BIOINFORMATICS

Vol. 20 no. 6 2004, pages 937–944
DOI: 10.1093/bioinformatics/bth010

The Hera database and its use in the characterization of endoplasmic reticulum proteins

M. Scott[1,3], G. Lu[2], M. Hallett[1] and D. Y. Thomas[3,*]

[1]McGill Center for Bioinformatics, Duff Medical Building, McGill University, 3755 University Street, Montreal, Quebec, Canada H3A 2B4, [2]Center for Biotechnology and School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, NE 68588, USA and [3]Biochemistry Department, Faculty of Medicine, McGill University, McIntyre Medical Sciences Building, 3655 Promenade Sir William Osler, Montreal, Quebec, Canada H3G 1Y6

Received on September 11, 2003; accepted on October 16, 2003
Advance Access publication January 29, 2004

ABSTRACT

Motivation: Information concerning endoplasmic reticulum (ER) proteins is widely dispersed and cannot be easily and rapidly processed by the biological community. We present a comprehensive database of human ER proteins, called Human ER Aperçu (Hera). The Hera database was constructed by exhaustively searching through public databases and the scientific literature for ER proteins.

Results: Hera was used for the analysis of characteristics common to all human ER proteins. Our results show that a high proportion of ER proteins (59%) have at least one transmembrane domain and display physical characteristics consistent with this observation. In addition, one-third of ER proteins contain known ER retrieval or retention signals and 70% of ER proteins contain a signal peptide or anchor. Finally, 85% of ER proteins contain at least one InterPro motif. The most abundant InterPro motifs in ER proteins represent many of the most well-characterized functions of the ER.

Availability: Hera is available at http://www.mcb.mcgill.ca/∼hera.

Contact: david.thomas@mcgill.ca

INTRODUCTION

The endoplasmic reticulum (ER) is an intracellular compartment in eukaryotic cells that harbors a wide variety of cellular activities. Nascent polypeptides destined for the secretory pathway are synthesized on membrane-bound ribosomes and translocate into the ER where they fold under the supervision of chaperone complexes (reviewed in Pelletier *et al*., 2001). The ER quality control process ensures proteins destined for all compartments of the secretory pathway as well as proteins marked for secretion or the cell surface can only leave the ER when they are correctly folded. The ER is also involved in other cellular activities, such as lipid biosynthesis, detoxification, calcium storage, calcium signaling and many aspects of cellular stress response. Dysfunction of processes in the ER has been shown to be involved in many human diseases including neurological and aging diseases, such as ischemia, epileptic seizures, Alzheimer's, Parkinson's (reviewed in Paschen and Frandsen, 2001) as well as cystic fibrosis, cancer and juvenile pulmonary emphysema (Lee, 2001).

A comprehensive description of the ER requires a catalog of resident proteins, their interactions and the pathways in which they are involved. In addition to experimental approaches, informatics provides important tools for defining ER proteins and their functions. However, recent information regarding ER proteins is dispersed; a small proportion of this data lies in various public and private databases but the majority of the data remains 'buried' in research papers in the literature. This makes comprehensive analyses of the ER proteome difficult. We report here the creation of a publicly available database of known and potential human ER proteins and their preliminary analysis via several bioinformatics approaches.

IMPLEMENTATION AND RESULTS

The Hera database: structure

The Human ER Aperçu (Hera) database is a publicly available catalog of known human ER proteins. The entity–relationship model of Hera (Fig. 1) is implemented in the Open Source relational database MySQL and the database can be queried via the scripting language PHP. The central table of Hera is the Protein table. Additional information and characterizations of each protein are encoded in several other tables and related to the Protein table in a standard manner. Hera currently holds

*To whom correspondence should be addressed.

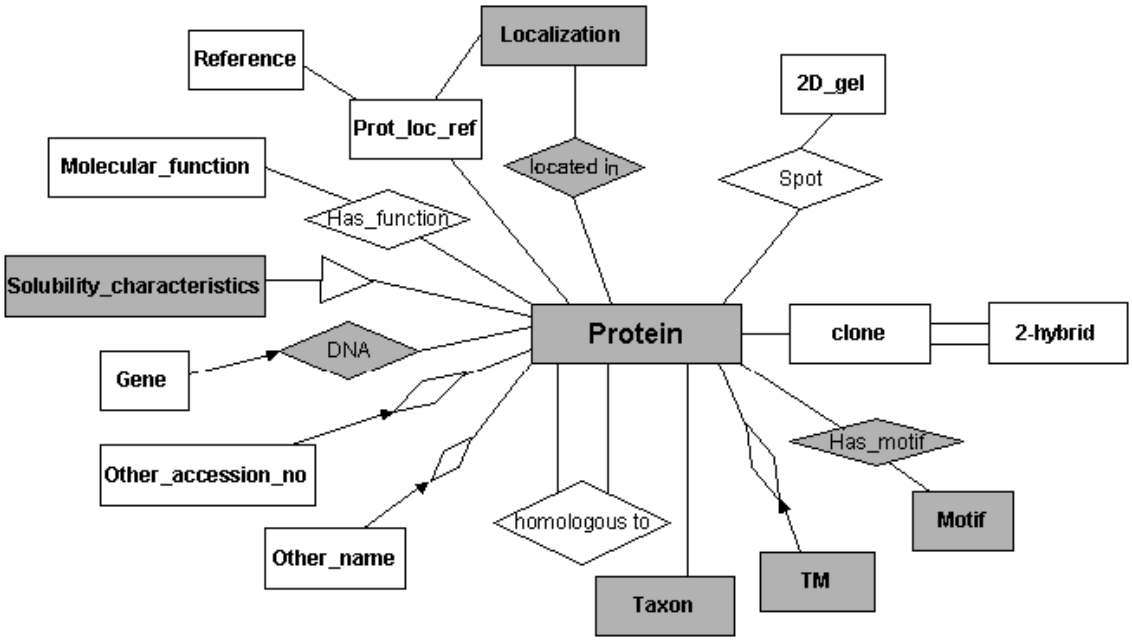

Fig. 1. Entity–relationship diagram of the Hera database. The Protein table is central to this database. It stores fundamental protein information including the amino acid sequence, physical characteristics, taxon, name and refseq accession number for the NCBI sequence repository. The Protein table is related to most other tables of the database, which further characterize the proteins. The tables colored in gray currently house information regarding all 499 ER proteins. The other tables contain incomplete information.

information relating to 499 known or potential human ER proteins, including protein name(s), sequence and several physical properties. This includes the number and location of transmembrane domains and the presence of signal peptides. The tables that currently hold information regarding all ER proteins in the database are colored in gray in Figure 1. At this time, Hera can be searched for information by protein name, NCBI refseq identification number or can be browsed by ER targeting sequences and protein characteristics. Links to other public databases including NCBI, InterPro, GeneCards and SwissProt are displayed when applicable. We encourage feedback regarding any entry in the Hera database as well as any type of information concerning ER proteins not present in the database. Accordingly, we have set up a user feedback/contribution interface accessible through the Hera web page. As the Hera project develops, it is expected that information regarding protein function, post-translational modifications, homology and orthology as well as protein–protein interactions (detected by, e.g. on going yeast two-hybrid experiments) will also be added to the database.

The Hera database: content

Data relating to human ER proteins were collected from diverse information sources by entering the keywords ‘ER’ and ‘endoplasmic reticulum’. Each entry was then manually screened to verify the localization evidence and to determine the classification criteria of the entry. When the information source was the scientific literature, each article was read to evaluate the evidence presented. These data sources and their associated criteria or type of evidence for the localization of these proteins to the ER are shown in Table 1. This information is important since the proteins present in Hera have been annotated as ER proteins according to vastly different criteria,

Table 1. Sources and types of protein localization classification

Source	Types of classification labels[a]
NCBI, GenBank	k, t, c, h
SwissProt (Boeckmann *et al*., 2003)	c, s, p
PubMed and scientific literature	e
GeneCards v2.26 (Safran *et al*., 2002)	c, s, k
PIR (Wu *et al*., 2002)	c
Proteome BioKnowledge® Library[b]	c

[a]c, classified as ER with high degree of certainty by the source; e, experimentally determined to be localized in the ER; h, high-sequence similarity with proteins known to be resident ER proteins in other species; k, the classification source contains the keyword ER in the description of the protein but it is unclear whether the protein is an ER resident or just passes through the ER; p, possible ER protein; s, classified ER by similarity by the source; t, the protein contains a known ER targeting signal.
[b]The data originating from the Proteome BioKnowledge® Library (http://www.incyte.com/bioknowledge) are not publicly available in the Hera database. This data represents 7.3% of the kernel ER proteins.

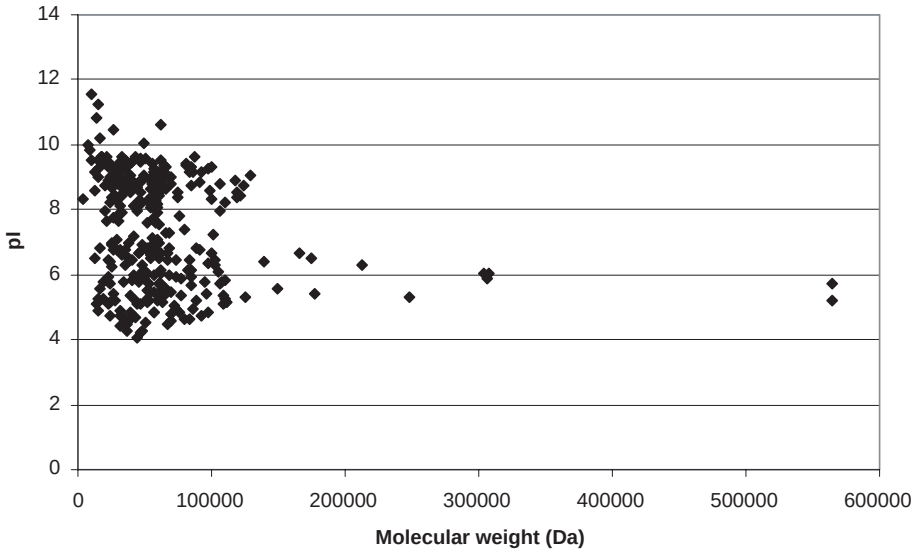

Fig. 2. Scatter plot of pI versus molecular weight estimated for all kernel ER proteins using the ProtParam tool. This plot reveals two general groups of proteins of roughly equal size. The first group contains 182 proteins whose pI is above 7.50. The second group is composed of 161 proteins with pI values below 7.50 and includes several very large proteins.

and these methods vary in stringency and levels of quality control. This collection process produced over 1000 ER proteins. Many of these proteins were present more than once under different names and accession numbers.

Duplicated sequences were removed before entry into the database using a specially designed sequence alignment program developed in our laboratory. The program implements the Smith–Waterman algorithm for pairwise alignment with affine gap costs in a linear space (Smith and Waterman, 1981). To score the alignment, we use the BLOSUM62 score matrix, and a gap cost of -8 per unaligned residue. The probability of score is calculated based on the methods of Altschul and Gish (1996). The program performs pairwise alignment of amino acid sequences and estimates the score of similarity and its probability, which are then used for human curation of the entries. Sequences identical to other entries were removed from the database.

Many classification sources label proteins as belonging to the ER even though they are only anchored on the cytoplasmic side of the ER or found in the ER periphery. By manually searching the scientific literature, our last curation step identified several such proteins in our database and relabeled them as belonging to a special class called 'ER periphery'. After curation, the database contained 499 human ER proteins.

Of the 499 ER proteins, 343 are classified as ER proteins based on experimental evidence (classification label 'e') or because the information source that provided this localization annotated the protein as being localized in the ER with a high degree of certainty (classification label 'c'). We refer to this set as the kernel ER proteins. This set of proteins does not include any ER proteins from the literature or public databases where authors employ categorizations such as 'potential ER protein'. Furthermore, proteins localized to the ER only via homology or by similarity search were not included in the kernel group. Our rationale for not allowing such localizations is based on several cases where such similarity searches predict human proteins to be classified as ER, whereas other sources predict these proteins to be non-ER. Such false predictions tend to occur when the protein is similar to an ER protein in another species. (However, whenever possible, we confirmed the ER localization of all these proteins by searching extensively through the literature.) Finally, no proteins belonging to the ER periphery group are part of the kernel ER set.

Characteristics of ER proteins

We used the Hera database to identify characteristics shared by all ER proteins or subgroups of ER proteins.

Physical properties The molecular weight and isoelectric point (pI) are important physical properties of proteins, since they determine the solubility of proteins and ultimately influence their subcellular localization and function. We used the ProtParam tool (http://ca.expasy.org/tools/protparam.html) to estimate the molecular weight and pI values of all proteins in our database. The average molecular weight and pI values of the 343 kernel ER proteins are respectively 60 539 Da and 7.458. When the pI values are plotted as a function of the molecular weights of these 343 proteins (Fig. 2), two general groups of proteins of roughly equal size become apparent. The first group contains proteins whose pI is above 7.50. This group consists of proteins with relatively small molecular weights (the highest molecular weight of this group is

129 kDa). Over 68% of proteins in this group have at least one transmembrane domain as predicted by TMHMM (Krogh *et al*., 2001). This is in agreement with a recent study, which reported that integral membrane proteins have comparatively high pI values that cluster around nine (Schwartz *et al*., 2001).

The second distinct group in Figure 2 is composed of proteins with pI values below 7.50. This group contains some very large proteins including seven proteins of molecular weight between 200 and 600 kDa. Six of these possess at least six transmembrane domains and function as ion exchangers or transporters. The majority of proteins in this second group however have a molecular weight below 125 kDa and pI values uniformly distributed between 4 and 7.5. Approximately 48% of the proteins in the second group contain at least one transmembrane domain as predicted by TMHMM. This group of proteins more closely resembles the cytoplasmic profile than either the nuclear or integral membrane protein profile previously defined (Schwartz *et al*., 2001).

Signal peptides and entry into the secretory pathway

N-terminal signal peptides are frequently responsible for the targeting of nascent polypeptides to the ER allowing for subsequent transport through the secretory pathway (Rapoport, 1992; von Heijne, 1990). As polypeptide synthesis progresses, the N-terminal signal peptide is recognized by the signal recognition particle (SRP) which causes the ribosome to come into contact with the ER and allows for the insertion of the nascent peptide chain into the ER. SignalP is a publicly available high accuracy tool that predicts the presence of signal peptides and anchors in proteins (Nielsen *et al*., 1997). We submitted the sequences of all the proteins in our database to SignalP (version 2.0). We also submitted sequences from 1756 human non-ER proteins with subcellular localization annotation in GeneCards (Safran *et al*., 2002). Figure 3A shows that 70% of the 343 kernel ER proteins contain either a signal peptide or a signal anchor. As well, our analysis also indicates that 80% of soluble kernel ER proteins contain a signal peptide (not shown in Fig. 3A). Other organelles of the secretory pathway also display a low percentage of proteins containing such signals: only 59% of all the proteins whose localization is the Golgi apparatus in GeneCards contain a signal peptide or anchor. By manually examining the Hera database, we have noticed that some proteins classified as ER in public databases or the published literature are simply bound to the cytoplasmic side of these organelles and thus do not require entry into the secretory pathway. We have relabeled these proteins as located at the ER periphery in Hera and, thus, they are not part of the present analysis (and are not part of the set of kernel ER proteins). It is possible that a small number of such proteins are still labeled as ER in Hera. However, we note that they would certainly not account for the 30% of kernel ER proteins that contain neither signal peptides nor anchors. It is equally unlikely that this high percentage of ER proteins carrying neither signal peptides nor anchors be due to other types of

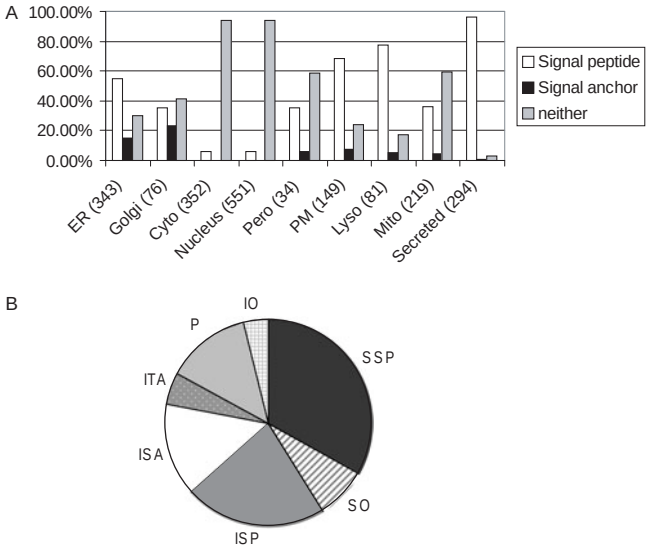

Fig. 3. (**A**) Distribution of signal peptides and anchors in proteins depending on their subcellular localization. SignalP was used to assess the presence of signal peptides and anchors in the kernel ER proteins as well as in 1756 non-ER proteins annotated as such in GeneCards (the exact number of proteins considered for each compartment is indicated in parentheses beside the compartment name). Some compartments of the secretory pathway display a low proportion of proteins that contain such signals. Cyto, cytosolic proteins; Pero, peroxisomal proteins; PM, integral plasma membrane proteins; Mito, mitochondrion proteins; Secreted, extracellular proteins. (**B**) Classification of the 343 kernel ER proteins depending on their solubility and signal type. SSP, soluble proteins containing a signal peptide; SO, other soluble proteins; ISP, insoluble proteins containing a signal peptide; ISA, insoluble proteins containing a signal anchor; ITA, tail-anchored insoluble proteins; P, polytopic proteins containing neither signal peptide nor signal anchor; IO, other insoluble proteins.

classification error in the Hera, SwissProt or GeneCards databases or prediction error due to SignalP. In contrast, a higher proportion of plasma membrane proteins and secreted proteins contain a signal peptide or anchor; these two localizations are more representative of what is expected for proteins of all compartments of the secretory pathway (Rapoport, 1992). As anticipated, a much smaller proportion of proteins in organelles that are not part of the secretory pathway contain a signal peptide or anchor.

Since a high proportion of proteins of the secretory pathway contain neither a signal peptide nor an anchor, it can be asked whether other insertion mechanisms into the ER are used in human cells. It is well known that *Saccharomyces cerevisiae* is capable of post-translational translocation of large precursors in an SRP-independent manner (Ng *et al*., 1996; Zheng and Gierasch, 1996). It has been shown that unlike mammalian systems, yeast cells depleted of SRP are viable (Kalies and Hartmann, 1998).

However, SRP-independent post-translational protein translocation across the ER membrane does occur in higher eukaryotes in the case of small proteins of fewer than 75 amino acids in length (Muller and Zimmermann, 1987; Kalies and Hartmann, 1998). Additionally, it has been shown that many tail-anchored proteins have the capacity for post-translational insertion into the ER (Kutay *et al*., 1995; Linstedt *et al*., 1995; Behrens *et al*., 1996). Tail-anchored proteins are characterized by their lack of a signal peptide or anchor and the presence of a single C-terminal transmembrane domain that acts as a membrane insertion sequence (Abell *et al*., 2003). A recent study provides evidence that tail-anchored proteins and signal-anchored proteins use membrane insertion pathways that share common elements (Abell *et al*., 2003). By searching the Hera database, we identified 17 such tail-anchored kernel ER proteins (8.4% of kernel ER membrane proteins).

We have further analyzed the remaining 86 kernel ER proteins which contain neither a signal peptide nor a signal anchor as predicted by SignalP and which are not tail-anchored proteins. Fifty-eight of these proteins are insoluble and 45 of these contain at least three transmembrane domains. Such polytopic proteins probably possess ER membrane insertion sequences in their membrane spanning regions. It has recently been shown that different transmembrane domains of a polytopic protein interact with different elements of the ER translocation machinery (Meacock *et al*., 2002). Some polytopic proteins have been shown to be inserted both co- and post-translantionally in microsomal membranes (Kanner *et al*., 2002). Finally, there are 28 soluble ER resident proteins that do not contain a signal peptide as predicted by SignalP. These include well-known ER luminal proteins such as DPM1 (Maeda *et al*., 2000). By using pattern recognition programs such as teireisias (Rigoutsos and Floratos, 1998), we have searched for subsequences common to several of these proteins that might serve as ER targeting signals, but have found none. Further analysis of these proteins will be required. The classification of the kernel ER proteins depending on their solubility and signal types is shown in Figure 3B.

Transmembrane domains It has been estimated that in most organisms, including *Homo sapiens*, between 20 and 35% of all cellular proteins contain at least one transmembrane domain (Stevens and Arkin, 2000). We analyzed the kernel ER proteins using TMHMM (Krogh *et al*., 2001). TMHMM has been shown to generally outperform other publicly available transmembrane domain predictors in a comparative study (Möller *et al*., 2001).

From the set of 343 kernel ER proteins, as many as 202 proteins (59%) contain at least one transmembrane domain. This number does not include the proteins predicted to have only one N-terminal transmembrane domain defined by SignalP as a cleavable signal peptide, since such proteins should be soluble once the signal peptide is cleaved. This high proportion of non-soluble ER proteins is not very surprising considering the high membrane composition of the ER and the need to maintain intercompartmental communications between the ER and other parts of the cell to support many different cellular processes. Table 2 shows the distribution of transmembrane domains in ER proteins in the database. Twenty-four of the kernel ER proteins contain more than seven transmembrane domains. These proteins include enzymes involved in protein glycosylation and modification of sugar chains, calcium channels, other types of transporters and proteins involved in lipid biosynthesis. Most ER proteins, however, contain fewer than eight transmembrane domains.

Table 2. Number of transmembrane domains (TMDs) in the 343 kernel ER proteins

Number of TMDs	Number of proteins
1	66
2	39
3	18
4	14
5	10
6	14
7	17
8	8
9	6
10	3
>10	7

Table 3. Most prevalent KDEL-like C-terminal motifs of soluble kernel ER proteins

Motif	Number of occurrences
KDEL>	9
KEEL>	5
HDEL>	5
RDEL>	3
SDEL>	3
H-[VITEN]-EL>	6

Retention/retrieval signals The C-terminal KDEL signal is the most well-known ER retrieval sequence. Since its discovery (Munro and Pelham, 1987), it has been found to be widely used in many different organisms and has been extended to the much more general Prosite consensus pattern [KRHQSA]-[DENQ]-E-L> (Sigrist *et al*., 2002). We have searched Hera for proteins containing this signal, using a program we developed in our laboratory. In total, 18.4% of the soluble proteins in the kernel ER set have the Prosite KDEL-like sequence. If we extend the pattern to [KRHQSADEN]-[DENQTFIV]-E-[LF]>, 26.2% of the soluble proteins contain the signal. Table 3 shows the most prevalent C-terminal KDEL-like motifs in kernel ER

Table 4. Most prevalent basic C-terminal motifs of ER membrane proteins

Motif	Number of occurrences
KKxx>	19
KxKxx>	29
KxKxD>	10
KKxKxx>	9
KSKXX>	8
KKAx>	7
KRx>	7
KXKTX>	7
KxKAx>	7
KKKXX>	7

proteins. Among the remaining 73.8% of soluble kernel ER proteins that do not contain this signal, nine proteins contain C-terminal sequences somewhat similar to the extended KDEL motif. These sequences are QEDL>, PDAL>, KENL>, IERL>, FKHL>, RAKL>, AKAL>, QVLE> and KLYL>. It will be necessary to experimentally verify the functionality of these possible retrieval signals.

Other well-characterized ER sorting signals include the C-terminal di-lysine motif, which has been shown to be necessary and sufficient for ER retention for some proteins, and the N-terminal di-arginine motif (reviewed in Teasdale and Jackson, 1996). These motifs are believed to be used by type I and type II membrane proteins respectively. We analyzed the 202 membrane proteins in the kernel ER set for the presence of these motifs. The di-lysine motif was generalized to the pattern UUUXX> where at least two out of the three U's must be a lysine, X can be any amino acid and > indicates the C-terminal end of the protein. Although this motif is not believed to be widely used (Teasdale and Jackson, 1996), our results indicate that 22.8% of ER membrane proteins contain this C-terminal motif. If we extend the motif to UUUUX> where at least two out the four U's must be a basic residue (lysine, arginine or histidine), we obtain that 38.6% of ER membrane proteins contain this motif. It will be necessary to experimentally assess the biological relevance of this extended motif. Table 4 shows the most common C-terminal motifs in kernel in human ER proteins. Our results do not agree well with previous values obtained by a combinatorial screen in mammalian cells to determine the C-terminal tetrapeptides that are most efficient in retaining proteins in the ER (Zerangue *et al*., 2001), suggesting that other parts of the protein sequence also play an important part in determining the final localization of the protein.

Only 1.5% of ER membrane proteins contain the di-arginine motif <X(2,3)-RR [this increases to 4.5%, if the motif is extended to <X(2,3)-R/K-R/K]. This is in line with the belief that the motif is more widely used by Golgi proteins than ER proteins. In total, 36.2% of the 343 kernel ER set of proteins contain the KDEL-like, the di-lysine or the di-arginine retention motifs.

Table 5. Most common InterPro domains and motifs in kernel ER proteins

InterPro entry identifier	InterPro entry name	Number (%) of kernel ER proteins containing entry	Number of non-ER proteins containing entry[a]
IPR001128	Cytochrome P450	37 (10.8)	4
IPR002213	UDP-glucoronosyl/UDP-glucosyl transferase	13 (3.8)	0
IPR006663	Thioredoxin domain 2	10 (2.9)	5
IPR003608	MIR domain	8 (2.3)	0
IPR000719	Protein kinase	7 (2.0)	42
IPR002048	Calcium-binding EF-hand	7 (2.0)	10
IPR000379	Esterase/lipase/thioesterase, active site	7 (2.0)	9
IPR003388	Reticulon	6 (1.7)	0
IPR001682	Ca^{2+}/Na^{+} channel, pore region	6 (1.7)	0
IPR005821	Ion transport protein	6 (1.7)	0

[a]1756 proteins annotated in GeneCards as being in a subcellular compartment other than the ER were scanned for the presence of InterPro motifs.

Other ER retention motifs have been described, including the internal RXR motif for membrane proteins (Shikano and Li, 2003) and the cytochrome b5 transmembrane domain (Honsho *et al*., 1998). However, most of these motifs are either too general (i.e. equal proportions of ER and non-ER proteins contain these motifs and these motifs might be position-specific) or too specific (i.e. found in very few ER proteins, possibly only one) to be used in the analysis of the kernel ER set. In the case of soluble ER proteins, a KDEL-receptor independent ER retention has been described for the lysyl hydroxylase protein (Suokas *et al*., 2003). Other retention mechanisms have been proposed including retention by oligomerization and through membrane thickness (reviewed in Nilsson and Warren, 1994).

Protein domains and motifs We assessed the presence of different protein motifs and domains in ER proteins via InterProScan software (Mulder *et al*., 2003). This software provides a platform for an integrated use of many secondary protein databases. We also used this software to scan 1756 proteins annotated in GeneCards as being in compartments other than the ER. In total, 84.5% of the 343 kernel ER proteins contain at least one domain or motif corresponding to an InterPro entry. Table 5 shows the most abundant InterPro entries found in ER proteins. The InterPro entry recognized in the largest number of ER proteins is the cytochrome P450 motif, which is present in 10.8% of the ER proteins. Cytochrome P450 proteins are involved in the metabolism of many different compounds by acting as terminal oxidases in electron transfer chains. Most eukaryotic cytochrome P450s are believed to be localized

at microsomal membranes (Werck-Reichhart and Feyereisen, 2000). Three children motifs (motifs that further characterize a subgroup of proteins that contain the parent motif) of the InterPro cytochrome P450 motif were detected in ER proteins (not shown in Table 5). These are the E-class P450 group I, E-class P450 group II and E-class P450 group IV motifs.

The second most abundant motif in ER proteins (Table 5) is the UDP-glucosyl transferase motif, which is used by proteins involved in the folding and quality control pathway of the ER (Ellgaard and Helenius, 2003). This motif seems specific to ER proteins since none of the 1756 non-ER proteins contain it. The thioredoxin domain 2 motif is also present in several chaperone-like ER proteins. Thioredoxins are small protein disulphide oxidoreductases (Martin, 1995). Other motifs that appear to be ER-specific are related to calcium storage and regulation, transport and signaling pathways; these are all well-studied functions of the ER. In contrast, the protein kinase motif is contained indiscriminately in ER and non-ER proteins, indicating that it is not ER-specific. The calcium-binding EF-hand, the thioredoxin domain 2 and the esterase/lipase/thioesterase families are present in proteins in several different compartments indicating that different organelles share some similar functions. Most of the abundant motifs in ER proteins detected by InterProScan seem to be present in proteins involved in the well-characterized functions of the ER. The study of ER-specific motifs is important in the characterization of the ER and the determination of its detailed role in the cell.

Future improvements

The Hera database was built to store characteristics of human ER proteins. As the Hera project progresses, we will add new features to the database including information regarding protein isoforms, other names (synonyms), interaction partners and mechanisms of involvement in human disease where relevant. As well, we are in the process of adding yeast ER proteins to Hera and we will subsequently include ER proteins from other species. This will allow us to address better the question of the completeness of the database and will provide a platform for evolutionary studies of ER proteins and pathways. We will also further annotate the Hera database to provide functional classification of the proteins and integrate experimental protein interaction data. Hera is a unique resource that will allow a comprehensive characterization of ER proteins and thus improve our understanding of this organelle and its associated diseases. We believe the Hera database will help biologists in increasing the efficiency of finding useful and relevant information regarding ER proteins, as well as bioinformaticians in the characterization of the human proteome.

ACKNOWLEDGEMENTS

We are grateful to Martha Lopez for data entry during the development of Hera and Dr Scott Bunnell for critical reading of this manuscript. We wish to thank Dr Bettina Kemme, Zsuzsanna Bencsath-Makkai and Stephanie Pollock for help in the design of the Hera ER model and François Pepin for logistical support. This work was supported by grants to D.Y.T. and M.H. from Genome Quebec/Genome Canada as well as to D.Y.T. from the Canadian Institutes of Health Research (CIHR). M.S. is a recipient of a Canada Graduate Scholarship (CGS).

REFERENCES

Abell,B.M., Jung,M., Oliver,J.D., Knight,B.C., Tyedmers,J., Zimmermann,R. and High,S. (2003) Tail-anchored and signal-anchored proteins utilize overlapping pathways during membrane insertion. *J. Biol. Chem.*, **278**, 5669–5678.

Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Meth. Enzymol.*, **266**, 460–480.

Behrens,T.W., Kearns,G.M., Rivard,J.J., Bernstein,H.D. and Yewdell,J.W., Staudt,L.M. (1996) Carboxyl-terminal targeting and novel post-translational processing of JAW1, a lymphoid protein of the endoplasmic reticulum. *J. Biol. Chem.*, **271**, 23528–23534.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Ellgaard,L. and Helenius,A. (2003) Quality control in the endoplasmic reticulum. *Nat. Rev. Mol. Cell. Biol.*, **4**, 181–191.

Honsho,M., Mitoma,J.Y. and Ito,A. (1998) Retention of cytochrome b5 in the endoplasmic reticulum is transmembrane and luminal domain-dependent. *J. Biol. Chem.*, **273**, 20860–20866.

Kalies,K.-U. and Hartmann,E. (1998) Protein translocation into the endoplasmic reticulum (ER)—two similar routes with different modes. *Eur. J. Biochem.*, **254**, 1–5.

Kanner,E.M., Klein,I.K., Friedlander,M. and Simon,S.M. (2002) The amino terminus of opsin translocates "posttranslationally" as efficiently as cotranslationally. *Biochemistry*, **41**, 7707–7715.

Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Kutay,U., Ahnert-Hilger,G., Hartmann,E., Wiedenmann,B. and Rapoport,T.A. (1995) Transport route for synaptobrevin via a novel pathway of insertion into the endoplasmic reticulum membrane. *EMBO J.*, **14**, 217–223.

Lee,A.S. (2001) The glucose-regulated proteins: stress induction and clinical applications. *Trends Biochem. Sci.*, **26**, 504–510.

Linstedt,A.D., Foguet,M., Renz,M., Seelig,H.P., Glick,B.S. and Hauri,H.P. (1995) A C-terminally-anchored Golgi protein is inserted into the endoplasmic reticulum and then transported to the Golgi apparatus. *Proc. Natl Acad. Sci., USA*, **92**, 5102–5105.

Maeda,Y., Tanaka,S., Hino,J., Kangawa,K. and Kinoshita,T. (2000) Human dolichol-phosphate-mannose synthase consists of three subunits, DPM1, DPM2 and DPM3. *EMBO J.*, **19**, 2475–2482.

Martin,J.L. (1995) Thioredoxin—a fold for all reasons. *Structure*, **3**, 245–250.

Meacock,S.L., Lecomte,F.J., Crawshaw,S.G. and High,S. (2002) Different transmembrane domains associate with distinct endoplasmic reticulum components during membrane integration of a polytopic protein. *Mol. Biol. Cell*, **13**, 4114–4129.

Möller,S., Croning,M.D.R. and Apweiler,R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.

Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.

Muller,G. and Zimmermann,R. (1987) Import of honeybee prepromelittin into the endoplasmic reticulum: structural basis for independence of SRP and docking protein. *EMBO J.*, **6**, 2099–2107.

Munro,S. and Pelham,H.R. (1987) Sorting of membrane proteins in the secretory pathway. *Cell*, **48**, 899–907.

Ng,D.T., Brown,J.D. and Walter,P. (1996) Signal sequences specify the targeting route to the endoplasmic reticulum membrane. *J. Cell Biol.*, **134**, 269–278.

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.

Nilsson,T. and Warren,G. (1994) Retention and retrieval in the endoplasmic reticulum and the Golgi apparatus. *Curr. Opin. Cell Biol.*, **6**, 517–521.

Paschen,W. and Frandsen,A.J. (2001) Endoplasmic reticulum dysfunction—a common denominator for cell injury in acute and degenerative diseases of the brain? *J. Neurochem.*, **79**, 719–725.

Pelletier,M., Bergeron,J.J. and Thomas,D. (2001) Molecular chaperone systems in the endoplasmic reticulum. In Lund,P. (ed.), *Molecular Chaperones: Frontiers in Molecular Biology Series*. Oxford University Press, Oxford, pp. 180–199.

Rapoport,T.A. (1992) Transport of proteins across the endoplasmic reticulum membrane. *Science*, **258**, 931–936.

Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.

Safran,M., Solomon,I., Shmueli,O., Lapidot,M., Shen-Orr,S., Adato,A., Ben-Dor,U., Esterman,N., Rosen,N., Peter,I. *et al*. (2002) GeneCards(TM) 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, **18**, 1542–1543.

Schwartz,R., Ting,C.S. and King,J. (2001) Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Res.*, **11**, 703–709.

Shikano,S. and Li,M. (2003) Membrane receptor trafficking: evidence of proximal and distal zones conferred by two independent endoplasmic reticulum localization signals. *Proc. Natl Acad. Sci., USA*, **100**, 5783–5788.

Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform.*, **3**, 265–274.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Stevens,T.J. and Arkin,I.T. (2000) Do more complex organisms have a greater proportion of membrane proteins in their genomes? *Proteins*, **39**, 417–420.

Suokas,M., Lampela,O., Juffer,A.H., Myllyla,R. and Kellokumpu,S. (2003) Retrieval-independent localization of lysyl hydroxylase in the endoplasmic reticulum via a peptide fold in its iron-binding domain. *Biochem. J.*, **370**, 913–920.

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.

von Heijne,G. (1990) The signal peptide. *J. Membr. Biol.*, **115**, 195–201.

Werck-Reichhart,D. and Feyereisen,R. (2000) Cytochromes P450: a success story. *Genome Biol.*, **1**, REVIEWS3003.

Wu,C.H., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z.-Z., Ledley,R.S., Lewis,K.C., Mewes,H.-W., Orcutt,B.C. *et al*. (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, **30**, 35–37.

Zerangue,N., Malan,M.J., Fried,S.R., Dazin,P.F., Jan,Y.N., Jan,L.Y. and Schwappach,B. (2001) Analysis of endoplasmic reticulum trafficking signals by combinatorial screening in mammalian cells. *Proc. Natl Acad. Sci., USA*, **98**, 2431–2436.

Zheng,N. and Gierasch,L.M. (1996) Signal sequences: the same yet different. *Cell*, **86**, 849–852.