Nucleic Acids Research, 2005, Vol. 33, No. 8 **2355–2363**
doi:10.1093/nar/gki520

Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals

Mitsuteru Nakao[1,2], Roberto A. Barrero[3], Yuri Mukai[2], Chie Motono[2], Makiko Suwa[2] and Kenta Nakai[1,*]

[1]Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan, [2]Computational Biology Research Center, National Institute of Advanced Industry Science and Technology, Tokyo, Japan and [3]Center for Information Biology and DNA Data Bank Japan, National Institute of Genetics, Shizuoka, Japan

Received November 18, 2004; Revised December 28, 2004; Accepted March 29, 2005

ABSTRACT

We investigated human alternative protein isoforms of >2600 genes based on full-length cDNA clones and SwissProt. We classified the isoforms and examined their co-occurrence for each gene. Further, we investigated potential relationships between these changes and differential subcellular localization. The two most abundant patterns were the one with different C-terminal regions and the one with an internal insertion, which together account for 43% of the total. Although changes of the N-terminal region are less common than those of the C-terminal region, extension of the C-terminal region is much less common than that of the N-terminal region, probably because of the difficulty of removing stop codons in one isoform. We also found that there are some frequently used combinations of co-occurrence in alternative isoforms. We interpret this as evidence that there is some structural relationship which produces a repertoire of isoformal patterns. Finally, many terminal changes are predicted to cause differential subcellular localization, especially in targeting either peroxisomes or mitochondria. Our study sheds new light on the enrichment of the human proteome through alternative splicing and related events. Our database of alternative protein isoforms is available through the internet.

INTRODUCTION

In higher eukaryotic cells, the repertoire of gene sets coded in their genomes is greatly expanded through transcriptional and post-transcriptional events, such as alternative splicing (1,2), alternative promoter usage (3) and alternative polyadenylation (4,5). Recent studies have estimated that 40–79% of multi-exon human genes produce a set of different mRNAs (6–10). Such diversity is regarded as a key concept in explaining how vertebrate cells attain a high functional complexity from a relatively small number of genes. In this variation of alternative isoforms, not only the coding regions but also untranslated regions can be affected.

Alternative isoforms have been implicated to play some role in a number of cellular processes from development (11) to pathology (12). They are also important in modulating temporal molecular function and spatial subcellular localization (13) (differential subcellular localization; DSL) of gene products. These functional changes are typically caused by insertions and/or deletions of sequence segments corresponding to either functional domains, subcellular sorting signals or transmembrane regions (14). For example, antigen genes *CD-8*α and β code two protein isoforms, membrane-bound and secreted ones, respectively, through alternative splicing (15). In another example, *Bfl-1*, an apoptotic Bcl-2 family member, codes two alternative protein isoforms, a long form (*Bfl-1*) and a short form (*Bfl-1S*), in which their localization is switched between the nucleus and the mitochondrial membrane due to the gain/loss of a nuclear localization signal (NLS) or a transmembrane helix (TMH) by a frameshift (Figure 1) (16).

Although recent studies have revealed the effect of functional domains upon alternative splicing (14,17,18), these mainly were focused on the analysis around alternative splice sites and, to our knowledge, there has been no comprehensive analysis that clarifies the diversity of alternative isoforms in terms of their amino acid sequences. In this study, we constructed a comprehensive dataset of human alternative protein isoforms that are the union of two sequence sets: a set of translated sequences of full-length cDNAs collected by the 'H-Invitational' international project (19) and a set of variant

*To whom correspondence should be addressed. Tel: +81 3 5449 5131; Fax: +81 3 5449 5133; Email: knakai@ims.u-tokyo.ac.jp

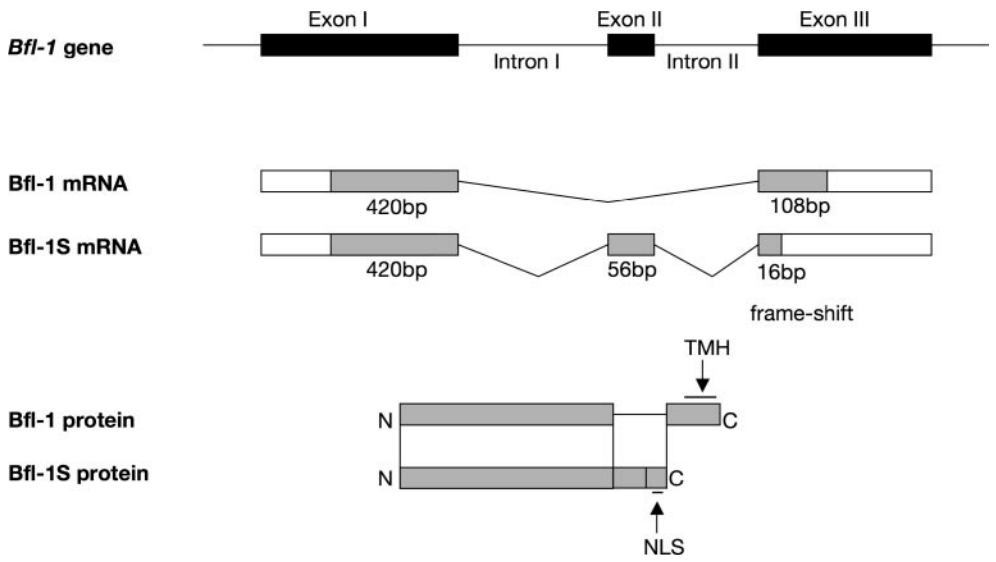


Figure 1. Example of differential subcellular localization by alternative protein isoforms. An example of the *BFl-1* gene products (16). In an isoform *Bfl-1S*, an exon is inserted and the coding region, shown in gray, in the last exon is shortened because of a frame-shift. This causes a loss of the transmembrane helix (TMH) as well as the creation of a nuclear localization signal (NLS) in this variant.

protein sequences stored in SwissProt (20). Next, we classified the structural diversity of these variants based on our novel coding system of sequence changes, where the amino acid sequences of each pair of variants are globally aligned and their terminal variation patterns are mainly examined. A severe restriction by the translation mechanism was observed in the occurrence of a specific isoformal pattern. The co-occurrence of coded patterns in each gene was also analyzed. Lastly, the potential effects of alternative isoforms on the gain/loss of sorting signals (DSL) were explored.

MATERIALS AND METHODS

Construction of a large-scale dataset of alternative protein isoforms

We constructed a database of alternative isoforms from two sources: 42 149 human full-length cDNA sequences from the H-Invitational project (19) and 13 745 sequences from the VARSPLIC collection of SwissProt release 41 (20–22). For the H-invitational data, we restricted our study to 8553 sequences of 3181 genes showing alternative splicing (AS unique isoforms). This selection was carried out with the combination of a computer program that detects alternative splicing variants and visual inspection by experts in the H-invitational project (19). The AS unique isoforms set was further screened on the condition that the translated sequences should start from methionine and that there should be at least two isoforms.

From the VARSPLIC collection, we selected human sequences with at least two isoforms: 1487 genes (4133) remained. To remove the redundancy between the two datasets, sequences with >98.0% identity (100% coverage) were detected using the BLASTP program (23). As a result, we obtained a set of 2624 human genes corresponding to 6876 isoforms.

Recognition method for alternative isoform pattern

We adopted a novel method for classifying protein isoforms. The patterns were defined in two ways: one pattern defined from the pair-wise alignments between the amino acid sequences of isoformal clones (pair-wise statistics) and one defined for each locus (locus-wise statistics). The latter was defined based on the former: we aligned all pairs of isoforms for each locus and assigned a pattern for each pair-wise alignment.

For each pair of clones, a global alignment was constructed using local alignments made with the (ungapped) BLASTP program. To remove the effect of simple sequencing errors, mismatches between amino acid pairs made by 1 nt change were regarded as identical. We defined the aligned segments of identical residues as 'blocks'. The obtained alignments were divided into at most three regions based on the status of the blocks (Figure 2); the N-terminal block, for example, may take one of the three types of states: the 'A(ligned)' state which means the pair shares an identical N-terminal side; the 'I(nsertion)' state which means one of the pair has an N-terminal extension; and the 'M(utually-exclusive)' state which means both sequences have unaligned N-terminal extensions. The states of the C-terminal block are defined in a similar way. The remaining internal region, if it exists, has similar 'A', 'I' and 'M' flags if they have an (identically) matched segment, an inserted segment or an unaligned segment, respectively. Note that the internal region may take these three states simultaneously because this region may be translated from multiple exons. Sometimes, the 'A' state for this region is hard to define because there is no boundary between the state and the terminal block(s). In other words, some simple patterns contain both terminal blocks only and lack internal blocks (e.g. Pattern #1 '___.A___.____.___M'). Therefore, we cannot directly compare the occurrences of types between internal regions and terminal regions.

More specifically, locally identical regions were enumerated as candidates of A-type blocks. These regions were obtained as

'HSPs' of 'BLASTP' with the unitary scoring matrix and without gaps and filters. We set the minimal length of these HSPs to six residues after several trials (note that this length is not the minimum block length; see Discussion). According to our observation, this value was successful for most cases but there remains some problems for proteins with repetitive sequences, which were handled manually. In the next step, those A-block candidates were extended in both directions by regarding two amino acids that are interchangeable with a single base substitution as identical. This condition was set so that the presence of simple sequencing errors and/or polymorphisms should not strongly affect classification. The remaining regions in each sequence were treated as candidates of I-type blocks. Next, the candidate blocks were aligned with the two sequences. In the case of multiple possibilities, the longer blocks were selected. Two successive I-type blocks across two sequences were interpreted as the M-type. All pair-wise alignments were checked by visual inspection. In the above alignment algorithm, the most important parameter was the initial minimal block size. When we made this value smaller, the average block size tended to become smaller due to the influence of repeats.

There were also significant numbers of identical pairs and totally unaligned pairs at the amino acid level designated as 'F' and 'f', respectively. Thus, each pair is characterized with the combination of 11 flags (`Ff.AIM.AIM.AIM`) for the flag positions: global appearance, N-terminal region, Internal region and C-terminal region as shown in Table 1.

RESULTS

Pattern analysis of alternative protein isoforms

As described in Materials and Methods, we constructed a database of alternative protein isoforms that contain 6876 isoforms of 2624 genes (loci). As shown in Figure 2, the two source databases are quite complementary. There are 2.62 variants per gene on average. The dataset is accessible through the Internet (http://www.jbirc.aist.go.jp/wgi/).

For each gene, each pair of its isoforms was globally aligned and the pattern of their difference was identified. To describe such patterns, we developed a coding method as explained in the Materials and Methods section. Its main idea is to identify the variation from the positions of (un)aligned blocks and their alignment states. Unless a pair of sequences are totally identical or different, which is often the case, there should be at least one aligned region without gaps ('block') as well as one unaligned block. The positions of these blocks are classified into one of the three categories: the N-terminal region, the C-terminal region and the internal region (if any). For each position of the block, one of the following alignment states is assigned: 'aligned' (A-type), 'inserted' (I-type) or 'mutually exclusive' (M-type, indicating that the sequences are totally different within this block) (see Figure 3). In total, there are 39 possible patterns of variations, including the cases where two sequences are entirely identical or different. In our dataset, the two most complicated patterns were missing (Table 1).

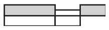
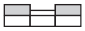
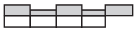
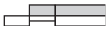
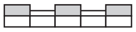
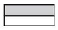
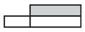
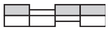
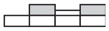
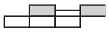
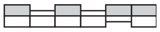
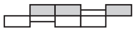
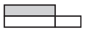
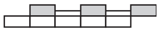
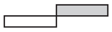
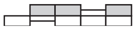
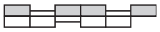
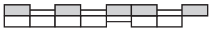
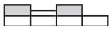
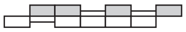
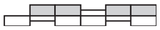
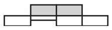
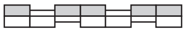
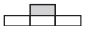
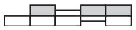
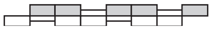
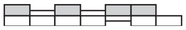
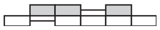
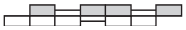
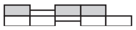
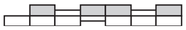
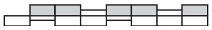
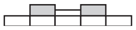
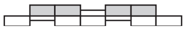
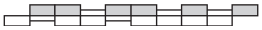
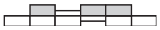
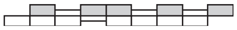
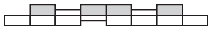
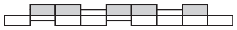
In Table 1, the patterns of variations are listed along with the ascending order of their frequency. The two most abundant patterns, the M-type at the C-terminus (Rank #1: 23%) and the I-type at the internal region (Rank #2: 20%), represent almost half of the total. In our notation, these patterns are represented as '`__.A__.___.__M`' and '`__.A__._I_.A__`', respectively (see Materials and Methods). The cumulative ratio amounts to over 80% if we consider the 10 most common patterns (see column 4 of Table 1). It is clear that relatively simple patterns occur most frequently. Since the above 'pair-wise' statistics tends to exaggerate the effects of genes that have many isoforms, we also counted the number of genes having each of the pair-wise pattern ('locus-wise' statistics; columns 5 and 6 of Table 1). In this statistic, the overall tendency remains unchanged though two exceptional patterns ('totally identical' and 'different') are frequent. Column 7 ('Avg') of Table 1 shows the average frequency of each pattern per gene. Patterns #9 (`__._I_.AI_.A__`), #14 (`__._I_.AI_.__M`) and #19 (`__.A__.AI_._I_`) show relatively large average values, reflecting that their occurrences are biased to specific genes. On the other hand, the reason why Pattern #2 occupies the top in the locus-wise statistics is that it occurs in more genes than Pattern #1.

Both the pair-wise and locus-wise frequencies of each alignment state in each position are summarized into a cross table (Figure 3). It is evident that the C-terminal region is more variable than the N-terminal region. At the N-terminal region, the I-type and the M-type occur with almost the same frequency. On the other hand, the M-type is much more abundant than the I-type in the C-terminal region. These tendencies are also present in another dataset collected from RefSeq (24) human REVIEWED entries (data not shown).

To see the degree of amino acid changes, we calculated the length distribution of blocks on both terminal regions. The total average length of affected terminal sequences is 147 amino acids (aa), which is 25% of the average entire sequence length (581 aa). If there were some difference between the length distribution of A-type blocks on both sides, it would suggest a bias on the positions of internal changes. However, there was no significant difference between them (Figure 4a). As for the difference in the length distribution of M-type blocks, C-terminal blocks are clearly longer (Figure 4b). In addition, the distributions of the I-type blocks are almost the same on both sides (data not shown). Therefore, since the distributions of the I-type and the M-type are almost the same on the N-terminal region (Figure 4c), it appears that the extension of C-terminal M-type occurs rather easily.

Though we set the minimum HSP length as six residues in defining the isoform patterns, the length of derived blocks is sometimes very short. For example, there are 153 cases where the C-terminal M-type blocks are only one-residue in length. Such small blocks could easily be caused by sequencing errors, polymorphisms or corrupted data. However, after a more detailed analysis, we found that most, if not all, are not artifacts. In the above 153 cases, 126 overlap a splice site, where a stop codon of the short block is located at the second codon position of the corresponding exon. Another 20 of them were created by so-called intron retention. In these cases also, stop codons were located at the position corresponding to the second position of the in-frame codons. As for the remaining cases, we could not find clear explanations for their creation. Although we admit some possibility of artifacts, we prefer not to set a minimum length threshold to avoid missing such interesting cases.

Table 1. Patterns of alternative protein isoforms

ID	Pair-wise Rank	Fraq	%	Locus-wise Rank	Fraq	Avg	Isoform pattern Diagram	Code
#1	1	1898	23	2	833	2.3		A__.__.__M
#2	2	1627	43	1	912	1.8		A__.__I__.A__
#3	3	511	49	7	185	2.8		A__.AI__.__M
#4	4	501	55	4	290	1.7		__M.__.A__
#5	5	489	61	8	182	2.7		A__.AI__.A__
#6	6	485	67	3	362	1.3		Identical
#7	7	391	72	5	224	1.7		__I__.__.A__
#8	8	351	76	6	221	1.6		A__.__M.A__
#9	9	274	79	14	68	4.0		__I__.AI__.A__
#10	10	198	82	12	74	2.7		__I__.A__.__M
#11	11	195	84	13	73	2.7		A__.AIM.A__
#12	12	194	87	11	97	2.0		__M.A__.__M
#13	13	181	89	10	103	1.8		A__.__.__I__
#14	14	180	91	18	20	9.0		__I__.AI__.__M
#15	15	166	93	9	105	1.6		Totally different
#16	16	104	94	16	50	2.1		__M.AI__.A__
#17	17	102	95	15	62	1.6		A__.A__M.__M
#18	18	67	96	17	39	1.7		A__.AIM.__M
#19		67	97	23	16	4.2		A__.AI__.__I__
#20	20	36	97	20	18	2.0		__M.AI__.__M
#21	21	34	98	20	18	1.9		__M.A__M.A__
#22	22	28	98	18	20	1.4		__M.A__.__I__
#23	23	23	99	22	17	1.4		A__.A__M.A__
#24	24	21	99	23	16	1.3		__I__.A__.__I__
#25	25	20	99	27	11	1.8		__I__.A__M.A__
#26	26	16	99	23	16	1.0		__M.A__M.__M
#27	27	13	99	26	13	1.0		A__.AIM.__I__
#28	28	10	99	28	10	1.0		__M.AI__.__I__
#29		10	100	28	10	1.0		__I__.A__M.__M
#30	30	9	100	30	9	1.0		A__.A__M.__I__
#31	31	7	100	31	7	1.0		__I__.AIM.A__
#32	32	6	100	32	6	1.0		__M.AIM.A__
#33	33	4	100	33	4	1.0		__I__.AI__.__I__
#34	34	2	100	34	2	1.0		__M.A__M.__I__
#35		2	100	34	2	1.0		__.M.AIM.__M
#36	36	1	100	36	1	1.0		__I__.A__M.__I__
#37		1	100	36	1	1.0		__I__.AIM.__M
#38	38	0	100	38	0			__I__.AIM.__I__
#39		0	100	38	0			__M.AIM.__I__

The 'ID' column represents the ID of patterns; 'Rank' is the rank of its pair-wise and locus-wise frequency respectively. '%' is the cumulative ratio of the pattern frequency in the pair-wise statistics. The 'Avg' is the ratio of the two frequency statistics, which corresponds to the redundancy of pattern occurrences in each gene locus. The pattern codes are defined in Materials and Methods.

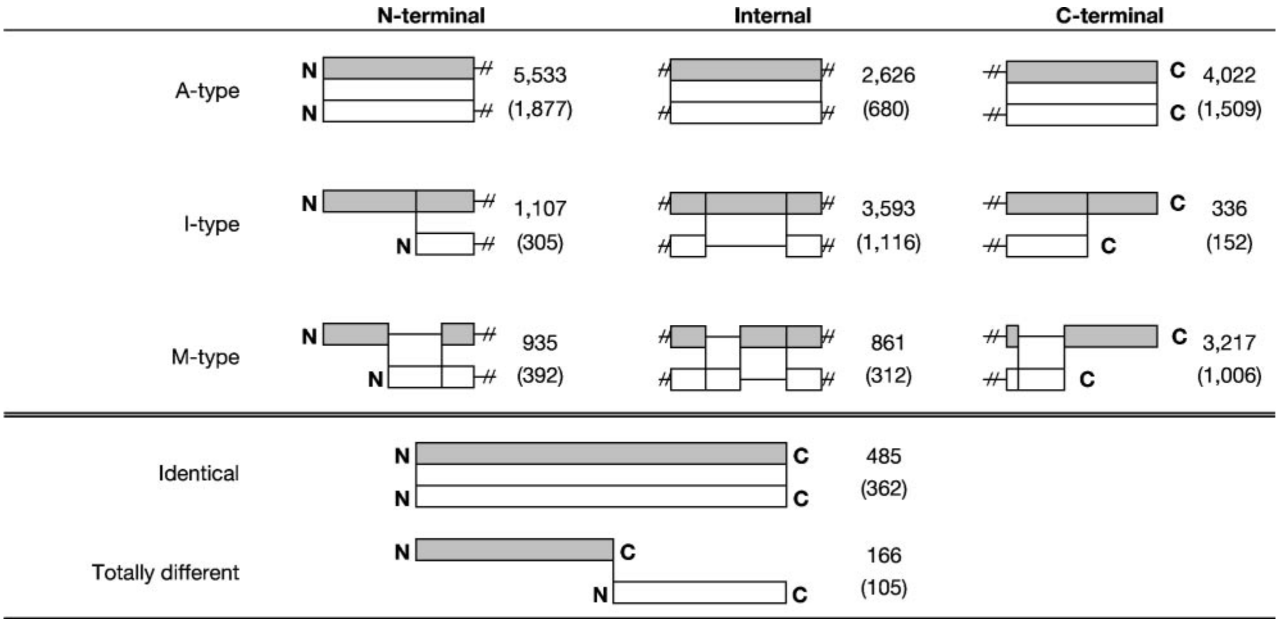


Figure 2. Statistics of alternative protein isoforms. Each isoform pair is classified in terms of the positions of the differences (N-terminal, internal and C-terminal) and the alignment states (A-type, I-type and M-type) except for the cases of totally identical or different (see Materials and Methods). The frequency of each type is shown pair-wise and locus-wise (in parenthesis). Since not all alignments have an internal region, the sum of the frequencies in that column is less than that of the two terminal columns.

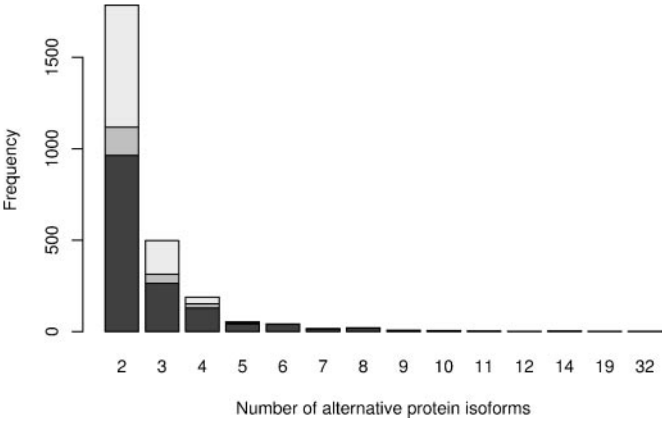


Figure 3. Distribution of the number of isoforms for each gene. The numbers of isoforms obtained from the H-invitational dataset only is shown in white, those from SwissProt only in black and those from both sources in gray, respectively.

Classification of alternative protein isoforms

Next, we tried to identify frequent combinations of isoform patterns in genes. For this purpose, we selected the combinations that occur significantly more frequently than expected by chance. The obtained frequent combinations were classified into seven classes rather arbitrarily: (i) internal, (ii) C-terminal, (iii) N-terminal, (iv) bi-terminal, (v) identical, (vi) totally different and (vii) miscellaneous ones (Table 2).

Internal. For over a quarter of the gene set, isoform variation is restricted to the internal region (i.e. their terminal regions are unchanged). The combination of patterns #2 and #5 was the most common in this class, which occurs in most cases (146 out of 182 genes) in which the gene shows Pattern #5; in other words, when there is a variant that has two insertions, the gene is very likely to have another variant with a single insertion, as well. We found that the combination #2-#5-#8-#11 (13 genes) is also frequent.

C-terminal. The class with changes in the C-terminal region includes a third. A combination of three Patterns #1-#2-#3 appears in 97 genes. About a half (97/182) of the genes having Pattern #3 show this combination. Thus, the occurrence of Pattern #3 is correlated with the occurrences of Patterns #1 and #2. There is another frequent combination #1-#8-#17, in which the internal I-type change is combined with the C-terminal M-type.

N-terminal. In this class of 381 genes (16%), their C-terminal region is constitutive (unchanged). Combinations #2-#4-#16 and #2-#7-#9 are typical variants.

Bi-terminal. A minor fraction (3%) of genes include variants where both of their terminal regions are changed simultaneously.

Identical and totally different. In many cases, genes have a variant with totally identical or different amino acid sequence (Patterns #6 and #15) and do not have other types of variants.

Miscellaneous. The rest (144 genes; 6%) show other miscellaneous combinations. Combinations #1-#4-#12 and #1-#7-#10 are relatively frequent. In most cases, genes in this group seem to have relatively complex gene structures.

In summary, it seems that Patterns #1 (C-terminal M-type) and #2 (internal I-type) work as a seed while Patterns #4 (N-terminal M-type) and #7 (N-terminal I-type) work as additive factors. For example, #1-#4-#12 (38 genes) and #1-#7-#10 (39 genes) contain basic combinations of #1-#4 and #1-#7. Combinations #2-#4-#16 (34 genes) and #2-#7-#9 (37 genes)

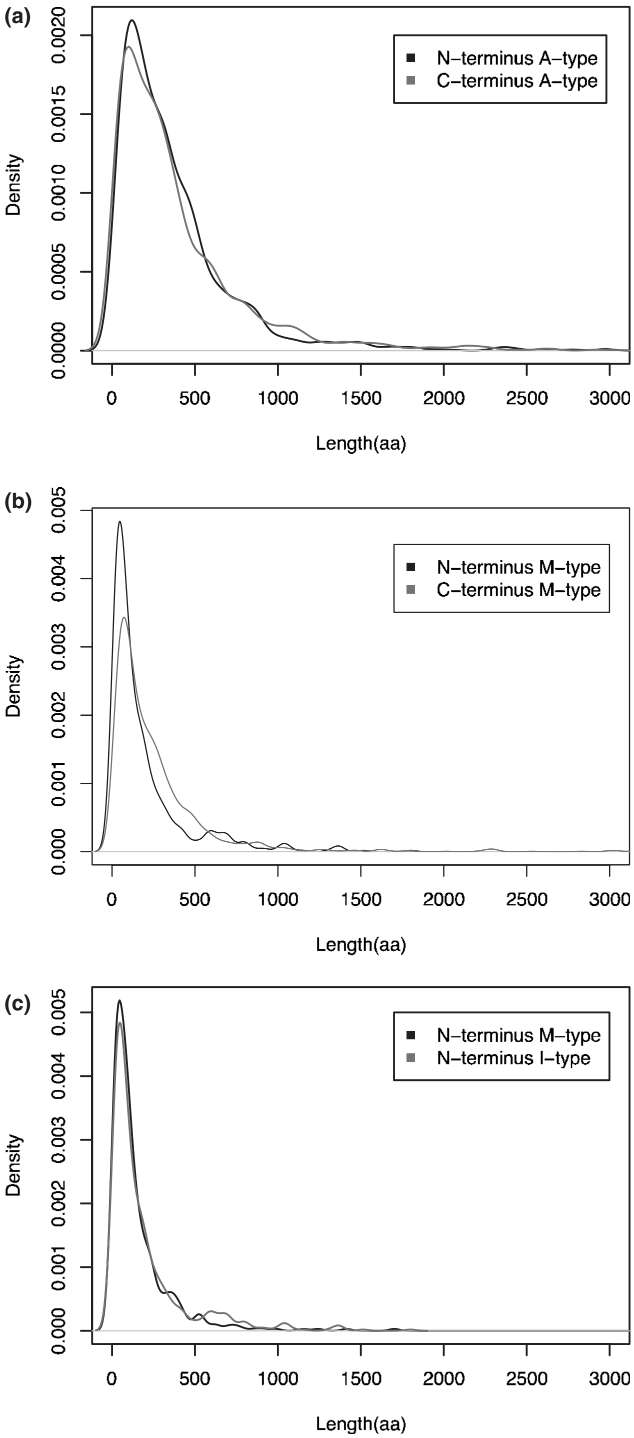
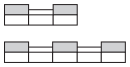
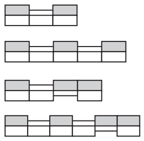
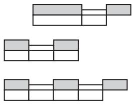
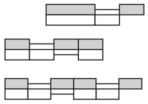
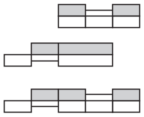
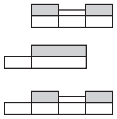
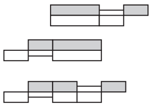
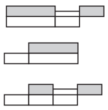


Figure 4. Length distribution of aligned blocks. (**a**) Lengths of the A-type blocks. Black line shows the density of length of N-terminal region while gray line shows that of C-terminal region. (**b**) Lengths of the M-type blocks. Black line: N-terminal region; gray line: C-terminal region. Values for blocks longer than 3000 residues are not shown. (**c**) Lengths of N-terminal blocks: black line are the M-type blocks while gray line are the I-type blocks. One long block that is >3000 residues is not shown.

Table 2. Classification of the combination of isoforms appeared in the same gene

Class	Number of genes	Examples	Genes families
Internal (constitutive both sides)	739 (31%)	#2-#5  #2-#5-#8-#11 	Ikaros ELAV Hu-antigen Plakophilin Clathlin light chain
C-terminal (constitutive N-terminal)	801 (34%)	#1-#2-#3  #1-#8-#17 	DnaJ 2′–5′ Oligodenylate sythetase Integrin alpha precursor Splicing factor
N-terminal (constitutive C-terminal)	381 (16%)	#2-#4-#16  #2-#7-#9 	Apolipoprotein L1, L3, L4 Wnt Carnitine O-acetyltransferase Endotheline-converting enzymes C-X-C chemokine receptor Retinoic acid receptor
Biterminal	76 (3%)		Solute carrier family
Identical	214 (9%)		
Totally different	33 (1.4%)		
Miscellaneous	144 (6%)	#1-#4-#12  #1-#7-#10 	Amyloid beta A4 Nuclear Factor 1 Tropomysin

For each class, typical combinations are shown with examples of gene families. A complete list of the family genes is available from the Supplementary web Material.

are also interpreted to be derived from basic combinations, #2-#4 and #2-#7. Except for totally identical or different patterns, all combinations contain either #1 or #2. In addition, we observed that members of the same gene family often belong to one of the above classes. For example, members of the DNA binding Ikaros family, the Plakophilin family and the ELAV Hu-antigen family belong to the 'internal' group (Table 2); while members of the SFR splicing factor family, the Dual-specificity tyrosine kinase CLK family and the Paired box protein Pax family (1, 2, 3, 4, 7 and 8) tend to belong to the 'C-terminal' class. A complete list of this protein family analysis based on the Ensembl Protein Family is available on the web material (http://www.jbirc.aist.go.jp/wgi/).

Analysis of subcellular localization signals located on the alternative terminals

Since many subcellular localization signals reside on the terminal regions of proteins, it is of special interest to analyze the possibility of how alternative isoforms can modulate subcellular localization sites. By using TargetP (25), we predicted the gain/loss of N-terminal mitochondrial targeting signals (mTPs) and signal peptides (SPs) with alternative variations. Sixty-three genes (108 pairs) were predicted to show the gain/loss of mTP while 51 genes were predicted to be unchanged (Table 3). There is not a significant difference between the frequencies of the I-type and the M-type. It appears contradictory that nine A-type genes are also predicted to show differential localization because mTP resides on the N-terminus of proteins. One possible explanation is that internal I-/M-types are likely to affect the targeting signal when the N-terminal A-type blocks are short (i.e. <15 residues). All pairs that were predicted to be unchanged with respect to mitochondrial localization were either A-type or identical type.

When the presence of SPs is predicted by TargetP, 124 genes (214 pairs) were predicted to be affected. The changes were significantly more frequent in the M-type variants than the I-type ones ($P < 3.93e-3$) (Table 3).

The peroxisomal targeting signal type 1 (PTS1) is interesting in that it exists on the C-termini of proteins (26). By using the PTS1 program (26,27) to predict the presence of this signal, 61 genes (102 pairs) were predicted to be subject to DSL. In most cases, these changes were predicted on the M-type isoforms (87 pairs for the M-type while 7 pairs for the I-type). No A-type isoforms were predicted to have DSL.

Table 3. Potential relationship between subcellular localization signals and terminal variation patterns

Subcellular localization signal	Block type	Signal changed Pairs (loci)	Signal unchanged Pairs (loci)
Peroxisomal targeting signal type 1 (PTS1) (C-terminus)	A-type	0 (0)	97 (35)
	I-type	7 (5)	0 (0)
	M-type	87 (50)	0 (0)
	Identical	0 (0)	6 (5)
	T.D.	8 (6)	8 (6)
	Total	102 (61)	111 (48)
Mitochondrial targeting peptide (mTP) (N-terminus)	A-type	16 (9)	118 (44)
	I-type	35 (18)	0 (0)
	M-type	47 (29)	0 (0)
	Identical	0 (0)	10 (7)
	T.D.	10 (7)	0 (0)
	Total	108 (63)	128 (51)
Signal peptide (SP) (N-terminus)	A-type	54 (18)	1406 (367)
	I-type	49 (28)	9 (8)
	M-type	82 (57)	29 (16)
	Identical	0 (0)	66 (38)
	T.D.	29 (21)	7 (3)
	Total	214 (124)	1517 (432)
TMH (TMHMM)		326 (183)	1551 (484)
TMH (SOSUI)		523 (246)	1779 (578)

Predicted frequency of the gain/loss of various subcellular localization signals (including the change of the number of TMH) is classified with the types of isoforms. 'T.D.' means 'totally different'. For the abbreviation of each localization signal, see text.

Lastly, the presence of transmembrane helices (TMH) was predicted using the TMHMM program (28). It was predicted that 27% (=183/667) of transmembrane proteins change the number of transmembrane helices in alternative isoforms. This tendency was also seen in predictions made by the SOSUI program (29) (Table 3).

DISCUSSION

To our knowledge, this study is the first large-scale analysis of the variation patterns of amino acid sequences in alternative protein isoforms. Since our method uses only the information of amino acid sequences, we could combine two different sources: the H-invitational full-length cDNA annotation (19) and the SwissProt amino acid sequence database (20). In principle, no information on the genome sequence is necessary in our analysis. By combining two sources of data, we can use the maximal set of obtainable full-length sequences. Moreover, we expect that this will minimize the effect of inherent biases included in each data source. Namely, the SwissProt VARSPLIC is constructed from a compilation of a number of works. Thus, the variants of well-studied proteins are likely to be included more frequently. In contrast, the H-invitational set was obtained from systematic studies but the clones that have the same 5′-end sequence as previously determined ones were often discarded. Therefore, it is likely that the variants around their 3′ ends are less frequently sequenced than those around their 5′ ends. Nevertheless, we observed an opposite tendency, which suggests that our observation is biologically significant.

We believe that the overall tendency we observed will also hold in the entire human proteome for the following three reasons: (i) The top two patterns occupy the majority in Table 1; such a conspicuous tendency is not likely to be reversed. (ii) The same result was obtained when we used the RefSeq-based data in terms of the tendencies we observed (For example, the rank correlation coefficient between the frequency of 36 patterns between these two kinds of data was 0.92). (iii) As discussed earlier, the combined use of two data sources is expected to reduce the effect of their inherent biases.

There are, however, two potential problems: it is possible that some alternative isoforms are stored in separate entries of SwissProt since it is constructed independently of the genome data. We checked the positions on the human genome for each SwissProt entry using the Ensembl database (30). Seven genes seem to have multiple entries. We leave them in our dataset because many of them do not have further evidence of their origin and besides, they do not significantly affect our conclusions. Another potential problem is that we ignored the cases where a novel isoform appears as the single sequence of a gene in one data source but does not appear in the other. Considering the risk of accidentally merging paralogous sequences, we chose to ignore these sequences. The protein-based approach implemented in this study is complementary to previous methods that are based on the comparison of messenger alternative splicing patterns. One of the advantages of our analysis is that the changes of structure/function of genes are directly ascribed to their amino acid sequences; it is hard to grasp the entire effect of frame-shifts at the nucleotide level.

Structural diversity of alternative protein isoforms is non-random. The diversity of alternative protein isoforms seems to be biased in many aspects: pair-wise, only 10 patterns, which are relatively simple in structure, cover over 80% of the total dataset (Table 1); The average frequency of some patterns (e.g. #14, #19 and #9) per gene is very high, which suggests that these patterns occur quite often in a relatively small number of genes (Table 1); Variations are disproportionally frequent in the C-terminal region and the nature of the alignment states is also different between the two terminal regions (Figure 3; discussed below).

Our result shows that the majority of observed isoform patterns are made from at most a few selection events of alternative promoters, alternative splicing or alternative polyA-addition. Moreover, we found that some isoform patterns are inter-related; that is, some combinations of them are frequently observed in many genes (Table 3).

Translation mechanism leads to disproportionate variation at the C-terminal side

The C-terminal side turned out to be more variable (Figure 3) as well as longer (Figure 4b) than the N-terminal side. This is plausible because changes in the N-terminal region are more harmful in that these may lead to substantial changes of amino acid sequences due to early frame-shifts. Non-functional isoforms may have been prevented evolutionarily.

It is also notable that the M-type (unaligned blocks) is 10 times more frequent than the I-type (simple extension) in the C-terminal region (Figure 3). This also makes sense in light of the fact that the formation of a C-terminal extension without changing the original sequence is difficult because the stop codon must not remain as an internal codon. One way in which such a situation might arise is if a splice junction is located just on or at the upstream neighbor of the stop codon; in such an isoform, the stop codon is not included in the extension.

On the contrary, there is no strong need to remove the start codon in the N-terminal region because it can be used as an internal codon. Thus, this may explain why the frequencies of the I-type and the M-type are almost the same in this region.

Terminal variation and differential localization

We predicted a number of genes that are subject to DSL. These data will be a useful resource for further experimental verification.

Predictive studies suggest that the change of N-terminal regions is often associated with the DSL into mitochondria while that of C-terminal regions is often associated with the DSL into peroxisomes (Table 3). However, the association was not so strong between N-terminal changes and the gain/loss of signal peptides, probably because their structure is more insensitive to small sequence changes. Other possible explanations are that the isoformal changes may modulate more subtle fates of proteins because there are several subtypes of signal peptides that the TargetP program cannot discriminate or that signal peptides are only used in the first step of the vesicular sorting system (31).

It is generally assumed that sequence similarity correlates with functional similarity of proteins. Moreover, homology-based methods are regarded as one of the most reliable ways to predict the subcellular localization of proteins. Since our result suggest that there are plenty of genes whose isoform change subcellular localization, we should be more careful when predicting protein function/localization from sequence similarity.

Concluding remarks

Based on two sources, we constructed a large database of protein alternative isoforms that contain over 2600 genes. The database is open to the public and we hope that it will become a useful resource for exploring the human proteome. The types of isoforms were classified using our new coding scheme of sequence changes. One finding was that there are very few cases of C-terminal extension probably because of the difficulty in removing the stop codon. We also found that there are some co-occurring combinations of isoforms within the same gene. They are likely to reflect the structurally compatible relationships between isoforms. Finally, we present a number of candidate genes that are predicted to change their subcellular localization between their isoforms. This is the first work that characterizes how alternative isoforms enrich the human proteome comprehensively.

ACKNOWLEDGEMENTS

The authors thank the members of the H-Invitational consortium for providing the manually constructed annotation. The authors thank Goro Terai, Riu Yamashita, Yutaka Suzuki, Jyunichi Takeda, Tadashi Imanishi, Hideki Nagasaki and Kengo Kinoshita for valuable comments and suggestions. We are thankful to Hidemasa Bono and Paul Horton for critical reading of the manuscript. We thank Yasuyuki Fujii for his support with the arrangement of the web materials. This work is partly supported by the 'National Project on Protein and Functional Analysis' and Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology of Japan. Funding to pay the Open Access publication charges for this article was provided by the National Project on Protein Structural and Functional Analyses, MEXT, Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*., **17**, 100–107.
2. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet*., **30**, 13–19.
3. Landry,J.R., Mager,D.L. and Wilhelm,B.T. (2003) Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet*., **19**, 640–648.
4. Gautheret,D., Poirot,O., Lopez,F., Audic,S. and Claverie,J.-M. (1998) Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res*., **8**, 524–530.
5. Beaudoing,E. and Gautheret,D. (2001) Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res*., **11**, 1520–1526.
6. Mironov,A.A., Fickett,J.F. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res*., **9**, 1288–1293.

7. Brett,D., Hanke,J., Lehmann,G., Haase,S., Delbruck,S., Krueger,S., Reich,J. and Bork,P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett*., **474**, 83–86.
8. Kan,Z., Rouchka,E.C., Gish,W.R. and States,D.J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res*., **11**, 889–900.
9. Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res*., **29**, 2850–2859.
10. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **30**, 2141–2144.
11. Grabowski,P.J. (1998) Splicing regulation in neurons: tinkering with cell-specific control. *Cell*, **92**, 709–712.
12. Wang,Z., Lo,H.S., Yang,H., Gere,S., Hu,Y., Buetow,K.H. and Lee,M.P. (2003) Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res*., **63**, 655–657.
13. Stanford,D.R., Martin,N.C. and Hopper,A.K. (2000) ADEPTs: information necessary for subcellular distribution of eukaryotic sorting isozymes resides in domains missing from eubacterial and archaeal counterparts. *Nucleic Acids Res*., **28**, 383–392.
14. Resch,A., Xing,Y., Modrek,B., Gorlick,M., Riley,R. and Lee,C. (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome. *J. Proteome Res*., **3**, 76–83.
15. Norment,A.M., Lonberg,N., Lacy,E. and Littman,D.R. (1989) Alternatively spliced mRNA encodes a secreted form of human CD8 alpha. characterization of the human CD8 alpha gene. *J. Immunol*., **142**, 3312–3319.
16. Ko,J.K., Lee,M.J., Cho,S.H., Cho,J.A., Lee,B.Y., Koh,J.S., Lee,S.S., Shim,Y.H. and Kim,C.W. (2003) Bfl-1S, a novel alternative splice variant of Bfl-1, localizes in the nucleus via its C-terminus and prevents cell death. *Oncogene*, **22**, 2457–2465.
17. Liu,S. and Altman,R.B. (2003) Large scale study of protein domain distribution in the context of alternative splicing. *Nucleic Acids Res*., **31**, 4828–4835.
18. Kriventseva,E.V., Koch,I., Apweiler,R., Vingron,M., Bork,P., Gelfand,M.S. and Sunyaev,S. (2003) Increase of functional diversity by alternative splicing. *Trends Genet*., **19**, 124–128.
19. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al*. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol*., **2**, 0856–0875.
20. Bairoch,A., Boeckmann,B., Ferro,S. and Gasteiger,E. (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform*., **5**, 39–55.
21. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
22. Kersey,P., Hermjakob,H. and Apweiler,R. (2000) VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. *Bioinformatics*, **16**, 1048–1049.
23. Altschul,S., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
24. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*., **29**, 137–140.
25. Emanuelsson,O., Nielsen,H., Brünak,S. and vonHeijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol*., **300**, 1005–1016.
26. Neuberger,G., Maurer-Stroh,S., Eisenhaber,B., Hartig,A. and Eisenhaber,F. (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol*., **328**, 581–592.
27. Neuberger,G., Maurer-Stroh,S., Eisenhaber,B., Hartig,A. and Eisenhaber,F. (2003) Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *J. Mol. Biol*., **328**, 567–579.
28. Krogh,A., Larsson,B., vonHeijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*., **305**, 567–580.
29. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
30. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al*. (2004) An overview of Ensembl. *Genome Res*., **14**, 925–928.
31. Nakai,K. (2002) Signal peptides. In Langel,U. (ed.), *Cell-Penetrating Peptides: Processes*, *Applications*, *1st edn*. CRC Press, pp. 295–324.