4828–4835 *Nucleic Acids Research, 2003, Vol. 31, No. 16*
DOI: 10.1093/nar/gkg668

Large scale study of protein domain distribution in the context of alternative splicing

Shuo Liu and Russ B. Altman*

Department of Genetics, Stanford Medical Informatics, 251 Campus Drive, MSOB X-215, Stanford, CA 94305-5479, USA

Received April 21, 2003; Revised and Accepted June 16, 2003

ABSTRACT

Alternative splicing plays an important role in processes such as development, differentiation and cancer. With the recent increase in the estimates of the number of human genes that undergo alternative splicing from 5 to 35–59%, it is becoming critical to develop a better understanding of its functional consequences and regulatory mechanisms. We conducted a large scale study of the distribution of protein domains in a curated data set of several thousand genes and identified protein domains disproportionately distributed among alternatively spliced genes. We also identified a number of protein domains that tend to be spliced out. Both the proteins having the disproportionately distributed domains as well as those with spliced-out domains are predominantly involved in the processes of cell communication, signaling, development and apoptosis. These proteins function mostly as enzymes, signal transducers and receptors. Somewhat surprisingly, 28% of all occurrences of spliced-out domains are not effected by straightforward exclusion of exons coding for the domains but by inclusion or exclusion of other exons to shift the reading frame while retaining the exons coding for the domains in the final transcripts.

INTRODUCTION

Alternative splicing was first predicted to occur in 1978 (1). With only 32 000 genes predicted from the genome sequence and an estimated 100 000 genes based on EST clustering, alternative splicing may be a major mechanism for producing genomic complexity (2). Prior estimates of the prevalence of alternative splicing were as low as 5% (3) and have been revised to 35–59% of all genes having at least one alternative splice form (4–8). The increased expectation of alternative splicing raises intriguing questions about the identification, functional roles and regulation of alternative splice variants across the whole genome.

Alternative splicing plays a major role in sex determination in *Drosophila*, antibody response in humans and other tissue or developmental stage-specific processes (9). Coordinated changes in alternative splicing patterns of pre-mRNAs are an integral component of gene expression programs such as those involved in nervous system differentiation (10) and apoptotic cell death (11,12). Alternative splicing is also implicated in a large number of human pathologies, such as cancer and Alzheimer's disease (13).

The diverse outcomes of alternative splicing fall into two major categories: protein-level alterations and transcript-level modifications. On the protein level, alternative splicing generates splice variants that give rise to different protein products, for example, a shortened protein product due to a frame shift introduced by an alternate exon or a protein product with a different functional domain due to the inclusion of a specific exon from a mutually exclusive group of exons. On the transcript level, alternative splicing produces splice variants that have different translation or stability profiles, for example, a transcript with a longer life span would prolong the availability of the corresponding protein product.

There are many examples of protein-level alterations introduced by alternative splicing (14–20). In some cases, the proteins produced by the splice variants play opposite roles in the cellular processes. These studies tend to focus on specific genes and not on classes of proteins that are, as a group, disproportionately affected by alternative splicing. We would like to find protein domains that are disproportionately distributed among alternatively spliced or constitutively spliced genes. We would also like to find domains that are frequently spliced out among members of a protein family or across protein families.

Most large scale studies of alternative splicing focus on the DNA and RNA level. Loraine *et al.* identified genes whose alternative transcripts produce different protein domain structures and developed a graphical tool to study protein domain differences for the splice variants of individual genes (21). Their work did not address the distribution of protein domains more generally. In this paper, we used manually reviewed LocusLink and RefSeq records to systematically study protein domains in the context of alternative splicing.

MATERIALS AND METHODS

Data source

The July 26, 2002 release of LocusLink was downloaded from NCBI web site (ftp://ftp.ncbi.nih.gov/refseq/LocusLink/). The flat file was parsed with a parser program and selected fields

*To whom correspondence should be addressed. Tel: +1 650 725 3394; Fax: +1 650 725 7944; Email: russ.altman@stanford.edu were loaded into a relational database. Only loci with a 'reviewed' status were used for this study. Reviewed status is given to a locus after a careful review by the NCBI staff. Reviewed loci that have more than one RefSeq transcript associated with them constituted the set of reviewed alternatively spliced loci.

Identification of disproportionately distributed protein domains

The NCBI Conserved Domain Database (CDD, http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) protein domain annotations were used to identify protein domains. These annotations are generated by the NCBI staff using the NCBI CDD search tool (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). For each of the CDD protein domains annotated on reviewed loci, the numbers of constitutively spliced and alternatively spliced reviewed loci that were annotated with it were determined. A 2 × 2 contingency table and Fisher's exact test were used to determine if a protein domain is disproportionately distributed among alternatively spliced and constitutively spliced loci. A sample contingency table for Cadherin domain is shown in Table 1. To avoid false positives generated by the testing of multiple hypothesis, the false discovery rate correction method (22) was applied.

Characterization of gene ontology (GO) annotations for a set of genes

GO terms and annotations were obtained in a mySQL database developed by BDGP from http://www.godatabase.org/dev/database/. We downloaded the December 2002 release and imported it into the Oracle database previously loaded with LocusLink data. Given a set of genes, all the GO terms for two sub-ontologies of GO (molecular function and biological process) were extracted from the GO database. The count of the number of genes for each GO term was recorded. We defined the sub-tree covering a given set of GO terms as the one rooted by the deepest common ancestor of all the given GO terms. For gene coverage, the number of genes represented by the sub-tree rooted in a GO term was divided into the total number of annotated genes to arrive at the percentage coverage. For GO term coverage, the number of GO terms represented by the sub-tree rooted in a GO term was divided into the total number of unique GO terms to arrive at the percentage coverage.

The *P*-value assigned to each GO term was calculated as described on the ProToGo website (http://www.protonet.cs.huji.ac.il/ProToGO/Introduction.html). First, the number of reviewed loci assigned to each GO term was calculated. This count includes all the loci assigned to a GO term and those assigned to its descendents. Duplicated loci were counted only once. This served as the background distribution. Then, for a given set of loci, a similar calculation was performed to derive the number of loci from this set assigned to each GO term. The *P*-value calculated for each GO term is expressing the probability to receive such a count from the given set, or higher, assuming the null hypothesis under hypergeometric distribution.

Identification of spliced-out protein domains

To identify spliced-out protein domains, we used CDD protein domain annotations. In particular, annotations corresponding to each splice variant were extracted from the database. Domains that were not present on all splice variants were considered spliced-out protein domains.

Table 1. Sample contingency table for use in Fisher's test

Cadherin domain	Constitutively spliced	Alternatively spliced	Total
With domain	50	50	100
Without domain	3566	882	4448
Total	3616	932	4548

Classification of transcript level modifications to spliced-out domains

Two types of transcript level modifications can be employed to produce a spliced variant without a protein domain. One is to exclude one or more exons coding for the domain from the processed transcript. The other is to retain all the exons coding for the domain in the transcript but shift the reading frame resulting in the loss of the domain. To estimate the prevalence of each type of modification, we classified each occurrence of alternatively spliced domains in our data set. For each locus that has a specific spliced-out domain, all the transcripts for the locus were first aligned with BLAT against the June 2002 release of the human genome draft sequence at UCSC (http://genome.ucsc.edu). Only transcripts that align to the draft genome without internal gaps and with at most 50 nt missing at either end were kept. The location of the domain was determined using the NCBI CDD search tool available at http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml for the transcripts having the domain. The location information was then used to extract the exons coding for the domain. Exon structures of the transcripts without the domain were compared with the exons coding the domain to classify the type of modifications.

RESULTS

Disproportionately distributed protein domains

In our study, all disproportionately distributed domains are domains that are disproportionately distributed in the direction of alternatively spliced genes. They are not necessarily the domains that are spliced in or out. Thus, these domains can be considered 'innocent bystanders' of the splicing process which affects other areas of the genes.

The July 26, 2002 release of LocusLink contains 47 310 human loci. 4548 of these have a 'reviewed' status. 932 of the reviewed loci have more than one RefSeq transcript associated. Thus, 20% of all the reviewed loci are annotated with alternative splicing. A total of 2501 unique CDD protein domains were annotated on all the human loci in LocusLink. From these, a total of 1675 were annotated on all the reviewed loci, or 67%.

Disproportionately distributed domains are summarized in Table 2. Twenty four domains were found to be disproportionately distributed and all of them showed higher frequency of occurrences in alternatively spliced loci. The proteins harboring these domains are engaged in diverse cellular processes such as adhesion and morphogenesis (23) (cadherin domains), growth and differentiation (24) (phosphatase

Table 2. Protein domains disproportionately distributed among alternatively spliced genes

Domain name	Domain ID	InterPro ID	No. of constitutively spliced loci with this domain	No. of alternatively spliced loci with this domain	No. of constitutively spliced loci without this domain	No. of alternatively spliced loci without this domain	*P*-value
Cadherin domain	pfam00028	IPR002126	50	50	3566	882	0
Cadherin repeats	smart00112	IPR002126	47	43	3569	889	0
Protein tyrosine phosphatase, catalytic domain	smart00194	IPR000242	18	21	3598	911	0.000004
Protein tyrosine phosphatase, catalytic domain motif	smart00404	IPR003595	20	21	3596	911	0.000011
Protein-tyrosine phosphatase	pfam00102	IPR000242	22	22	3594	910	0.000011
B-box zinc finger	pfam00643	IPR000315	12	16	3604	916	0.000021
Caspase, interleukin-1 beta converting enzyme (ICE) homologs	smart00115	IPR002398	2	9	3614	923	0.000023
ICE-like protease (caspase) p10 domain	pfam00655	IPR002138	2	9	3614	923	0.000023
ICE-like protease (caspase) p20 domain	pfam00656	IPR001309	2	9	3614	923	0.000023
Domain in SPla and the RYanodine receptor	smart00449	IPR003877	6	12	3610	920	0.000028
Ankyrin repeats	smart00248	IPR002110	5	11	3611	921	0.00004
B-box-type zinc finger	smart00336	IPR000315	12	15	3604	917	0.000061
Dual specificity phosphatase, catalytic domain	smart00195	IPR000340	12	15	3604	917	0.000061
Fibronectin type III domain	smart00060	IPR003961	16	17	3600	915	0.000071
Ank repeat	pfam00023	IPR002110	14	15	3602	917	0.000179
Dual specificity phosphatase	pfam00782	IPR000340	18	17	3598	915	0.000181
Ring finger	smart00184	IPR001841	16	16	3600	916	0.000183
SPRY domain	pfam00622	IPR003877	11	13	3605	919	0.000269
Somatotropin hormone family	pfam00103	IPR001400	0	5	3616	927	0.000358
Zinc-binding domain seen in both chromatinic and cytoskeletal proteins	LOAD_zz		0	5	3616	927	0.000358
Serine/threonine protein kinases, catalytic domain	smart00220	IPR002290	46	28	3570	904	0.000406
Tyrosine kinase, catalytic domain	smart00219	IPR001245	46	28	3570	904	0.000406
Double-stranded RNA binding motif	pfam00035	IPR001159	1	6	3615	926	0.000422
Zinc finger, C3HC4 type (ring finger)	pfam00097	IPR001841	16	15	3600	917	0.000459

domains), transcription regulation (25) (B-box zinc finger domain), apoptosis (26) (caspase domains) and intracellular Ca^{2+} signaling (SPRY domain).

Process and function GO annotations for alternatively spliced genes with disproportionately distributed protein domains

To more accurately assess the roles played by alternatively spliced genes with disproportionately distributed protein domains, we extracted GO annotations on these genes and sought to identify common themes from both the perspective of GO terms and that of genes. The 24 domains identified are found on 169 alternatively spliced genes. For the biological process sub-ontology of GO, 108 genes were annotated with 75 unique GO terms (229 total).

Figure 1 shows the top 10 significant terms in the biological process hierarchy with respect to gene coverage. Gene coverage allows the identification of processes that a majority of these genes participate in. The top three significant terms—cell communication, signal transduction and protein metabolism—account for 53, 32 and 25% of this set of genes, respectively. The 10 most significant terms ranked by *P*-value are shown in Table 3. All these processes require exquisite control of responses to extracellular or intracellular signals and it is our belief that alternative splicing is one of the regulatory mechanisms employed.

Top terms with respect to term coverage were also identified. Term coverage allows the identification of processes that these genes frequently participate in without being biased by the number of genes annotated by each term. The top two terms—cell growth and/or maintenance and cell communication—account for 52 and 28% of all biological processes participated in by this set of genes, respectively.

Similar analyses of these genes were carried out for the molecular function sub-ontology. Out of 169 alternatively spliced genes, 121 were annotated with 69 unique GO terms (176 total). For term coverage, three first level terms—enzyme, binding and signal transducer—account for 50, 27 and 18% of all molecular functions carried out by this set of genes, respectively. For gene coverage, the top three significant terms—enzyme, hydrolase and protein kinase—account for 51, 30 and 20% of all the genes, respectively. These functions are critical for the processes identified above. The 10 most significant terms ranked by *P*-value are shown in Table 4. They include seven terms for phosphatase and one term each for serine/threonine kinase, cell adhesion molecule and caspase. Proteins with these functions are involved in signal transduction, development and apoptosis.

Spliced-out protein domains

Spliced-out domains, in contrast to disproportionately distributed domains, are those which not only occur in alternatively spliced genes, but are also the target of splicing in or out.

Out of 932 reviewed loci that have more than one associated RefSeq transcript, we identified 202 domains as spliced out. This represents 26% of a total of 773 domains. 139 loci (15%)

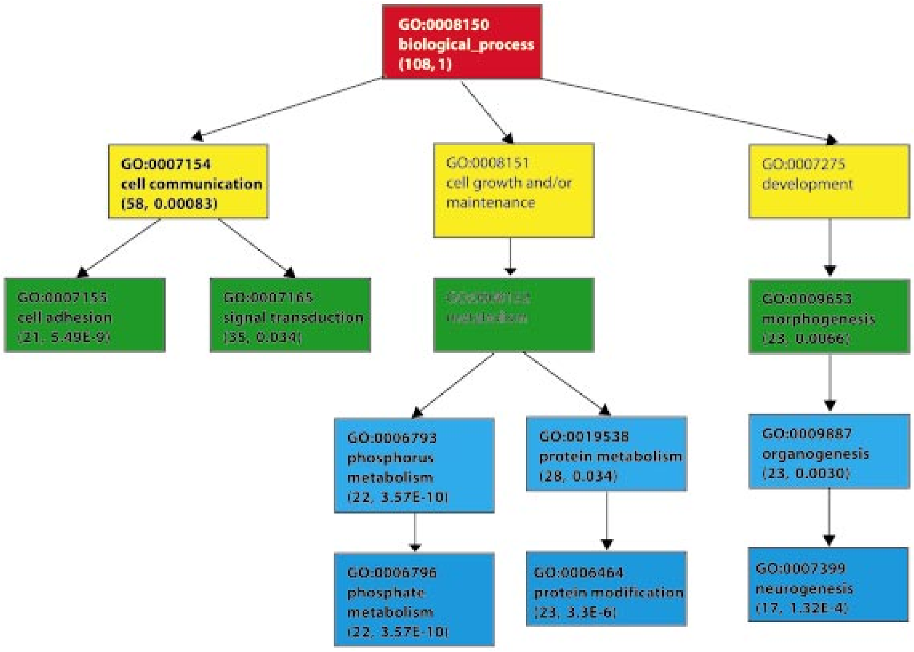

Figure 1. Top 10 significant terms in the biological process hierarchy with respect to gene coverage for alternatively spliced genes having the disproportionately distributed domains. The terms are color coded according to their depth in the hierarchy and the numbers in parentheses after the name of each term are the number of genes covered in the sub-tree with the term as the root and the *P*-value. The terms in bold font are the top terms. The terms not in bold and with no numbers of genes and *P*-values are shown because they are on the path leading to the top significant terms.

Table 3. Ten most significant biological process GO terms for alternatively spliced genes with disproportionately distributed domains

GO term accession	GO term name	No. of genes annotated	Depth of term	Percent of gene occurrences	*P*-value
GO:0006470	Protein dephosphorylation	14	6	12	9.96E–11
GO:0016311	Dephosphorylation	14	5	12	9.96E–11
GO:0006796	Phosphate metabolism	22	4	20	3.57E–10
GO:0006793	Phosphorus metabolism	22	3	20	3.57E–10
GO:0007155	Cell adhesion	21	2	19	5.49E–09
GO:0006464	Protein modification	23	4	21	3.33E–06
GO:0007185	Transmembrane receptor protein tyrosine phosphatase signaling pathway	3	5	2	7.28E–05
GO:0007399	Neurogenesis	17	4	15	1.32E–04
GO:0007154	Cell communication	58	1	53	8.30E–04
GO:0007089	Mitotic start control point	3	5	2	0.001327

were found to have one or more of these domains. Two interesting subsets are shown in Tables 5 and 6. Table 5 shows 14 domains that are present on four or more alternatively spliced loci. Table 6 shows 13 domains that are present on at least two loci and are always alternatively spliced. Fourteen of the 24 disproportionately distributed domains were also identified as spliced-out domains.

GO term analyses were carried out for the 40 alternatively spliced genes with spliced-out domains shown in Table 5. These results are fundamentally the same as previous analyses on genes with disproportionately distributed domains. For the biological process sub-ontology, 28 genes were annotated with 40 unique GO terms (58 total). With respect to gene coverage, the top three significant terms—protein metabolism, development and organogenesis—account for 35, 35 and 28% of all the genes, respectively. Protein metabolism and organogenesis are also the top significant terms for genes with disproportionately distributed domains as shown in Figure 1.

For the molecular function sub-ontology, 24 genes were annotated with 31 unique GO terms (38 total). Figure 2 shows the top 10 significant terms with respect to gene coverage for this set of alternatively spliced genes. The top three significant terms—enzyme, hydrolase and cysteine-type peptidase—account for 70, 50 and 25% of all the genes, respectively. Table 7 shows the 10 most significant molecular function GO

Table 4. Ten most significant molecular function GO terms for alternatively spliced genes with disproportionately distributed domains

GO term accession	GO term name	No. of genes annotated	Depth of term	Percent of gene occurrences	*P*-value
GO:0004721	Protein phosphatase	22	5	18	2.07E–16
GO:0004725	Protein tyrosine phosphatase	19	6	15	1.40E–14
GO:0016302	Phosphatase	22	2	18	1.92E–14
GO:0016791	Phosphoric monoester hydrolase	22	4	18	2.75E–14
GO:0005001	Transmembrane receptor protein tyrosine phosphatase	11	7	9	3.01E–11
GO:0019198	Transmembrane receptor protein phosphatase	11	6	9	3.01E–11
GO:0004674	Protein serine/threonine kinase	18	6	14	9.69E–10
GO:0005194	Cell adhesion molecule	22	1	18	1.27E–09
GO:0016788	Hydrolase, acting on ester bonds	22	3	18	1.84E–09
GO:0004199	Caspase	7	6	5	2.83E–09

Table 5. Spliced-out domains that are present on at least four loci

Domain name	Domain ID	InterPro ID	No. of alternatively spliced loci where this domain is spliced out	No. of alternatively spliced loci where this domain is present	Percent of loci where this domain is spliced out
Domain in SPla and the RYanodine receptor	smart00449	IPR003877	7	12	58
SPRY domain	pfam00622	IPR003877	7	13	54
Intermediate filament proteins	pfam00038	IPR001664	6	11	55
ICE-like protease (caspase) p20 domain	pfam00656	IPR001309	5	9	56
Eukaryotic protein kinase domain	pfam00069	IPR000719	5	42	12
EGF-like domain. pfam00053 is very similar	pfam00008	IPR006209	4	6	67
Ezrin/radixin/moesin family	pfam00769	IPR000798	4	9	44
ICE-like protease (caspase) p10 domain	pfam00655	IPR002138	4	9	44
Immunoglobulin domain	pfam00047	IPR003006	4	14	29
Serine/threonine protein kinases, catalytic domain	S_TKc	IPR002290	4	15	27
Tyrosine kinase, catalytic domain	TyrKc		4	15	27
Dual specificity phosphatase	pfam00782	IPR000340	4	17	24
Fibronectin type III domain	pfam00041	IPR003961	4	18	22
Tyrosine kinase, catalytic domain	smart00219	IPR001245	4	28	14

Table 6. Domains that are spliced out for all alternatively spliced genes in which they are found

Domain name	Domain ID	InterPro ID	No. of alternatively spliced loci where this domain is spliced out	No. of alternatively spliced loci where this domain is present	Percent of loci where this domain is spliced out
Calcium-binding EGF-like domain	EGF_CA	IPR001881	3	3	100
Domain abundant in complement control proteins	CCP	IPR000436	2	2	100
Epidermal growth factor-like domain	EGF	IPR006209	2	2	100
Conserved domain seen in Groovin and Gas2 proteins	LOAD_gas2groo		2	2	100
Growth-arrest-specific protein 2 domain	smart00243	IPR003108	2	2	100
Growth-arrest-specific protein 2 domain	pfam02187	IPR003108	2	2	100
Elongation factor Tu GTP binding domain. This domain contains a P-loop motif	pfam00009	IPR000795	2	2	100
von Willebrand factor (vWF) type C domain	smart00214	IPR001007	2	2	100
von Willebrand factor type C domain	pfam00093	IPR001007	2	2	100
Immunoglobulin C-type	smart00407	IPR003597	2	2	100
Neuregulin family	pfam02158	IPR002154	2	2	100
Zinc-binding domain present in Lin-11, Isl-1, Mec-3	smart00132	IPR001781	2	2	100
Shikimate kinase	pfam01202	IPR000623	2	2	100

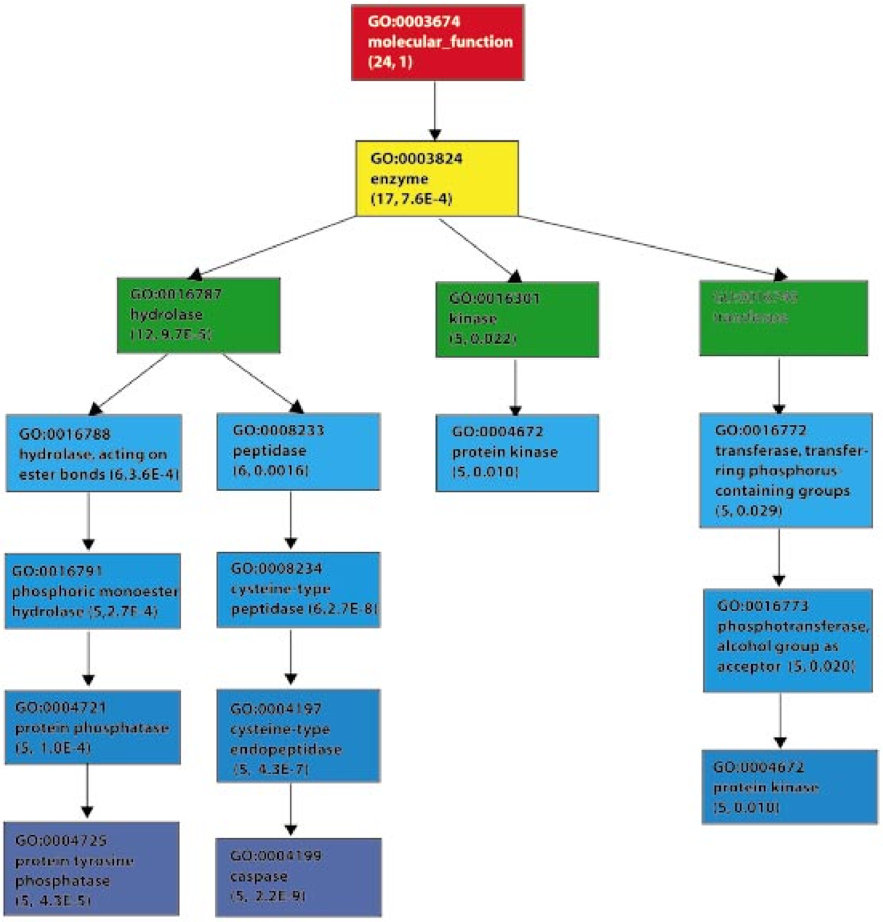

Figure 2. Top 10 significant terms in the molecular function hierarchy with respect to gene coverage for alternatively spliced genes having the 14 spliced-out domains that are present on four or more alternatively spliced loci. The terms are color coded according to their depth in the hierarchy and the number in parentheses after the name of each term is the number of genes covered in the sub-tree with the term as the root and the *P*-value. The terms in bold font are the top terms. The terms not in bold and with no numbers of genes and *P*-values are shown because they are on the path leading to the top significant terms. More than 10 terms are shown because of ties.

terms ranked by *P*-value. They include terms that lead from hydrolase to signaling caspase and tyrosine phosphatase as shown in Figure 2. Proteins with these functions are involved in apoptosis and signal transduction.

Transcript level modifications to splice out protein domains

The straightforward way to splice out a protein domain would be to exclude the exons coding the domain. However, we found several examples of another transcript level modification to achieve this, i.e. inclusion or exclusion of other exons to shift the reading frame (15,17,18).

We performed a systematic analysis of the types of transcript level modification used to splice out domains. Somewhat surprisingly, out of a total of 314 instances of a domain being spliced out, only 225 instances (72%) were effected with the straightforward exclusion of one or more exons coding for the domain, while 89 instances (28%) were effected by retaining all the exons coding for the domain but shifting the reading frame to remove the original amino acid sequence of the domain from the final protein product.

DISCUSSION

One of the direct effects of alternative splicing is the production of different protein products. The current study identifies two types of protein domain that are of special interest. The first type includes protein domains disproportionately distributed among the alternatively or constitutively spliced genes (the ‘innocent bystanders’). The second includes spliced-out protein domains (the ‘victims’). The disproportionately distributed and spliced-out domains are associated with genes playing important roles in processes such as signal processing, development, differentiation and apoptosis. Alternative splicing may be an important mechanism to regulate the functions of these genes. It has long been hypothesized that alternative splicing plays a regulatory role in these processes. Our results provide objective evidence to support this hypothesis.

Although we understand the significance of some domains on our list of disproportionately distributed and spliced-out domains, the functional significance of all the domains is not well understood. For example, caspase catalytic domains, elongation factor Tu GTP binding domain and von Willebrand

Table 7. Ten most significant molecular function GO terms for alternatively spliced genes with spliced-out domains shown in Table 5

GO term accession	GO term name	No. of genes annotated	Depth of term	Percent of gene occurrences	*P*-value
GO:0004199	Caspase	5	6	20	2.23E–09
GO:0004200	Signaling (initiator) caspase	4	7	16	2.65E–08
GO:0008234	Cysteine-type peptidase	6	4	25	2.68E–08
GO:0004197	Cysteine-type endopeptidase	5	5	20	4.34E–07
GO:0004725	Protein tyrosine phosphatase	5	6	20	4.34E–05
GO:0016787	Hydrolase	12	2	50	9.66E–05
GO:0004721	Protein phosphatase	5	5	20	1.04E–04
GO:0016302	Phosphatase	5	2	20	2.51E–04
GO:0016791	Phosphoric monoester hydrolase	5	4	20	2.69E–04
GO:0016788	Hydrolase, acting on ester bonds	6	3	25	3.56E–04

factor (vWF) type C domain play important roles in the modification of protein functions, especially across members of a protein family. Isoforms with or without these domains may be involved in the development of diseases such as leukemia, neuroblastoma and gastric carcinoma (14–18,27,28). The cadherin domain, a disproportionately distributed domain, is present on all members of the large protocadherin family, whose splice variants play important roles during nervous system development, and the exact functions of the variant cytoplasmic domains are active areas of research (29). Our results suggest that genes containing these domains are more likely than others to use splicing as a regulatory mechanism.

Somewhat surprisingly, a high percentage (28%) of the spliced-out domains were not effected by straightforward exclusion of exons coding the affected domain. Presumably, some exons coding the domains are under very high selection pressure to be conserved, so other exons either upstream or in the middle of these exons are used to modulate the exclusion of the domains in protein isoforms. As such, it is very important to supplement the traditional exon structure study of alternative splicing with studies from the perspective of protein domains. Otherwise, the changes to protein sequences as a result of reading frame shifts while retaining most of the exon structures will not be readily detected. This phenomenon demonstrates the importance of proteomics since genome and transcriptome studies may not readily detect such changes. It also presents an interesting challenge to gene prediction and eventually alternative splicing prediction algorithms as most algorithms use reading frame consistency as a guide to predict the exon structures.

The current study adopted a protein domain centric approach as opposed to a gene centric approach. We looked at the functional diversity of proteins grouped together with common patterns of alternative splicing via GO molecular function annotations. We found that some groups had similar and consistent annotations (e.g. proteins having a cadherin domain) and others that had diverse annotations (e.g. proteins having a B-box zinc finger domain or a Fibronectin type III domain). The significance of these findings is difficult to assess because function assignments are made to multi-domain genes.

A relatively small percentage of domains (1.4%) were found to be disproportionately distributed among the alternatively spliced and constitutively spliced loci. The number of occurrences of each domain among reviewed loci is very low. As more loci are reviewed, the number of occurrences of each domain will increase correspondingly and more domains could be found to be disproportionately distributed.

To ensure the quality of our analyses with limited data, our study focused on the set of reviewed loci in LocusLink. Reviewed status is given to a locus after a careful review by the NCBI staff. Such a focus probably improves the quality of the annotation but introduces potential biases in our findings. For example, the reviewers could prioritize their reviews based on the availability of publications, and researchers could be more interested in studying genes involved in signal transduction, development and apoptosis. Since the percentage of loci that have multiple RefSeq annotation among the provisional and predicted loci is low, the choice of reviewed loci should not result in any significant losses of domains from either the disproportionately distributed list or the spliced-out list.

The percentage (20%) of alternatively spliced loci among reviewed loci in LocusLink is much lower than those estimates (40–60%) produced by several studies of alternative splicing based on mRNA and/or EST data. Since ESTs are only partial fragments of the full-length transcripts, it is difficult to reliably infer the full-length transcripts from a collection of ESTs. As our results show, full-length transcripts are required before alternate protein sequences can reliably be determined. Methods that find and annotate alternative splicing and generate full-length transcripts are needed before all the disproportionately distributed and alternatively spliced domains can be identified. There are some promising recent developments in this area (5,30–37).

Given the important roles alternative splicing is playing, large scale studies of its functional consequences and regulatory mechanisms can provide insights into many important normal and abnormal processes such as differentiation and cancer development and provide us with a better understanding of why nature chooses to play the combinatorial game.

ACKNOWLEDGEMENTS

We thank Drs Sam Karlin, Michael Walker, Robert Tibshirani and Luciano Brocchieri for comments and suggestions. R.A. and S.L. are supported by NIH GM61374. S.L. was a Stanford Graduate Fellowship James Clark Fellow and is also supported by NIH LM07033. We thank SUN, Inc. for a hardware grant.

REFERENCES

1. Gilbert,W. (1978) Why genes in pieces? *Nature*, **271**, 501.
2. Brett,D., Pospisil,H., Valcarcel,J., Reich,J. and Bork,P. (2002) Alternative splicing and genome complexity. *Nature Genet.*, **30**, 29–30.
3. Sharp,P.A. (1994) Split genes and RNA splicing. *Cell*, **77**, 805–815.
4. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet.*, **30**, 13–19.
5. 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.
6. 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.
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. Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.
9. Lopez,A.J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.*, **32**, 279–305.
10. Grabowski,P.J. (1998) Splicing regulation in neurons: tinkering with cell-specific control. *Cell*, **92**, 709–712.
11. Jiang,Z.H. and Wu,J.Y. (1999) Alternative splicing and programmed cell death. *Proc. Soc. Exp. Biol. Med.*, **220**, 64–72.
12. Wu,J.Y., Tang,H. and Havlioglu,N. (2003) Alternative pre-mRNA splicing and regulation of programmed cell death. *Prog. Mol. Subcell. Biol.*, **31**, 153–185.
13. Ho,L., Guo,Y., Spielman,L., Petrescu,O., Haroutunian,V., Purohit,D., Czernik,A., Yemul,S., Aisen,P.S., Mohs,R. *et al.* (2001) Altered expression of a-type but not b-type synapsin isoform in the brain of patients at high risk for Alzheimer's disease assessed by DNA microarray technique. *Neurosci. Lett.*, **298**, 191–194.
14. Droin,N., Bichat,F., Rebe,C., Wotawa,A., Sordet,O., Hammann,A., Bertrand,R. and Solary,E. (2001) Involvement of caspase-2 long isoform in Fas-mediated cell death of human leukemic cells. *Blood*, **97**, 1835–1844.
15. Droin,N., Beauchemin,M., Solary,E. and Bertrand,R. (2000) Identification of a caspase-2 isoform that behaves as an endogenous inhibitor of the caspase cascade. *Cancer Res.*, **60**, 7039–7047.
16. Eckhart,L., Henry,M., Santos-Beneit,A.M., Schmitz,I., Krueger,A., Fischer,H., Bach,J., Ban,J., Kirchhoff,S., Krammer,P.H. *et al.* (2001) Alternative splicing of caspase-8 mRNA during differentiation of human leukocytes. *Biochem. Biophys. Res. Commun.*, **289**, 777–781.
17. Horiuchi,T., Himeji,D., Tsukamoto,H., Harashima,S., Hashimura,C. and Hayashi,K. (2000) Dominant expression of a novel splice variant of caspase-8 in human peripheral blood lymphocytes. *Biochem. Biophys. Res. Commun.*, **272**, 877–881.
18. Waltereit,R. and Weller,M. (2002) The role of caspases 9 and 9-short (9S) in death ligand- and drug-induced apoptosis in human astrocytoma cells. *Brain Res. Mol. Brain Res.*, **106**, 42.
19. Ando,S., Sarlis,N.J., Krishnan,J., Feng,X., Refetoff,S., Zhang,M.Q., Oldfield,E.H. and Yen,P.M. (2001) Aberrant alternative splicing of thyroid hormone receptor in a TSH-secreting pituitary tumor is a mechanism for hormone resistance. *Mol. Endocrinol.*, **15**, 1529–1538.
20. Jiang,Z.H., Zhang,W.J., Rao,Y. and Wu,J.Y. (1998) Regulation of Ich-1 pre-mRNA alternative splicing and apoptosis by mammalian splicing factors. *Proc. Natl Acad. Sci. USA*, **95**, 9155–9160.
21. Loraine,A.E., Helt,G.A., Cline,M.S. and Siani-Rose,M.A. (2002) Protein based analysis of alternative splicing in the human genome. *IEEE Computer Society Bioinformatics Conference*. IEEE Computer Society, pp. 118–124.
22. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.
23. Takeichi,M. (1990) Cadherins: a molecular family important in selective cell-cell adhesion. *Annu. Rev. Biochem.*, **59**, 237–252.
24. Fischer,E.H., Charbonneau,H. and Tonks,N.K. (1991) Protein tyrosine phosphatases: a diverse family of intracellular and transmembrane enzymes. *Science*, **253**, 401–406.
25. Borden,K.L. (1998) RING fingers and B-boxes: zinc-binding protein–protein interaction domains. *Biochem. Cell Biol.*, **76**, 351–358.
26. Nicholson,D.W. and Thornberry,N.A. (1997) Caspases: killer proteases. *Trends Biochem. Sci.*, **22**, 299–306.
27. Tanaka,S., Sugimachi,K., Saeki,H., Kinoshita,J., Ohga,T., Shimada,M. and Maehara,Y. (2001) A novel variant of WISP1 lacking a Von Willebrand type C module overexpressed in scirrhous gastric carcinoma. *Oncogene*, **20**, 5525–5532.
28. Schurmann,A., Brauers,A., Massmann,S., Becker,W. and Joost,H.G. (1995) Cloning of a novel family of mammalian GTP-binding proteins (RagA, RagBs, RagB1) with remote similarity to the Ras-related GTPases. *J. Biol. Chem.*, **270**, 28982–28988.
29. Frank,M. and Kemler,R. (2002) Protocadherins. *Curr. Opin. Cell Biol.*, **14**, 557–562.
30. Fairbrother,W.G., Yeh,R.F., Sharp,P.A. and Burge,C.B. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.
31. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res.*, **31**, 101–105.
32. Xu,Q., Modrek,B. and Lee,C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.*, **30**, 3754–3766.
33. Spingola,M., Grate,L., Haussler,D. and Ares,M.,Jr. (1999) Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA*, **5**, 221–234.
34. Stamm,S., Zhu,J., Nakai,K., Stoilov,P., Stoss,O. and Zhang,M.Q. (2000) An alternative-exon database and its statistical analysis. *DNA Cell Biol.*, **19**, 739–756.
35. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genet.*, **29**, 412–417.
36. Clark,T.A., Sugnet,C.W. and Ares,M., Jr. (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, **296**, 907–910.
37. Lim,L.P. and Burge,C.B. (2001) A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA*, **98**, 11193–11198.