Mouse Proteome Analysis

Alexander Kanapin,[1,11] Serge Batalov,[4] Melissa J. Davis,[3] Julian Gough,[5]
Sean Grimmond,[3] Hideya Kawaji,[2,8] Michele Magrane,[1] Hideo Matsuda,[2]
Christian Schönbach,[6] Rohan D. Teasdale,[3] RIKEN GER Group[7] and
GSL Members,[9,10] and Zheng Yuan[3]

[1]EMBL Outstation—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK; [2]Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Toyonaka, Osaka 560-8531, Japan; [3]Institute for Molecular Bioscience and ARC Special Research Centre for Functional and Applied Genomics, University of Queensland, St. Lucia, Queensland 4072, Australia; [4]Genomic Institute of the Novartis Research Foundation (GNF), San Diego, California 92121, USA; [5]Department of Structural Biology, Stanford University, Stanford, California 94305, USA; [6]Knowledge Discovery Team, Bioinformatics Group, RIKEN Genomic Sciences Center, Yokohama 230-0045, Japan; [7]Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan; [8]NTT Software Corporation, Naka-ku, Yokohama, Kanagawa, 231-8554, Japan; [9]Genome Science Laboratory, RIKEN, Hirosawa, Wako, Saitama 351-0198, Japan

A general overview of the protein sequence set for the mouse transcriptome produced during the FANTOM2 sequencing project is presented here. We applied different algorithms to characterize protein sequences derived from a nonredundant representative protein set (RPS) and a variant protein set (VPS) of the mouse transcriptome. The functional characterization and assignment of Gene Ontology terms was done by analysis of the proteome using InterPro. The Superfamily database analyses gave a detailed structural classification according to SCOP and provide additional evidence for the functional characterization of the proteome data. The MDS database analysis revealed new domains which are not presented in existing protein domain databases. Thus the transcriptome gives us a unique source of data for the detection of new functional groups. The data obtained for the RPS and VPS sets facilitated the comparison of different patterns of protein expression. A comparison of other existing mouse and human protein sequence sets (e.g., the International Protein Index) demonstrates the common patterns in mammalian proteomes. The analysis of the membrane organization within the transcriptome of multiple eukaryotes provides valuable statistics about the distribution of secretory and transmembrane proteins

The Mouse Gene Encyclopedia project (FANTOM Consortium 2002) provides a unique opportunity for researchers to investigate a mammalian proteome from its functional perspective. The data provide a snapshot of proteins present in the living cell and can therefore be used for functional analysis and classification.

The following paper summarizes a general analysis of the mouse proteome sets deduced from the transcriptome DNA sequences based on various algorithms and approaches. We used protein domain databases, namely InterPro (Apweiler et al. 2001) and Superfamily (Gough and Chothia 2002), to carry out initial functional annotation of the protein sequences and to classify these sequences according to existing biological resources, such as Gene Ontology (GO). The general coverage of proteins in the representative proteins set is about 92% for both InterPro and Superfamily, and this provides a comprehensive overview of the proteome. InterPro analysis has also been used for comparison of the different proteomes produced; this analysis highlights interesting differences between various mouse sequencing projects. New domains which are not included in existing resources have been detected using algorithms implemented in the MDS database (Kawaji et al. 2002), and seven new domain candidates have been discovered. Determination of the membrane organization within the secretory pathway, namely whether a protein is secreted into the extracellular media, a membrane-spanning protein (transmembrane protein), or a nonsecretory protein, is essential for understanding its function. This information complements other computational annotation projects, as it provides the context by determining the membrane topography of predicted functional protein units and is essential for the prediction of subcellular localization, which depends on the class of protein.

RESULTS AND DISCUSSION

Two protein sets have been produced as a result of the FANTOM2 sequencing project. The representative proteins set (RPS) is derived from the representative set of transcriptional units. The variant-based proteins set (VPS) combines RPS and complete protein sequences representing splice variants not included in RPS. The VPS includes variant forms of known genes identified by sequencing of the FANTOM2 clones. We summarized the characterization of the sets in the main FANTOM2 paper (FANTOM Consortium 2002). We describe

[10]Takahiro Arakawa, Piero Carninci, Jun Kawai, and Yoshihide Hayashizaki.
[11]Corresponding author.
E-MAIL alex@ebi.ac.uk; FAX 44 0 1223 494610.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.978703.

here the different characteristics of the variants and provide comparisons with other available sequence data for mouse and human.

InterPro Matches Statistics

The major goal of the domain/site/motif composition analysis was to obtain a general functional overview of the proteome and to use these results for initial functional assignments. We used InterPro as a standard tool to determine the domain/site/motif composition of different mouse protein sequence data sets. In addition to the RPS and VPS described earlier, we also analyzed a mouse sequence data set of hypothetical proteins computationally predicted by Celera and the nonredundant mouse protein set produced as part of the International Protein Index (IPI) (http://www.ebi.ac.uk/IPI). The human protein set provided by IPI was also analyzed.

The general number of proteins for both FANTOM2 proteome sets having matches for InterPro entries is about 72% (92% for combined InterPro and Superfamily databases). This amount is quite similar to other existing proteomes analyzed in the Proteome Analysis Database (http://www.ebi.ac.uk/proteome); about 60%–75% for complete proteomes in the database. This provides some evidence of the high quality of the FANTOM2 data. We also analyzed amino-acid frequency distribution for the mouse protein sequences (data not shown). The difference in the frequencies between the different mouse datasets is only about 0.3%, which is far less than the difference between various eukaryotic proteomes (about 3%).

Comparative Proteomics

The algorithms implemented in the Proteome Analysis Database also include several InterPro-based statistical analyses, including a list of the top 20 InterPro entries. Table 1 presents statistics for the described mouse proteome data and also includes human IPI statistics. The analyses suggest that the general domain/site/motif composition is similar for all four mouse proteome sets. The statistics of the InterPro entries can be used to infer some functional information about the proteome. The most commonly represented functional groups are nucleic acid binding proteins and proteins belonging to the immunoglobulin family. The other major group of InterPro entries includes serine/threonine and tyrosine protein kinase domains. The RPS and VPS proteome sets have similar statistics for InterPro entry composition, which describes the protein sets from the point of view of functional domains/sites/motifs. This can be considered evidence of the relative stability of the functional potential of the transcriptome, which maintains a constant ratio of proteins of different functions despite the presence of splice variants. The InterPro entry-matches distribution is very similar for the human and mouse proteomes at the level of the top entries and can, therefore, provide valuable information about conserved functional domains/sites/motifs across different mammalian species.

Functional Classification and Assignment

InterPro analysis also provides a basis for the functional assignment of proteins to standard biological classification resources, such as Gene Ontology (GO). We used the existing curated mapping of InterPro entries to GO terms to classify the mouse and human proteome sets described above. A modified version of GO called "GO Slim" which is implemented in the Proteome Analysis Database was used to compare the functional composition of the proteomes. GO Slim comprises a selection of high-level terms from each of the three GO sections (molecular function, biological process, and cellular component), which were chosen to cover most aspects of the three ontologies without overlapping in the GO hierarchy. The molecular function terms of GO Slim were used here to provide an overview of the functional composition of the proteomes. The results are presented in Figure 1. The diagram shows that FANTOM sets are similar to each other rather than to other sequence data. They differ mainly in quantitative order, but not in the "pattern" of proteins synthesized. There is also a great degree of similarity between two genome sequence data sets—Celera and IPI—they show more similarity to each other rather than to FANTOM2 data. The number of G-protein coupled receptor (GPCR) proteins was higher for the Celera set, possibly indicating the lacking annotation of pseudogenes, abundant for this class of proteins. At the time of publication, Celera, through the process of expert curation, retired 20% of their gene predictions and annotated 6% as pseudogenes (Release R13d).

We also compared the GO Slim statistics for the mouse and human proteome sets with other less closely related eukaryotes, namely *Drosophila melanogaster* and *Arabidopsis thaliana*. The resulting diagram is presented in Figure 2. The general functional overview for the different eukaryotic proteomes is quite similar despite some obvious differences between the mammalian proteomes and those of the plant and insect species, which were used for comparison across a wider range of species. The statistics provide an insight into the conserved functional groups common to all eukaryotic genomes.

Superfamily Domain Analysis

The SUPERFAMILY hidden Markov model library (Gough and Chothia 2002), representing all proteins of known structure, was run against the complete FANTOM2 mouse cDNA collection (FANTOM Consortium 2002). The results in this section correspond to the VPS set of sequences, because this is the closest available representation of the actual mouse proteome. Detailed results are available at the SUPERFAMILY web site (http://supfam.org/SUPERFAMILY/cgi-bin/gen_list.cgi?genome=mr).

The SUPERFAMILY analysis is used to detect and classify evolutionarily related groups of domains for which there is a known structural representative. All assigned domains are classified at the superfamily level. An accurate superfamily level of classification is obtained by a detailed hand analysis by an expert of structural, sequence, and functional evidence of a common evolutionary ancestor (Murzin et al. 1995). The superfamily classification used is that defined by the SCOP database.

Functional Annotation

These assignments provide information which has been used as part of the MATRICS, annotation of the FANTOM2 mouse cDNA sequences. Because proteins with the same structure usually have the same or a related function, the structural domain assignments are a useful component of the information used in the functional annotation. Furthermore, the SUPERFAMILY analysis provides assignments for many sequences where there is little or no other significant information. At least one domain was assigned to 59% of the sequences, and the domains cover 42% of all residues. This is

Table 1. Top 20 InterPro Entries for Different Mouse Protein Sequence Sets and Human Proteome

	Mouse FANTOM RPS sequences		*Mouse FANTOM VPS sequences*		*Celera R13d*		*M. musculus IPIs*		*H. sapiens IPI*		
IntroPro	**Proteins matched (Proteome coverage)**	**Rank***	**Proteins matched (Proteome coverage)**	**Rank***	**Proteins matched (Proteome coverage)**	**Rank***	**Proteins matched (Proteome coverage)**	**Rank***	**Proteins matched (Proteome coverage)**	**Rank***	***Description***
IPR000694	1001 (5.3%)	1	1701 (5.0%)	2	1156 (5.6%)	3	544 (0.8%)	3	219 (0.7%)	25	Proline-rich region
IPR000822	460 (2.5%)	2	795 (2.3%)	4	682 (3.3%)	4	437 (1.7%)	6	942 (3.0%)	1	Zn-finger, C2H2 type
IPR000719	419 (2.2%)	3	831 (2.4%)	5	553 (2.7%)	7	448 (1.7%)	5	696 (2.2%)	3	Eukaryotic protein kinase
IPR002290	400 (2.1%)	4	777 (2.3%)	7	484 (2.3%)	9	283 (1.1%)	7	459 (1.5%)	5	Serine/Threonine protein kinase
IPR001245	383 (2.0%)	5	735 (2.1%)	8	445 (2.2%)	10	179 (0.7%)	18	288 (0.9%)	15	Tyrosine protein kinase
IPR000276	354 (1.9%)	6	544 (1.6%)	9	1176 (5.7%)	2	1347 (5.2%)	1	587 (1.9%)	4	Rhodopsin-like GPCR superfamily
IPR003599	313 (1.7%)	7	821 (2.4%)	6	507 (2.5%)	8	164 (0.6%)	20	351 (1.1%)	10	Immunoglobulin subtype
IPR003006	291 (1.6%)	8	832 (2.4%)	3	538 (2.6%)	6	604 (2.3%)	4	877 (2.8%)	2	Immunoglobulin/major histocompatibility complex
IPR001680	218 (1.2%)	9	457 (1.3%)	10	268 (1.3%)	13	224 (0.9%)	11	369 (1.2%)	6	G-protein beta WD-40 repeat
IPR001356	214 (1.1%)	10	343 (1.0%)	15	239 (1.2%)	20	234 (0.9%)	10	256 (0.8%)	19	Homeobox
IPR001841	199 (1.1%)	11	400 (1.2%)	13	251 (1.2%)	18	207 (0.8%)	13	360 (1.2%)	9	Zn-finger, RING
IPR001611	190 (1.0%)	12	340 (1.0%)	17	238 (1.2%)	21	139 (0.5%)	27	238 (0.8%)	21	Leucine-rich repeat
IPR000504	185 (1.0%)	13	407 (1.2%)	12	287 (1.4%)	11	205 (0.8%)	11	339 (1.1%)	8	RNA-binding region RNP-1 (RNA recognition motif)
IPR001849	182 (1.0%)	14	317 (0.9%)	19	214 (1.0%)	25	172 (0.7%)	18	329 (1.1%)	11	Pleckstrin-like
IPR002048	181 (1.0%)	15	324 (0.9%)	14	241 (1.2%)	14	193 (0.7%)	13	282 (0.9%)	14	Calcium-binding EF-hand
IPR002110	177 (0.9%)	16	303 (0.9%)	22	225 (1.1%)	23	156 (0.6%)	20	294 (0.9%)	15	Ankyrin
IPR001452	170 (0.9%)	17	312 (0.9%)	20	191 (0.9%)	26	188 (0.7%)	15	283 (0.9%)	17	SH3 domain
IPR003598	170 (0.9%)	18	340 (1.0%)	17	233 (1.1%)	22	136 (0.5%)	27	282 (0.9%)	18	Immunoglobulin C-2 type
IPR001909	152 (0.8%)	19	241 (0.7%)	27	223 (1.1%)	24	114 (0.4%)	31	328 (1.1%)	12	KRAB box
IPR005225	145 (0.8%)	20	262 (0.8%)	25	7 (0.0%)	774	140 (0.5%)	25	181 (0.6%)	29	Small GTP-binding protein domain
	Total: 5804 (30.9%)		Total: 11,082 (32.3%)		Total: 8158 (39.5%)		Total: 5570 (21.5%)		Total: 7960 (25.7%)		

The top 20 InterPro entries with the highest number of protein matches for the reference proteome. The number of proteins matched for each InterPro entry and the percentage of proteome coverage that this number represents are displayed. The rank represents the position of each InterPro entry in the list ordered according to the number of protein matches.

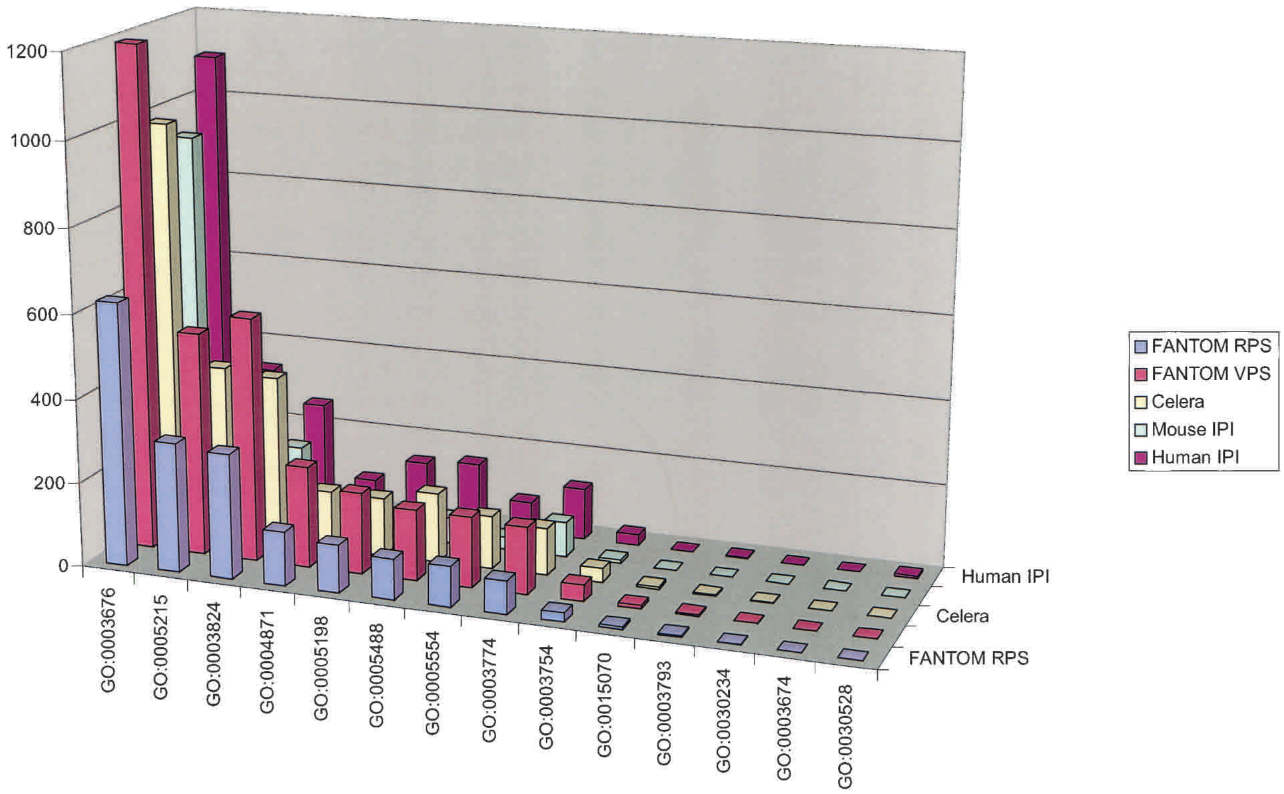

Figure 1 Comparative diagram of the GO Slim categories for different mouse and human proteomes. The *y*-axis indicates the number of proteins; the *x*-axis indicates the following GO categories: GO:0003676 nucleic_acid_binding; GO:0003754 chaperone; GO:0003774 motor; GO:0003793 defense/immunity_protein; GO:0003824 enzyme; GO:0004871 signal_transducer; GO:0005198 structural_molecule; GO:0005215 transporter; GO:0005488 binding; GO:0005554 molecular_function_unknown; GO:0015070 toxin; GO:0030234 enzyme_regulator; GO: 0030528 transcription_regulator.

close to the coverage of other eukaryote proteomes. The rest of the analysis shown in this section pertains to the subset of sequences and domains which were detected. The top 12 superfamilies are shown in Table 2. This is very similar to that which is observed in the human proteome based on gene predictions from the genomic sequence (Hubbard et al. 2002).

Structural Genomics

There are implications for experimental structural genomics projects (Gough 2002), most notably the discovery of novel domain combinations. Using strict criteria, pairwise structural domain combinations were enumerated, and compared to the already solved combinations in the Protein Data Bank (PDB) (Berman et al. 2000). Here, 335 structurally novel pairs were identified, 29 of which had not previously been found in any other proteome. These are listed at http://supfam.org/FANTOM2/domcombs.html). Although three-dimensional structures of the individual domains exist in the PDB, structures of these combinations of domains adjacent to each other on the polypeptide chain have not yet been solved. As well as being unique recombination events in evolution, these domain pairs provide targets for structural genomics projects which are assured to be novel. Solving the structures of novel domain-pair combinations will probably yield new 3D interfaces, which could be essential to or play an active role in the function of the protein as a whole.

Evolutionary Overview

The evolutionary relationships can give us a meaningful overview of a large proportion of the genome. The ancestral domain from each superfamily represents a genetic building block. These building blocks have been duplicated, recombined, and mutated to create the proteins which are currently observed in the genome. The assigned domain architecture for each sequence (available at the URL at the beginning of this section) shows the recombination of ancestral domains which has taken place during evolution. It can be seen that a small number of domains have been duplicated a very large number of times (see Fig. 3), and that a large number of domains have been duplicated very few times (see Fig. 4). In fact, 98% of the identified domains have been produced by duplication from 716 ancestral domains. This is very close to the pattern observed in the human proteome.

Novel Domains

We applied the MDS motif discovery method (Kawaji et al. 2002) to the FANTOM2 cDNA sequence set and identified seven new motif candidates that were deposited into the MDS database (http://motif.ics.es.osaka-u.ac.jp/fantom2/). Two candidate motifs (MDS00150, MDS00155) were found among hypothetical proteins, and five new motif candidates have been identified among proteins related to SCML2 (MDS00151), VPARP (MDS00152 and MDS00153), IAN4 (MDS00154), and ADMP (MDS00156). MDS00154 is a new

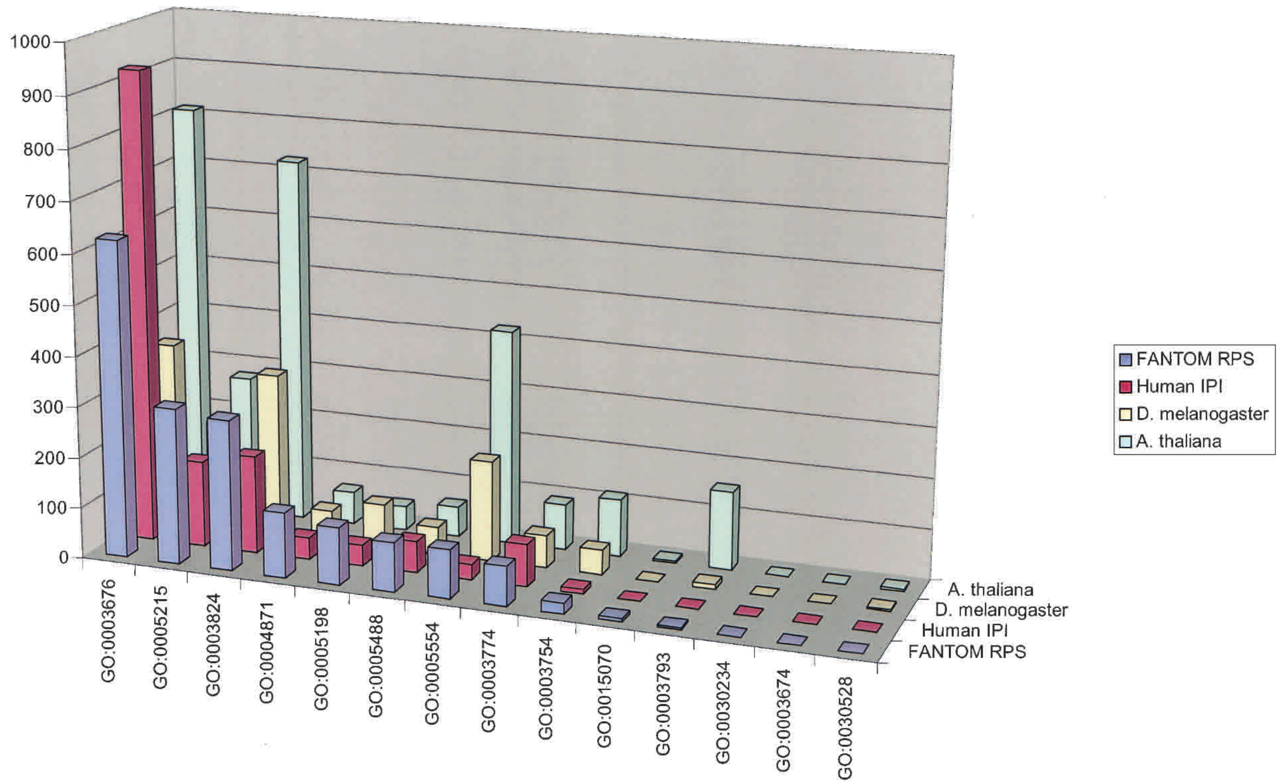

Figure 2 Comparative diagram of the GO Slim categories for different eukaryotic proteomes. The *y*-axis shows the number of proteins; the x-axis shows GO categories (see Fig. 1).

structural GTPase submotif that is specific for proteins of the immune associate nucleotide family (IAN), which is conserved in mammals and plants (Poirier et al. 1999). Interestingly, our motif appears to be restricted to mammals (FANTOM Consortium 2002). MDS00151 is a nuclear-localization signal containing a repeat motif for the transcriptional repressor gene sex-comb on midleg-like-2 (Scml2; Montini et al. 1999) and its homologs (AK016533, 6030439N15). The motif spans 24 amino acids and has two to six copies. Closer analysis of the flanking regions revealed that MDS00151 contains the NLS (nuclear localization signal) [KR]{3,5} and is flanked by the NLS KKPx{6,9}KxKR. The flanking NLS region and MDS00151 were not found in any other mammalian, suggesting an insertion/duplication event in the mouse lineage and a specific role for the nuclear import of mouse Scml2.

Table 2. The Top 12 Most Commonly Occurring Superfamily Domains

Rank	Domains	Proteins	Superfamily
1	3500	685	C2H2 and C2HC zinc fingers
2	2490	1201	Immunoglobulin
3	1550	1286	P-loop containing nucleotide triphosphate hydrolases
4	1449	307	EGF/Laminin
5	1235	182	Cadherin
6	912	893	Protein kinase-like (PK-like)
7	907	882	Membrane all-alpha
8	868	295	Fibronectin type III
9	791	449	RNA-binding domain, RBD
10	551	480	PH domain-like
11	509	421	Homeodomain-like
12	482	410	EF-hand

In addition, we performed, with the MDS motifs that were extracted from the FANTOM1 sequence set (Kawaji et al. 2002), Hidden Markov Model (HMM) searches against the FANTOM2 protein sequences and SWISS-PROT/TrEMBL nonredundant database (SWISS-PROT Release 40.27 of 30-Aug-2002, TrEMBL Release 21.12 of 13-Sep-2002, and TrEMBL_new of 13-Sep-2002). As a result, we obtained several new members with the FANTOM1 MDS motifs (see http://motif.ics.es.osaka-u.ac.jp/fantom2/). Proacrosin binding protein (E130112G13; AK053586) was detected as a new member of motif MDS00105.2 (the ING1-homolog subfamily motif)-containing proteins. The multiple alignment of the subfamily members (see http://motif.ics.es.osaka-u.ac.jp/fantom2/) shows that the clone is identical to a newly found splicing variant (mINGh-L, ING1-like protein long form; TrEMBL AAK63168) of the mouse ING1-homolog proteins (Ha et al. 2002).

Molecules interacting with CasL (MICALs; Suzuki et al. 2002; Terman et al. 2002) derived from human (TrEMBL Q8TDZ2), fruitfly (TrEMBL AAM55242, AAM55243, AAM55244, and MICAL-like protein, TrEMBL AAM55245), and mouse (TrEMBL AAH34682) were detected as new members of the leucine zipper-like motif MDS00113-containing proteins. Terman and coworkers (2002) showed that the fruitfly MICAL interacts with neuronal plexin A (PlexA) receptor in its C-terminal region including the MDS00113 motif, con-

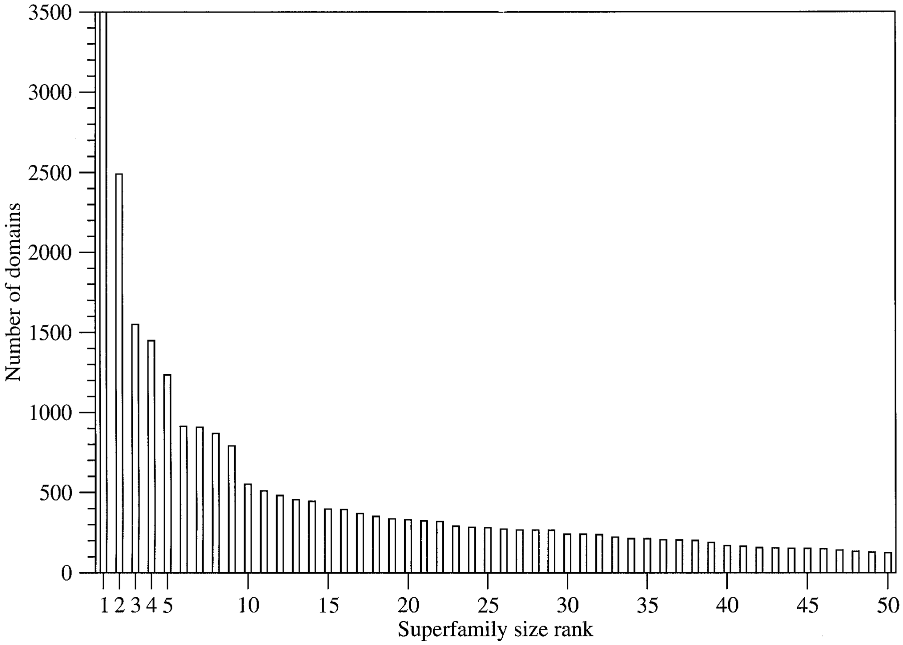

Figure 3 Observed evolutionary domains. The ordered sizes of superfamilies with greater than 100 members.

firming our previous prediction that this motif may act as a novel protein–protein interaction site.

Motif MDS00146, which previously comprised only hypothetical proteins, was expanded to the human Cdc42-activating protein zizimin1 (TrEMBL AAM90306; Meller et al. 2002). Zizimin1 contains a new domain named CDM (CED-5, DOCK180, MyoBlast city) zizimin homology domain 2 (CZH2) that mediates direct interaction with the Cdc42 Rho GTPase. Motif MDS00146 is included within the CZH2 domain and appears to be a submotif of CZH2 that is specific for two of the four CZH2 domain-containing protein subfamilies, namely zizimin, KIAA1395, DOCK180, and KIAA0299. Our HMM search with motif MDS00146 detected only members of the zizimin and KIAA1395 subfamilies.

Membrane Organization

In an attempt to annotate the membrane organization of entire proteomes from a range of species, we developed a computational strategy. Based on the prediction of two features, endoplasmic reticulum signal peptides (used for translocation into the secretory pathway) and membrane-spanning domains (transmembrane domains), the membrane organization of proteins can be classified. We have annotated 10 proteome databases from a range of species (see Table 3). Determination of the membrane organization of an individual protein is dependent on knowing the full-length protein open reading frame (ORF) and cannot be applied to partial protein ORFs. For each proteome dataset, we removed any readily identifiable partial protein sequences (see Table 3). As expected in both the human and mouse ENSEMBL proteome databases, significant numbers of partial ORFs were present (37% and 44% respectively.) This highlights the high level of partial protein sequences generated from predicted genes. These partial sequences would result in inaccurate predictions of their membrane organization if retained. Surprisingly, the *M. musculus* IPI proteome contained similar levels of partial sequences (45%), whereas the other proteomes analyzed contained less than 10%.

Prediction of Endoplasmic Reticulum Signal Peptides

We used two independent methods to predict endoplasmic reticulum signal peptides, Neural Networks (NN) and hidden Markov Models (HMM) methods from SignalP V2 (Nielsen and Krogh 1998). These methods were selected because they have low levels of false-negative predictions (1.0% and 1.1% respectively; Menne et al. 2000). A consensus approach was adopted. Where the two methods agreed, we annotated the protein as containing a signal peptide. When the two methods conflicted, we used a third independent signal peptide

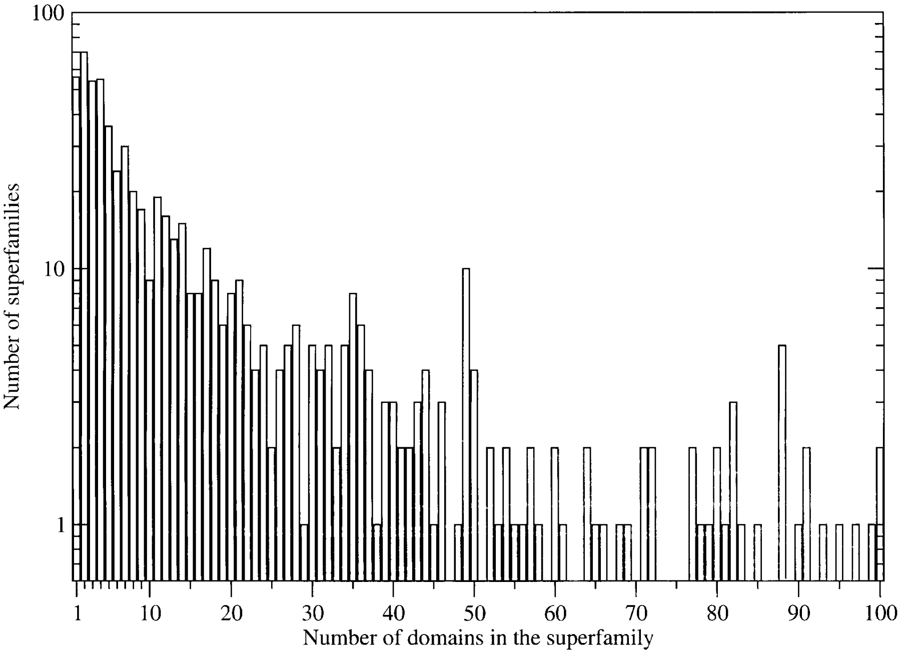

Figure 4 Observed evolutionary domains. The distribution sizes of superfamilies with less than 100 members.

Table 3. Predicted Signal Peptides and Transmembrane Domains in Eukaryotic Proteomes

		S. cerevisciae		*A. thaliana*		*D. melanogaster*		*C. elegans*		Riken RPS	
Feature	**Annotation**	**Number**	**%**	**Number**	**%**	**Number**	**%**	**Number**	**%**	**Number**	**%**
Signal Peptide predictions	Consensus annotation	481	8.0	3549	13.7	2417	17.5	3657	18.4	3461	20.1
	Non-consensus	28	0.5	234	0.9	124	0.9	178	0.9	174	1.0
	Total assigned	509	8.5	3783	14.6	2541	18.4	3835	19.3	3635	21.1
Trans-membrane predictions	Consensus annotation	1164	19.5	4789	18.5	2479	18.0	5591	28.1	3617	21.0
	Non-consensus	221	3.7	1181	4.6	497	3.6	680	3.4	676	3.9
	Total assigned	1385	23.1	5970	23.0	2976	21.6	6271	31.5	4293	24.9
Total number of proteins		6230		26,131		13,967		19,918		18,768	
Total full-length ORFs		5984		25,901		13,791		19,886		17,209	

		Riken VPS		*M. musculus* (IPI)		Mouse genome (Ensembl)		*H. sapiens* (IPI)		Human genome (Ensembl)	
Feature	**Annotation**	**Number**	**%**	**Number**	**%**	**Number**	**%**	**Number**	**%**	**Number**	**%**
Signal Peptide predictions	Consensus annotation	6630	20.4	4024	20.5	3199	20.2	4716	19.1	3494	20.4
	Non-consensus	300	0.9	177	0.9	137	0.9	253	1.0	175	1.0
	Total assigned	6930	21.4	4201	21.4	3336	21.1	4969	20.1	3669	21.5
Trans-membrane predictions	Consensus annotation	6766	20.9	4641	23.6	4641	23.5	4813	19.5	3437	20.1
	Non-consensus	1263	3.9	585	3.0	585	3.0	969	3.9	548	3.2
	Total assigned	8029	24.8	5226	26.6	5226	26.4	5782	23.4	3985	23.3
Total number of proteins		34,286		30,171		28,097		25,636		27,049	
Total full-length ORFs		32,437		19,655		15,832		24,701		17,103	

Total number of predictions of signal peptides and transmembrane domains predicted across 10 proteomes: *S. cerevisciae, A. thaliana, D. melanogaster, C. elegans,* Riken RPS, Riken VPS, *M. musculus* (IPI), Ensembl mouse genome, *H. sapiens* (IPI), and Ensembl human genome. Percentages for feature predictions are given as a percentage of the total full-length sequences. Prediction data available from http://microarray.imb.uq.edu.au/predictors/proteome/.

prediction method, SPScan (von Heijne 1987) to resolve the conflict. We considered this method suitable because of its lower false-positive rate compared to the other two methods (Menne et al. 2000). Typically less than 6% of the total number of annotated signal peptides required resolution using SPScan.

The results of this analysis using the various proteome datasets are presented in Table 3. The proportion of proteins predicted to contain signal peptides within the RIKEN RPS was 21.1%. Similar levels were annotated in the other higher eukaryotic proteomes (human and mouse). Lower proportions of signal peptides were annotated in the *D. melanogaster* (18.4%), *C. elegans* (19.3%), *A. thaliana* (14.6%), and *S. cerevisiae* (8.5%) proteome databases.

Prediction of the Membrane-Spanning Regions or Transmembrane Domains

Next we annotated the transmembrane domains for each protein. Although a consensus approach has been proposed (Nilsson et al. 2000), its application to entire genomes was not practical. To analyze all proteins, we selected two prediction methods that could be readily applied to large datasets. TMHMM 2.0, which was clearly the best performer in a recent comparative evaluation (Moller et al. 2001), was selected first. Secondly, SVMtm, a new transmembrane prediction method using a support vector machine (Z. Yuan and R. Teasdale, unpubl., http://microarray.imb.uq.edu.au/predictors/) was selected. When SVMtm was compared to TMHMM 2.0 it showed comparable accuracy levels (specificity 94.0% vs. 95.2% and sensitivity 91.8% vs. 90.8%, respectively).

Each protein within the different proteome databases was analyzed with both TMHMM 2.0 and SVMtm. Transmembrane domains were annotated when both methods positively predicted a membrane-spanning domain. Sequences containing conflicting predictions were further analyzed using three additional transmembrane prediction tools (SOSUI, HMMTOP, and MEMSAT). Only membrane-spanning regions that were positively predicted by more than two of these additional methods were annotated as transmembrane domains. Between 14% and 20% of the total number of transmembrane domains annotated were assigned by this method. In addition, an initial prediction was considered a false-positive prediction when not supported by any of the other transmembrane prediction methods. This approach is similar to the "majority vote" consensus method recently used by others (Nilsson et al. 2000; Ikeda et al. 2002).

Transmembrane domain prediction methods are known to incorrectly predict signal peptides as transmembrane domains. Therefore we adopted a filter for predicted N-terminal transmembrane segments: If the predicted transmembrane domain's starting point was within the first 15 residues of the ORF and a signal peptide was predicted, then this region was regarded as a signal peptide instead of a transmembrane domain. This filtering procedure was applied to the results of all transmembrane prediction tools. The results from this analysis using the various proteome datasets are presented in Table 3.

In contrast to the signal peptide analysis, the proportion of proteins with predicted transmembrane domains varied little between proteomes, (21.6%–26.6%), with the exception of the *C. elegans* proteome, where the proportion was higher (31.5%). These results are consistent with the similar attempts to annotate membrane-spanning domains in eukaryotes (Wallin and von Heijne 1998; Krogh et al. 2001; Liu and Rost 2001; Ward 2001). For example, using an earlier version of TMHMM, Krogh and others predicted transmembrane domains for *S. ceresiviae*, *D. melanogaster*, and *C. elegans*, at 20.7%, 20.1%, and 30.3% respectively.

Classification of Proteins Into Distinct Classes Based on Their Predicted Membrane Organization

Here we propose an alternative broad classification scheme for protein classes based on their predicted membrane organization. This approach utilizes the combined annotation within individual full-length protein ORFs of both signal peptides and transmembrane domains (see Table 4). Transmembrane-negative soluble proteins are classified as intracellular or extracellular based on the signal peptide predictions. Transmembrane-positive proteins are classified into three groups. The topology of single membrane-spanning proteins, Type I (Nout/Cin), or Type II (Nin/Cout), is assigned based on the presence or absence of a signal peptide. Proteins with more than one membrane-spanning domain are classified as multispan membrane proteins. Based on the above annotation of signal peptides and membrane-spanning regions, we obtained six groups of proteins (see Table 4).

In contrast to simply comparing the total proportion of transmembrane domains, which do not vary significantly across eukaryotic genomes, our classification scheme highlighted variation in the membrane organization of the proteome datasets analyzed. Overall, comparison of the results from the predicted membrane organization across the 10 proteome databases revealed that higher eukaryotes have a greater proportion of soluble secreted proteins and Type I

Table 4. Membrane Organization of Protein Classes Assigned Based on the Prediction of Signal Peptides and Transmembrane Domains

		S. cerevisciae		*A. thaliana*		*D. melanogaster*		*C. elegans*		Riken RPS	
Class	Description	Number	%	Number	%	Number	%	Number	%	Number	%
A	Soluble nonsecreted proteins	4107	68.6	16,467	63.6	8474	61.4	10,672	53.7	10,161	59.0
B	Soluble secreted proteins	252	4.2	2354	9.1	1815	13.2	2469	12.4	2040	11.9
C	Type I membrane proteins (single span, secreted)	117	2.0	777	3.0	335	2.4	636	3.2	935	5.4
D	Type II membrane proteins (single span, nonsecreted)	300	5.0	1542	6.0	651	4.7	1072	5.4	804	4.7
E	Multimembrane spanning proteins	820	13.7	2810	10.8	1676	12.2	3816	19.2	2096	12.2
F	Membrane organization not annotated	388	6.5	1951	7.5	840	6.1	1221	6.1	1173	6.8
Total full-length ORFs		5984		25,901		13,791		19,886		17,209	

		Riken VPS		*M. musculus* (IPI)		Mouse genome (Ensembl)		*H. sapiens* (IPI)		Human genome (Ensembl)	
Class	Description	Number	%	Number	%	Number	%	Number	%	Number	%
A	Soluble nonsecreted proteins	19,267	59.4	11,288	57.4	9117	57.6	15,057	61.0	10,336	60.4
B	Soluble secreted proteins	3862	11.9	2378	12.1	1908	12.1	2776	11.2	2119	12.4
C	Type I membrane proteins (single span, secreted)	1843	5.7	977	5.0	751	4.7	1361	5.5	864	5.1
D	Type II membrane proteins (single span, nonsecreted)	1497	4.6	779	4.0	598	3.8	953	3.9	655	3.8
E	Multimembrane spanning proteins	3839	11.8	3017	15.3	2478	15.7	2843	11.5	2030	11.9
F	Membrane organization not annotated	2129	6.6	1216	6.2	980	6.2	1711	6.9	1099	6.4
Total full-length ORFs		32,437		19,655		15,832		24,701		17,103	

Protein classes are determined by the following annotations: Class A, signal peptide (SP)-negative, transmembrane domain (TD)-negative; Class B, SP-positive, TD-negative; Class C, SP-positive, single TD-positive; Class D, SP-negative, single TD-positive; Class E, SP-positive or negative, transmembrane domains multiple positives per protein; Class F, no consensus in SP and/or TD predictions. Percentages are given as a percentage of the total full-length ORFs. Results used to compile Table 4 are available from http://microarray.imb.uq.edu.au/predictors/proteome/.

membrane proteins, whereas the proportions of Type II membrane and multi-span membrane proteins remained similar. For example, Riken RPS compared to *S. cerevisiae* had 2.8- and 2.7-fold increases in soluble secreted proteins and Type I membrane proteins, respectively, whereas the other classes of membrane proteins remained essentially unchanged. Comparison of the RPS and VPS proteomes revealed no difference in the degree of alternative splicing among the different membrane organization classes. The other result of note from this comparison, as previously observed (Krogh et al. 2001), is the higher proportion of multi-span membrane proteins in *C. elegans*.

METHODS

Databases

We analyzed the following proteome databases available from EBI on June 15th 2002 (http://www.ebi.ac.uk/proteome/; Apweiler et al. 2001): *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens* (International Protein Index, IPI), *Mus musculus* (IPI), and *Saccharomyces cerevisiae*. In addition, we analyzed the RIKEN Mouse Representative Transcript and Protein Sets (Riken RPS; http://genome.gsc.riken.go.jp/), RIKEN Mouse Variable Protein Set (Riken VPS; http://genome.gsc.riken.go.jp/), and the predicted protein ORFs from the ENSMBL human genome database (Human Build 29) and mouse genome database (MGSC Mouse Assembly 3; http://www.ensembl.org/; Hubbard et al. 2002). The sequences were filtered so that partial ORFs that did not contain a methionine at position 1 or were clearly annotated as partial or fragments were removed.

The set of predicted proteins produced by Celera (Release R13b, April 2002, http://www.celeradiscoverysystem.com/) was used as one of the whole-genome sets of computational predictions. The Celera sequencing, assembly, and transcript prediction methods are described (Mural et al. 2002). The complete Celera protein set contained 47,256 transcripts and protein sequences, corresponding to 46,941 gene predictions. Of this set, 14,994 weak-confidence predictions were excluded for compatibility with the analyses of the Ensembl set and of the Chr.16 (Mural et al. 2002). The remaining 32,262 high- to medium-confidence protein predictions were used in this study; 15,548 of these hypothetical proteins had BLAST hits to the publicly available protein sequences, 7085 were in common with the RefSeq (Pruitt and Maglott 2001) mouse protein set, and 19,089 had InterPro assignments.

InterPro version 5.1 (May 2002) and InterProScan version 3.1 (Zdobnov and Apweiler 2001) were used for the functional sites and domains composition analysis.

Prediction Methods

Signal P V2 (NN and HMM; Nielsen and Krogh 1998), SPScan (von Heijne 1987), TMHMM 2.0 (Krogh et al. 2001), SVMtm (Z. Yuan and R. Teasdale, in prep.; http://microarray.imb.uq.edu.au/predictors/), MEMSAT 1.5 (Jones et al. 1994), HMMTOP (Tusnady and Simon 2001), and SOSUI (Hirokawa et al. 1998) were applied using their default values except for selection of organism group. SPScan analysis was performed using the Genetics Computer Group (GCG) Wisconsin Package (version 8.1) located at the Australian National Genomic Information Service (ANGIS). SOSUI analysis was performed using its Web interface (http://sosui.proteome.bio.tuat.ac.jp/~sosui/proteome/welcomeE.html).

ACKNOWLEDGMENTS

We thank N. Mulder and W. Fleischmann (European Bioinformatics Institute) for consultations and support, A. Krogh and S. Brunak (Technical University of Denmark) for providing the TMHMM and SignalP software packages, G.E. Tusnady (Hungarian Academy of Sciences) for providing the HMMTOP software package, D.T. Jones (University College) for providing the MEMSAT software package, and K. Miranda (IMB, University of Queensland) for technical assistance. This work was supported in part by the Australian Research Council and the National Health and Medical Research Council of Australia; also in part by ACT-JST of JST (Japan Science and Technology Corp.), and a Grant-in-Aid for Scientific Research on Priority Areas "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan to H.M.

REFERENCES

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., et al. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29:** 37–40.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, W., Shindyalov, I.N., and Bourne, P.E. 2000. The protein data bank. *Nucleic Acids Res.* **28:** 235–242.

FANTOM Consortium 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.

Gough, J. 2002. The superfamily database in structural genomics. *Acta Crystallogr. D Biol. Crystallogr.* **58:** 1897–1900.

Gough, J. and Chothia, C. 2002. Superfamily: HMMs representing all proteins of known structure. *Nucleic Acids Res.* **30:** 268–272.

Ha, S., Lee, S., Chung, M., and Choi, Y. 2002. Mouse ING1 homologue, a protein interacting with A1, enhances cell death and is inhibited by A1 in mammary epithelial cells. *Cancer Res.* **62:** 1275–1278.

Hirokawa, T., Boon-Chieng, S., and Mitaku, S. 1998. SOSUI: Classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14:** 378–379.

Hubbard, D., Barker, E., Birney, G., Cameron, Y., Chen, L., Clark, T., Cox, J., Cuff, V., Curwen, T., Down, R., et al. 2002. The ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.

Ikeda, M., Arai, M., Lao, D.M., and Shimizu, T. 2002. Transmembrane topology prediction methods: A reassessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. *In Silico Biol.* **2:** 19–33.

Jones, D.T., Taylor, W.R., and Thornton, J.M. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33:** 3038–3049.

Kawaji, H., Schönbach, C., Matsuo, Y., Kawai, J., Okazaki, Y., Hayashizaki, Y., and Matsuda, H. 2002. Exploration of novel motifs derived from mouse cDNA sequences. *Genome Res.* **12:** 367–378.

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.

Liu, J. and Rost, B. 2001. Comparing function and structure between entire proteomes. *Protein Sci.* **10:** 1970–1979.

Meller, N., Irani-Tehrani, M., Kiosses, W.B., Del Pozo, M.A., and Schwartz, M.A. 2002. Zizimin1, a novel Cdc42 activator, reveals a new GEF domain for Rho proteins. *Nat. Cell Biol.* **4:** 639–647.

Menne, K.M.L., Hermjakob, H., and Apweiler, R. 2000. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics* **16:** 741–742.

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

Montini, E., Buchner, G., Spalluto, C., Andolfi, G., Caruso, A., den Dunnen, J.T., Trump, D., Rocchi, M., Ballabio, A., and Franco, B. 1999. Identification of SCML2, a second human gene homologous to the *Drosophila* sexcomb on midleg (Scm): A new gene cluster on Xp22. *Genomics* **58:** 65–72.

Mural, R.J., Adams, M.D., Myers, E.W., Smith, H.O., Miklos, G.L.G., Wides, R., Halpern, A., Li, P.W., Sutton, G.G., Nadeau, J., et al. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296:** 1661–1671.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Nielsen, H. and Krogh, A. 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6:** 122–130.

Nilsson, J., Persson, B., and von Heijne, G. 2000. Consensus predictions of membrane protein topology. *FEBS Lett.* **486:** 267–269.

Poirier, G.M., Anderson, G., Huvar, A., Wagaman, P.C., Shuttleworth, J., Jenkinson, E., Jackson, M.R., Peterson, P.A., and Erlander, M.G. 1999. Immune-associated nucleotide-1 (IAN-1) is a thymic selection marker and defines a novel gene family conserved in plants. *J. Immunol.* **163:** 4960–4969

Pruitt, K.D. and Maglott, D.R., 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.

Suzuki, T., Nakamoto, T., Ogawa, S., Seo, S., Matsumura, T., Tachibana, K., Morimoto, C., and Hirai, H. 2002. MICAL, a novel CasL interacting molecule, associates with vimentin. *J. Biol. Chem.* **277:** 14933—14941.

Terman, J.R., Mao, T., Pasterkamp, R.J., Yu, H.H., and Kolodkin, A.L. 2002. MICALs, a family of conserved flavoprotein oxidoreductases, function in plexin-mediated axonal repulsion. *Cell* **109:** 887–900.

Tusnady, G.E. and Simon, I. 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17:** 849–850.

von Heijne, G. 1987. *Sequence analysis in molecular biology: Treasure trove or trivial pursuit*, pp. 1–188. Academic Press, San Diego, CA.

Wallin, E. and von Heijne, G. 1998. Genome-wide analysis of integral membrane proteins. *Protein Sci.* **7:** 1029–1038.

Ward, J.M. 2001. Identification of novel families of membrane proteins from the model plant *Arabidopsis thaliana. Bioinformatics* **17:** 560–563.

Zdobnov, E.M. and Apweiler, R. 2001. InterProScan—An integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17:** 847–848.

WEB SITE REFERENCES

http://www.celeradiscoverysystem.com; Celera Corporation.

http://www.ensembl.org/; ENSEMBL genome databases.

http://www.ebi.ac.uk/IPI; International Protein Index.

http://www.ebi.ac.uk/proteome; Proteome Analysis Database.

http://genome.gsc.riken.go.jp/; RIKEN Mouse Representative Transcript and Protein Sets.

http://sosui.proteome.bio.tuat.ac.jp/~sosui/proteome/welcomeE.html; SOSUI Web Interface.

http://supfam.org/SUPERFAMILY/cgi-bin/gen_list.cgi?genome=mr; SUPERFAMILY Database.

http://supfam.org/FANTOM2/domcombs.html; SUPERFAMILY FANTOM2 data.

http://microarray.imb.uq.edu.au/predictors; SVMtm Support Vector Machines to predict transmembrane domains.

http://motif.ics.es.osaka-u.ac.jp/fantom2/; MDS Motif Database for FANTOM2.

http://microarray.imb.uq.edu.au/predictors/proteome/; SRC microarray facility—Proteome supplementary material.

Received November 11, 2002; accepted in revised form March 5, 2003.