2003) Subcellular location prediction of apoptosis proteins. Proteins **50:** 44–48
85 Yuan Z. and Teasdale R. D. (2002) Prediction of golgi type II membrane proteins based on their transmembrane domains. Bioinformatics **18:** 1109–1115
86 Reinhardt A. and Hubbard T. (1998) Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. **26:** 2230–2236
87 Mott R., Schultz J., Bork P. and Ponting C. P. (2002) Predicting protein cellular localization using a domain projection method. Genome Res. **12:** 1168–1174
88 Nair R. and Rost B. (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. Proteins **53:** 917–930
89 Rost B., Liu J., Nair R., Wrzeszczynski K. O. and Ofran Y. (2003) Automatic prediction of protein function. Cell. Mol. Life Sci. **60:** 2637–2650
90 Rayner J. C. and Pelham H. R. (1997) Transmembrane domain-dependent sorting of proteins to the ER and plasma membrane in yeast. EMBO J. **16:** 1832–1841
91 Wattenberg B. and Lithgow T. (2001) Targeting of C-terminal (tail)-anchored proteins: understanding how cytoplasmic activities are anchored to intracellular membranes. Traffic **2:** 66–71
92 Ponting C. P. (2000) Proteins of the endoplasmic-reticulum-associated degradation pathway: domain detection and function prediction. Biochem. J. **351:** 527–535