FEBS Letters 555 (2003) 572–578 FEBS 27916

Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains

Yi Xing, Qiang Xu, Christopher Lee*

Molecular Biology Institute, UCLA Institute for Genomics and Proteomics, Department of Chemistry and Biochemistry, University of California at Los Angeles, Los Angeles, CA 90095-1570, USA

Received 15 September 2003; accepted 11 November 2003

First published online 27 November 2003

Edited by Robert B. Russell

Abstract We have investigated the effects of alternative splicing on transcripts encoding membrane proteins in 1001 human genes. Out of a total of 464 alternatively spliced genes encoding single-pass transmembrane (TM) proteins, in 188 we observed a splice form that specifically removed the TM domain, producing a soluble protein isoform. For example, in syndecan-4, the new alternative splice form closely parallels the proteolytic ectodomain shedding previously shown in this protein, and recognized as an important regulatory mechanism of receptor function. While many of the soluble isoforms produced by alternative splicing have already been validated, most are novel, and in 57 genes showed a statistically significant association (*P*-value < 0.01) with a specific tissue.

Key words: Alternative splicing; Isoform; Transmembrane protein; Proteolysis; Bioinformatics

1. Introduction

Alternative splicing is a well-known mechanism for modulating gene function in different tissues and cellular states. Previously believed to occur in only 5–15% of human genes, it has recently been found to be widespread in the human and other genomes, and according to studies of expressed sequence tags (ESTs) is observed in around 30–60% of human genes [1–7]. Moreover, more than 80% of the alternative splice forms identified in EST data appear to be novel; that is, they have not previously been reported in existing mRNA sequence databases (for example, in GenBank). Therefore identification and analysis of these alternatively spliced isoforms have received great interest in the past several years [8–11], in particular their impact on protein function [12], for example, globular protein domains [13].

One major question about alternative splicing is its impact on membrane protein function. It is estimated that 20–30% of all genes in most genomes encode integral membrane proteins [14–16]. Most of the integral membrane proteins contain membrane-spanning hydrophobic α-helices which are inserted into the membrane lipid bilayer [17] and are generally classified into single-pass and multi-pass membrane proteins according to their number of transmembrane (TM) segments. They play a variety of crucial biological roles including signal transduction, transportation, energy metabolism and cell adhesion [18–21], participate in many important forms of regulation, and are the major targets of therapeutic drugs [22–24].

Extensive biochemical studies have demonstrated that one way to modulate membrane protein function is to release such a protein from its membrane anchorage, turning it into a soluble form [25]. Proteolytic cleavage has been shown to be a widespread mechanism for generating soluble fragments of membrane proteins [26,27]. Cleavage can occur in the extracellular region (termed 'ectodomain shedding'), releasing it from the membrane [28], or in the intracellular region (termed regulated intracellular proteolysis, RIP) [29]. These two processes are known to affect a wide range of membrane proteins such as Notch [30], Erb-B4 [31], and TGF-α [32]. Almost all types of membrane proteins have known examples of these proteolytic regulatory events [28]. The resulting soluble protein fragments can act as diffusible signals that agonize or antagonize the function of the membrane-bound form, or can transmit a signal to another location [33–35].

In this paper we analyze the effects of alternative splicing on membrane protein anchoring. There is some precedent for considering this to be a biologically interesting question. There are several examples of known membrane-anchored proteins (CD46, cadherin-7, and IL4 receptor) where it has been shown that alternative splicing removes the exonic sequence coding the TM domain, generating a soluble form of the protein [36–38] that functions differently from the membrane-bound form.

To analyze the effects of alternative splicing on membrane anchoring in a large sample of the human proteome, we constructed a database of alternatively spliced protein isoforms (ASP) based on our analysis of human expressed sequences, generating a collection of 13 384 isoforms for 4422 human genes [39,40]. We used this database to identify genes where alternative splicing removed a putative single-pass TM domain, producing a soluble isoform of a protein that is normally membrane-anchored. Our results demonstrate that this form of regulation is widespread, is present in a surprisingly large fraction of membrane proteins, and appears to be analogous to proteolytic cleavage in its functional impact. This database should be of interest to researchers studying regulation of membrane protein function, and hopefully will stimulate many new experimental studies of these novel protein isoforms.

2. Materials and methods

We identified alternative splicing events in human genes as previ-

*Corresponding author. Fax: (1)-310-267 0248.
E-mail address: leec@mbi.ucla.edu (C. Lee).

0014-5793/03/$22.00 © 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)01354-1

ously described [6], using NCBI's January 2002 draft human genomic sequence [5] and UniGene human EST data [41] downloaded in January 2002. Individual exons were identified by their start and end positions in the genomic sequence; individual splices were identified by their 5′ and 3′ splice site positions in the genomic sequence. A transcript was defined as a list of exons and a list of associated splices connecting them. The major transcript was identified as the transcript with largest number of consistent expressed sequence evidence. The longest open reading frame (ORF) was identified in each candidate transcript.

The set of transcripts was filtered by a variety of criteria: (1) Only major–minor isoform pairs resulting in a change to the protein product were retained. (2) No transcripts incorporating non-consensus splice sites were permitted. (3) The transcript's longest ORF must be full-length, i.e. begin with AUG and end with a STOP codon. (4) Minor-form protein products must have at least 50% identity to the major form (that is, no more than half of the major-form protein sequence can be changed or removed), and a minimum length of 50 amino acids. This produced a subset (ASP database) of 13 384 protein isoforms in 4422 human genes. Potential non-sense-mediated decay (NMD) targets in ASP were identified by checking for a STOP codon located over 50 bp upstream of the last exon–exon junction [42]. Full details and validation results of the isoform generation procedure have been presented elsewhere [40].

We have also performed a number of tests to assess the transcript sequences for the 188 genes described in this paper. First, we tested whether the region predicted to be a TM domain genuinely exists in a human-curated protein sequence from the SwissProt database. Out of 73 cases where our data could be compared with a SwissProt entry with the same gene symbol, in 64 cases our protein sequence was an exact match to the SwissProt entry, and in six cases our sequence was somewhat shorter than the SwissProt sequence, typically due to lack of EST coverage over the full length of the gene. However, in all six of these cases the region predicted to be a TM domain was matched by the SwissProt sequence. Finally, in three cases the TM isoform was novel. Thus, the relevant protein sequence region was validated in 70 of 73 test cases (96%). Next, we assessed whether the TM domain was really removed by alternative splicing, using a conservative criterion that at least 70% of the amino acids composing the TM domain must be deleted by an observed alternative splice. Out of the 188 cases, 159 deleted at least this fraction of the residues of the TM region. Thus, the data indicate that in at least 85% of the cases we report, alternative splicing genuinely removes the possibility of a TM region from the protein sequence.

We used the program TMHMM to identify putative TM regions in ASP protein isoforms. TMHMM uses a hidden-Markov model approach to predict TM protein topology, and is evaluated as having the best overall performance by several recent independent tests [16,43–45]. Protein isoform sequences were submitted to the THHMM 2.0 server, and putative TM regions in each isoform were identified. Using this approach, we identified the set of genes encoding single-pass TM proteins (group A), and the subset that also had alternative splice forms lacking a TM region (group B).

We tested the accuracy of TMHMM for a random sample of 25 single-pass membrane proteins and a random sample of 25 soluble proteins in human from SwissProt. Among the 25 single-pass membrane proteins, TMHMM missed the TM region in two proteins and predicted an additional TM region in one protein, yielding the correct result in 22 single-pass membrane proteins out of the 25 proteins tested (88%). The boundaries predicted by TMHMM agree well with the human-curated TM annotations in SwissProt (see Appendix 2, table B1). Among the 25 soluble proteins, TMHMM made a single false positive prediction (Appendix 2, table B2), yielding an accuracy of 96%.

We used Gene Ontology (downloaded in August 2002) to check whether the production of soluble forms by alternative splicing is more frequent in certain type of alternatively spliced TM proteins, using a similar approach as described before [46,47]. We compared group B with group A to look for GO keywords t found at much higher frequency in group B than group A. We define:

N = total number of genes in group A annotated by GO

n = number of genes in group A annotated by GO keyword t

M = total number of genes in group B annotated by GO

m = number of genes in group B annotated by GO keyword t

and calculated the enrichment ratio of GO keyword t in group B, and its P-value according to the hypergeometric distribution:

$$\text{Enrichment ratio} = \frac{m/M}{n/N}$$

$$P\text{-value} = \sum_{k=m}^{n} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

If the enrichment ratio is less than 1, then the P-value is calculated as:

$$P\text{-value} = \sum_{k=0}^{m} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

3. Results and discussion

3.1. Identification of alternative splice forms altering membrane anchoring

We searched for TM domains in protein isoform sequences using the program TMHMM [16,43–45]. TMHMM identified TM domains in 1001 of the 4422 human genes in our dataset (22.6%). In 533 of these genes (53.2%) alternative splicing changed the number of TM domains (Fig. 1). Out of a total of 464 genes coding single-pass membrane proteins (with a single TM domain) in our dataset, 188 had isoforms that produced a soluble protein isoform (no predicted TM domain according to TMHMM; we will refer to this as 'TM removal'). Thus alternative splicing frequently converted single-pass TM proteins to soluble forms (40.5% of the cases in our dataset). In 117 out of the 188 UniGene clusters (62.2%) TM regions are encoded by a cassette exon which can be removed by alternative splicing. Among genes with multiple evidence for each isoform (that is, each alternative splice form is supported by an mRNA or at least two ESTs), we found soluble isoforms in 130 genes out of 342 alternatively spliced single-TM genes (38.0%)

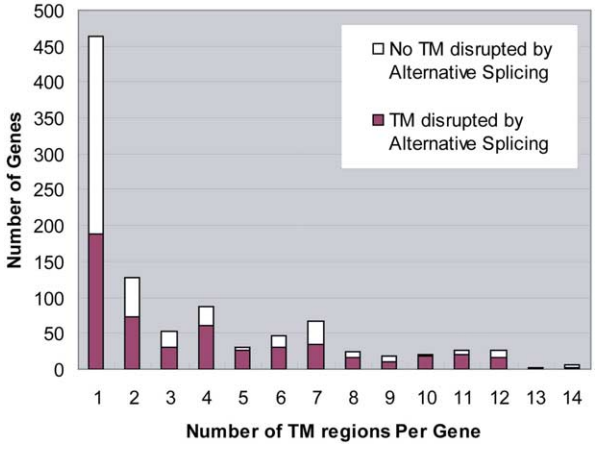

Fig. 1. Removal of TM domains by alternative splicing. A histogram of the number of genes in the ASP dataset encoding 1, 2 or more predicted TM domains. The filled portion of each bar (red) indicates the number of genes in which a TM domain was removed by alternative splicing.

Removal of TM domains by alternative splicing occurred more frequently than expected by random chance. Defining 'TM deletion' as meaning that at least 70% of the residues of the TM domain were deleted by alternative splicing, we observed TM deletion in 159 of the 464 single-TM genes in our dataset (34%). By contrast, under a random model (in which TM domains were repositioned randomly in the protein sequence), only 98 TM deletions occurred, indicating that the observed rate of TM deletion is significantly higher than random (P-value of 5×10^{-6} by Fisher's Exact Test), consistent with a similar finding for globular protein domains [13]. In 103 genes (64.8%), the region removed by alternative splicing extended not further than 80 amino acids from the plasma membrane, including 60 (37.7%) within 40 amino acids to the plasma membrane.

In many cases, the TM removal identified by our analysis from EST data can be directly validated by human-curated mRNA sequences from GenBank. For 57 of these genes (30.3% of the 188 single-TM genes in our dataset), we found full-length mRNA sequences in GenBank supporting both the membrane-anchored isoform and also the soluble isoform. We also performed extensive validation of our protein isoform sequences and TM domain identification and removal (see Section 2).

Consistent with previous genomics studies of alternative splicing, the majority of alternative splice forms we identified appear to be novel [1–7]; that is, they have not previously been reported in any mRNA deposited in GenBank. For the 166 genes in which at least one of the membrane-anchored form or soluble form has full-length mRNA evidence, the soluble form is novel in 90 genes, but the membrane-anchored form is novel in only 19 genes.

3.2. *Alternative splicing of syndecan-4 mimics proteolytic ectodomain shedding*

Our analysis also identified many novel soluble isoforms of known TM proteins with interesting functional implications. For example, we identified a novel isoform of syndecan-4, supported by three ESTs (Fig. 2). Syndecan-4 belongs to a family of TM heparan sulfate proteoglycans (HSPGs). It has the ability to bind to a variety of ligands including extracellular matrix components, growth factors, cell adhesion molecules, cytokines, proteinases, etc. [48–50]. Syndecan-4 has been shown to undergo ectodomain shedding. The proteolytic cleavage at the extracellular region of syndecan-4 releases a soluble form of syndecan-4 from the membrane which retains its ligand binding ability [51,52]. The ectodomain of syndecans has been suggested to play a role in a variety of biological processes. Ectodomain shedding of syndecan-1 and -4 is accelerated during injury and inflammation, generating a soluble fragment which can facilitate signaling by acting as a co-receptor; the soluble fragment is also reported to antagonize or agonize FGF-2 in different situations [28]. While the full-length syndecan-4 mRNA has five exons [53] and most of the ESTs in its UniGene cluster match this canonical gene structure, we detected an alternative splice in three ESTs (Hs#S3789360 from epithelioid carcinoma cell line; Hs#S541656 from pregnant uterus; Hs#S604747 from pooled human melanocyte, fetal heart, and pregnant uterus) which splice from the fourth exon to a new exon located between the fourth exon and last exon in syndecan-4 mRNA. The new exon (which we call exon V-b) contains an in-frame STOP codon. The new syndecan-4 isoform supported by three ESTs truncates the protein's C-terminus, removing the TM domain coded by the last exon (exon V-a in Fig. 2). It is striking that the alternative splice form of syndecan-4 encodes a soluble isoform that is closely analogous to the soluble form of syndecan-4 known to be produced by ectodomain shedding [51,52].

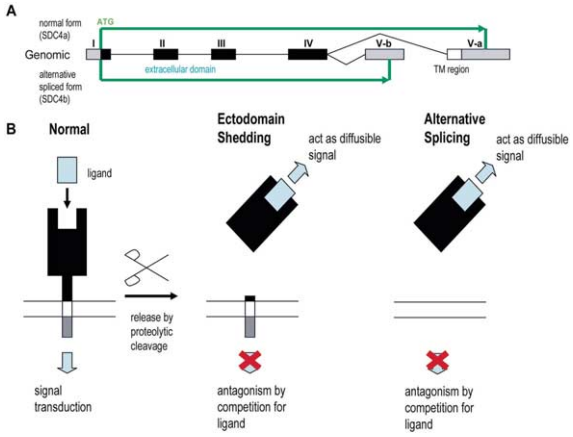

Fig. 2. Production of a soluble protein isoform of syndecan-4 by alternative splicing. A: Gene structure of syndecan-4. Exons are shown as boxes. The membrane-bound form (SDC4a) uses exon V-a as the last exon, which codes for the TM domain (white). The novel soluble form (SDC4b) uses exon V-b as the last exon. B: Schematic models of syndecan-4 TM regulation. In the normal, full-length form of syndecan-4, ligand binding to the extracellular domain leads to signal transduction by the cytoplasmic domain. In ectodomain shedding, proteolytic cleavage of syndecan-4 releases the ligand binding region from the membrane, which can act as a diffusible signal or antagonize normal syndecan-4 signaling. Alternative splicing of syndecan-4 appears to produce a similar soluble form, which could have similar functional impacts as ectodomain shedding.

3.3. *Functional impact of TM domain removal by alternative splicing*

TM removal by alternative splicing showed a notable asymmetry in our dataset: membrane-anchored forms were more likely to be the more common, ubiquitous isoform, while the soluble forms were often localized to a specific tissue. Using previously described methods for calculating the statistical confidence of apparent tissue specificity [54], we evaluated both the membrane-bound and soluble isoforms of all 188 genes in our TM removal set. In 67 genes (36%) one or both of the isoforms showed significant tissue specificity (P-value < 0.01), suggesting that regulation of membrane-anchoring is an important category of tissue-specific alternative splicing. Moreover, in 57 out of these 67 genes (85.1%), the soluble isoform was tissue-specific by these criteria, whereas the membrane-bound isoform was tissue-specific in only 29 genes (43.3%). For example, we identified a soluble isoform for *HLA-DMB* (Hs.1162) which in the EST data appears to be placenta-specific (see Table 1). It should be noted that these rates of observable tissue specificity probably underestimate the true extent of this phenomenon, since most genes lack sufficient EST sequences (from enough tissues) to detect tissue specificity. Our analysis probably does not detect most real tissue-specific splicing patterns, simply because of inadequate EST coverage [55].

What membrane protein functions are specifically modu-

Table 1
20 examples of genes where alternative splicing generates soluble protein isoforms

UniGene Id	Gene description	Gene name	Membrane-bound form has mRNA evidence	Soluble form has mRNA evidence
Hs.1162	major histocompatibility complex, class II, DM beta	HLA-DMB	+	–
Hs.129708	tumor necrosis factor (ligand) superfamily, member 14	TNFSF14	+	+
Hs.1311	CD1C antigen, c polypeptide	CD1C	+	–
Hs.13225	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 4	B4GALT4	+	–
Hs.159428	BCL2-associated X protein	BAX	+	+
Hs.167246	P450 (cytochrome) oxidoreductase	POR	+	–
Hs.173936	interleukin 10 receptor, beta	IL10RB	+	–
Hs.180338	tumor necrosis factor receptor superfamily, member 12 (translocating chain-association membrane protein)	TNFRSF12	+	+
Hs.181097	tumor necrosis factor (ligand) superfamily, member 4 (tax-transcriptionally activated glycoprotein 1, 34 kDa)	TNFSF4	+	–
Hs.212680	tumor necrosis factor receptor superfamily, member 18	TNFRSF18	+	+
Hs.2175	colony-stimulating factor 3 receptor (granulocyte)	CSF3R	+	+
Hs.252189	syndecan-4 (amphiglycan, ryudocan)	SDC4	+	–
Hs.271411	beta-site APP-cleaving enzyme 2	BACE2	+	+
Hs.305890	BCL2-like 1	BCL2L1	+	–
Hs.66	interleukin 1 receptor-like 1	IL1RL1	+	+
Hs.77572	BCL2/adenovirus E1B 19 kDa-interacting protein 1	BNIP1	+	–
Hs.79187	coxsackie virus and adenovirus receptor	CXADR	+	–
Hs.83532	membrane cofactor protein (CD46, trophoblast–lymphocyte cross-reactive antigen)	MCP	+	+
Hs.86386	myeloid cell leukemia sequence 1 (BCL2-related)	MCL1	+	+
Hs.93213	BCL2-antagonist/killer 1	BAK1	+	–

Description of GO category	Number of Genes annotated in TM-removing list	Number of Genes annotated in total TM list	enrichment ratio	p value
cell death(GO:0008219)	18.1%(13)	8.5%(17)	↑ 2.1	4.70E-04
death(GO:0016265)	18.1%(13)	8.5%(17)	↑ 2.1	4.70E-04
apoptosis(GO:0006915)	16.7%(12)	8.0%(16)	↑ 2.1	1.10E-03
programmed cell death(GO:001250)	16.7%(12)	8.0%(16)	↑ 2.1	1.10E-03
response to external stimulus(GO:0009605)	30.6%(22)	22.0%(44)	↑ 1.4	2.30E-02
obsolete(GO:0008369)	6.9%(5)	13.5%(27)	↓ 1.9	3.10E-02
transferase(GO:0016740)	5.6%(4)	12.0%(24)	↓ 2.2	2.60E-02
cell adhesion(GO:0007155)	5.6%(4)	13.0%(26)	↓ 2.3	1.30E-02

Fig. 3. Gene Ontology keywords which are significantly enriched or infrequent among genes in which alternative splicing removed a single TM domain, as compared to the set of all single-TM genes in which alternative splicing was observed.

lated by TM domain removal by alternative splicing? To answer this question, we analyzed our 188 candidate gene list using Gene Ontology (see Section 2 for details). We looked for GO keywords in the categories of *biological-process* and *molecular-function* which showed significantly higher representation in the set of 188 single-TM genes where alternative splicing removed membrane anchoring than in a control set consisting of all single-TM genes in ASP (464 genes) (Fig. 3). Since the control set also consists entirely of genes encoding TM proteins, this controls for the possibility that a GO term would receive a strong *P*-value simply because it is associated with membrane proteins. Five GO gene categories were enriched in the TM removal set with high significance ($P < 0.02$, number of hits in the whole set no less than 15), including: *cell death*; *death*; *apoptosis*; *programmed cell death*, *response to external stimulus*. Several examples are shown in Table 1, including BCL2-associated X protein (*BAX*), BCL2-antagonist/killer 1 (*BAK1*), BCL2-like 1, *CD38*, and myeloid cell leukemia sequence 1 (*MCL1*).

Our results have interesting implications for many experimental studies of the regulation of membrane protein function. Release of soluble forms of membrane proteins by proteolysis has been shown to be important to the function of many membrane proteins [25–35]. Since our data indicate that alternative splicing produces a similar effect in a large number of membrane proteins (188 found in this study, out of 464 single-pass membrane proteins included in our dataset), we propose that this phenomenon of 'TM domain splicing' may be another common switching mechanism of regulation of membrane protein function that deserves further experimental study. The functional consequences of such alternative splicing events can be either gain of function – by releasing functional fragments from the membrane (such as syndecan-4), mimicking membrane protein proteolysis such as protein ectodomain shedding or RIP – or simply loss of function due to disruption of membrane anchorage (such as HLA-DMB). It is striking that in some cases (e.g. syndecan-4), the production of a soluble isoform by alternative splicing matches the previously known *proteolytic* ectodomain shedding event for that protein.

As an example of how our database can suggest interesting directions for new experiments, Table 1 lists a number of genes that display TM domain splicing. Full lists of our results will be made available on the web upon publication (http://www.bioinformatics.ucla.edu/ASAP; for a short validation, see Appendix 1). In most of these cases, the soluble isoform produced by alternative splicing is novel (that is, it has not previously been reported by an mRNA deposited in GenBank). It has been shown in many cases that releasing protein domains from membrane anchorage can produce pos-

itive or negative regulators of the intact membrane-bound protein forms, or generate protein forms with new function [33–35]. In some cases it is already possible to infer the functional effect of our novel forms. For example, we have identified a novel isoform of *HLA-DMB* lacking its TM anchoring region. Coincidentally, it has been reported that an engineered form of *HLA-DMB* lacking the TM domain significantly reduces the rate of antigen processing [56]. However, in most cases new experiments will be required, and in our view the main value of our database is that it offers experimentalists many new specific directions for research. It should be useful for biologists interested in these receptors' roles (e.g. in growth control and apoptosis) to design experiments testing for the presence of these novel protein isoforms and their functional impact.

In cases where a protein has not previously been shown to be a membrane protein, biologists should test this experimentally, since our results are based on a prediction method, TMHMM. This program's accuracy in predicting TM domains has been reported to be 95.5% [44]. Our independent tests of TMHMM yielded an accuracy rate of 88% on known membrane proteins and 96% on known soluble proteins, and validated the accuracy of TMHMM's prediction of the boundaries of TM regions (see Section 2). When we tested a different TM prediction method, SOSUI [57], to cross-validate the results from TMHMM, we obtained very similar results (the SOSUI dataset indicated conversion of a membrane protein into a soluble form by alternative splicing in 37.2% of the genes, versus 40.5% in the TMHMM dataset). Overall, these data indicate that the pattern of widespread production of soluble forms through alternative splicing observed in our dataset is a confident result.

Acknowledgements: We wish to thank Drs. L. Iruela-Arispe, J. Bowie, A. van der Bliek, D. Papazian, A. Resch, M. Roy for their discussions and comments on this work. C.J.L. was supported by NIMH/NINDS Grant MH65166, and DOE grant DEFG0387ER60615.

Appendix 1.

Validation of 10 randomly selected genes in our 188 gene list

Table A1
Validation of 10 randomly selected UniGene clusters

UniGene ID	Gene description	Membrane-bound form evidence	Soluble-form evidence	NMD target	Matching GenBank/ SwissProt entry
Hs.12330	ectonucleoside triphosphate diphosphohydrolase 6	mRNA	mRNA	NO	Match
Hs.158200	EGF-like domain, multiple	mRNA	EST, single	NO	Shorter in N terminus
Hs.159428	BCL2-associated X protein	mRNA	mRNA	NO	YES
Hs.173936	interleukin 10 receptor, beta	mRNA	EST, single	NO	Longer in N terminus
Hs.180338	tumor necrosis factor receptor superfamily, member 25	mRNA	mRNA	NO	Match
Hs.181244	HLA-A	mRNA	EST, multiple	NO	Shorter in N terminus
Hs.190488	hypothetical protein DKFZp66M2411	EST, multiple	EST, single	soluble form is potential NMD target	N/A
Hs.2	*N*-acetyltransferase 2	mRNA	EST, single	NO	Match
Hs.2175	colony-stimulating factor 3 receptor	mRNA	mRNA	NO	Match
Hs.25887	SEMA4	mRNA	EST, single	soluble form is potential NMD target	Match

Appendix 2.

Test of TMHMM against SwissProt human-curated annotations of TM regions

Table B1
Single-pass TM proteins from SwissProt

Entry name	SwissProt annotation	TMHMM prediction	Overlap
1A01_HUMAN	309–332	308–330	22/24
AD02_HUMAN	687–707	689–711	19/21
APP1_HUMAN	581–603	581–603	23/23
ATRN_HUMAN	1280–1300	1279–1301	21/21
DSCA_HUMAN	1596–1616	1595–1617	21/21
FCAR_HUMAN	*228–246*	*soluble*	*N/A*
HB24_HUMAN	231–251	231–251	21/21
GAMP_HUMAN	631–653	631–653	23/23
GPBA_HUMAN	506–526	507–529	20/21
GS28_HUMAN	230–250	230–249	20/21
HATT_HUMAN	21–41	21–43	21/21
IPL2_HUMAN	355–375	357–379	19/21
IR18_HUMAN	330–350	331–353	20/21
K2S5_HUMAN	246–264	243–265	19/19
MCL1_HUMAN	330–349	327–349	20/20
NM16_HUMAN	565–585	565–587	21/21
MUC1_HUMAN	1159–1181	1159–1181	23/23
OX2G_HUMAN	*233–259*	*13–35, 237–259*	*N/A*
PIGL_HUMAN	2–22	5–22	18/21
STM2_HUMAN	*218–235*	*soluble*	*N/A*
SGCE_HUMAN	294–314	293–315	21/21
SILL_HUMAN	482–502	480–502	21/21
TTD1_HUMAN	266–288	266–288	23/23
XEDA_HUMAN	139–159	140–162	20/21
ZAN_HUMAN	2758–2778	2756–2778	21/21

Table B2
Soluble proteins from SwissProt

Entry name	TMHMM prediction
143B_HUMAN	soluble
2ABG_HUMAN	soluble
3BP1_HUMAN	soluble
8ODP_HUMAN	soluble
AGP2_HUMAN	soluble
AIM1_HUMAN	soloble
AMRP_HUMAN	*7–29*
BPL1_HUMAN	soluble
BRC1_HUMAN	soluble
CALI_HUMAN	soluble
CAN1_HUMAN	soluble
CENE_HUMAN	soluble
COXB_HUMAN	soluble
CUL2_HUMAN	soluble
DLX5_HUMAN	soluble
DPYD_HUMAN	soluble
ECT2_HUMAN	soluble
FXDL_HUMAN	soluble
GRWD_HUMAN	soluble
HM21_HUMAN	soluble
I5P2_HUMAN	soluble
K1CS_HUMAN	soluble
LMG3_HUMAN	soluble
MLL4_HUMAN	soluble
NEK1_HUMAN	soluble
OSF1_HUMAN	soluble

References

[1] Mironov, A.A., Fickett, J.W. and Gelfand, M.S. (1999) Genome Res. 9, 1288–1293.
[2] Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J. and Bork, P. (2000) FEBS Lett. 474, 83–86.
[3] Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P. and Mattick, J.S. (2000) Nat. Genet. 24, 340–341.
[4] Kan, Z., Rouchka, E.C., Gish, W.R. and States, D.J. (2001) Genome Res. 11, 889–900.
[5] Int. Human Genome Sequence Consortium (2001) Nature 409, 860–921.
[6] Modrek, B., Resch, A., Grasso, C. and Lee, C. (2001) Nucleic Acids Res. 29, 2850–2859.
[7] Modrek, B. and Lee, C. (2002) Nat. Genet. 30, 13–19.
[8] Boue, S., Vingron, M., Kriventseva, E. and Koch, I. (2002) Bioinformatics 18, S65–S73.
[9] Heber, S., Alekseyev, M., Sze, S.H., Tang, H. and Pevzner, P.A. (2002) Bioinformatics 18, S181–S188.
[10] Kan, Z., States, D. and Gish, W. (2002) Genome Res. 12, 1837–1845.
[11] Lewis, B.P., Green, R.E. and Brenner, S.E. (2003) Proc. Natl. Acad. Sci. USA 100, 189–192.
[12] Black, D.L. (2000) Cell 103, 367–370.
[13] Kriventseva, E.V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M.S. and Sunyaev, S. (2003) Trends Genet. 19, 124–128.
[14] Wallin, E. and von Heijne, G. (1998) Protein Sci. 7, 1029–1038.
[15] Stevens, T.J. and Arkin, I.T. (2000) Proteins 39, 417–420.
[16] Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) J. Mol. Biol. 305, 567–580.
[17] White, S.H. and Wimley, W.C. (1999) Annu. Rev. Biophys. Biomol. Struct. 28, 319–365.
[18] Medzhitov, R. (2001) Nat. Rev. Immunol. 1, 135–145.
[19] Hubner, C.A. and Jentsch, T.J. (2002) Hum. Mol. Genet. 11, 2435–2445.

[20] Schwartz, M.A. and Ginsberg, M.H. (2002) Nat. Cell Biol. 4, E65–E68.
[21] Ponta, H., Sherman, L. and Herrlich, P.A. (2003) Natl. Rev. Mol. Cell Biol. 4, 33–45.
[22] Raymond, V. and Sattelle, D.B. (2002) Natl. Rev. Drug Discov. 1, 427–436.
[23] Shawver, L.K., Slamon, D. and Ullrich, A. (2002) Cancer Cell 1, 117–123.
[24] Zuany-Amorim, C., Hastewell, J. and Walker, C. (2002) Natl. Rev. Drug Discov. 1, 797–807.
[25] Rose-John, S. and Heinrich, P.C. (1994) Biochem. J. 300, 281–290.
[26] Blobel, C.P. (2000) Curr. Opin. Cell Biol. 12, 606–612.
[27] Fortini, M.E. (2002) Natl. Rev. Mol. Cell. Biol 3, 673–684.
[28] Arribas, J. and Borroto, A. (2002) Chem. Rev. 102, 4627–4638.
[29] Mayer, R.J. (2000) Natl. Rev. Mol. Cell Biol. 1, 145–148.
[30] Chan, Y.M. and Jan, Y.N. (1998) Cell 94, 423–426.
[31] Ni, C.Y., Murphy, M.P., Golde, T.E. and Carpenter, G. (2001) Science 294, 2179–2181.
[32] Massague, J. and Pandiella, A. (1993) Annu. Rev. Biochem. 62, 515–541.
[33] Maliszewski, C.R. and Fanslow, W.C. (1990) Trends Biotechnol. 8, 324–329.
[34] Hoppe, T., Rape, M. and Jentsch, S. (2001) Curr. Opin. Cell Biol. 13, 344–348.
[35] Vermes, C., Jacobs, J.J., Zhang, J., Firneisz, G., Roebuck, K.A. and Glant, T.T. (2002) J. Biol. Chem. 277, 16879–16887.
[36] Mosley, B. et al. (1989) Cell 59, 335–348.
[37] Tsujimura, A. et al. (2001) J. Biochem. (Tokyo) 130, 841–848.
[38] Kawano, R., Matsuo, N., Tanaka, H., Nasu, M., Yoshioka, H. and Shirabe, K. (2002) J. Biol. Chem. 277, 47679–47685.
[39] Lee, C. (2003) Bioinformatics 19, 999–1008.
[40] Xing, Y., Resch, A. and Lee, C. submitted.
[41] Schuler, G. (1997) J. Mol. Med. 75, 694–698.
[42] Maquat, L.E. (2002) Curr. Biol. 12, R196–R197.
[43] Sonnhammer, E.L., von Heijne, G. and Krogh, A. (1998) ISMB 6, 175–182.
[44] Moller, S., Croning, M.D. and Apweiler, R. (2001) Bioinformatics 17, 646–653.
[45] Melen, K., Krogh, A. and von Heijne, G. (2003) J. Mol. Biol. 327, 735–744.
[46] Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C. and Krawetz, S.A. (2003) Genomics 81, 98–104.
[47] Zeeberg, B.R. et al. (2003) Genome Biol. 4, R28.
[48] Bernfield, M., Kokenyesi, R., Kato, M., Hinkes, M.T., Spring, J., Gallo, R.L. and Lose, E.J. (1992) Annu. Rev. Cell Biol. 8, 365–393.
[49] Bernfield, M., Gotte, M., Park, P.W., Reizes, O., Fitzgerald, M.L., Lincecum, J. and Zako, M. (1999) Annu. Rev. Biochem. 68, 729–777.
[50] Carey, D.J. (1997) Biochem. J. 327, 1–16.
[51] Subramanian, S.V., Fitzgerald, M.L. and Bernfield, M. (1997) J. Biol. Chem. 272, 14713–14720.
[52] Fitzgerald, M.L., Wang, Z., Park, P.W., Murphy, G. and Bernfield, M. (2000) J. Cell Biol. 148, 811–824.
[53] Takagi, A., Kojima, T., Tsuzuki, S., Katsumi, A., Yamazaki, T., Sugiura, I., Hamaguchi, M. and Saito, H. (1996) J. Biochem. (Tokyo) 119, 979–984.
[54] Xu, Q. and Lee, C. (2003) Nucleic Acids Res. 31, 5635–5643.
[55] Xu, Q., Modrek, B. and Lee, C. (2002) Nucleic Acids Res. 30, 3754–3766.
[56] Weber, D.A., Dao, C.T., Jun, J., Wigal, J.L. and Jensen, P.E. (2001) J. Immunol. 167, 5167–5174.
[57] Hirokawa, T., Boon-Chieng, S. and Mitaku, S. (1998) Bioinformatics 14, 378–379.