Journal of Bioinformatics and Computational Biology
Vol. 1, No. 2 (2003) 289–306
© Imperial College Press

Imperial College Press
www.icpress.co.uk

EXPLORING ALTERNATIVE TRANSCRIPT STRUCTURE IN THE HUMAN GENOME USING BLOCKS AND INTERPRO

ANN E. LORAINE*, GREGG A. HELT†, MELISSA S. CLINE‡
and MICHAEL A. SIANI-ROSE§

*Bioinformatics Department, Affymetrix, 6550 Vallejo St,
Emeryville, CA 94530, USA*
**ann_loraine@affymetrix.com*
†*gregg_helt@affymetrix.com*
‡*melissa_cline@affymetrix.com*
§*michael_siani-rose@affymetrix.com*

Received 15 September 2002
Revised 7 December 2002
Accepted 15 January 2003

Understanding how alternative splicing affects gene function is an important challenge facing modern-day molecular biology. Using homology-based, protein sequence analysis methods, it should be possible to investigate how transcript diversity impacts protein function. To test this, high-quality exon-intron structures were deduced for over 8000 human genes, including over 1300 (17 percent) that produce multiple transcript variants. A data mining technique (DiffMotif) was developed to identify genes in which transcript variation coincides with changes in conserved motifs between variants. Applying this method, we found that 30 percent of the multi-variant genes in our test set exhibited a differential profile of conserved InterPro and/or BLOCKS motifs across different mRNA variants. To investigate these, a visualization tool (ProtAnnot) that displays amino acid motifs in the context of genomic sequence was developed. Using this tool, genes revealed by the DiffMotif method were analyzed, and when possible, hypotheses regarding the potential role of alternative transcript structure in modulating gene function were developed. Examples of these, including: MEOX1, a homeobox-containing protein; AIRE, involved in auto-immune disease; PLAT, tissue type plasminogen activator; and CD79b, a component of the B-cell receptor complex, are presented. These results demonstrate that amino acid motif databases like BLOCKS and InterPro are useful tools for investigating how alternative transcript structure affects gene function.

Keywords: Alternative splicing; CD79b; AIRE; PLAT; MEOX1; BLOCKS; InterPro; visualization.

1. Introduction

Understanding how alternative splicing affects gene function is an important challenge facing modern-day molecular biology. Current estimates project that around a third of multi-exon, human genes give rise to more than one transcript variant, but the functional significance of this variation is unknown in most cases.[1,2]

Alternative splicing has been studied intensively at the level of individual genes, and examples of differential expression of transcript variants are common.[3,4] In some cases, expression of specific variants appears to be restricted to distinct cell types,[5] while in others, different variants are co-expressed in varying ratios,[6,7] and these ratios may change in response to extracellular signals or changes in the physiological state of the cell.

Other mechanisms besides alternative splicing contribute to transcript diversity, however. For example, alternative promoter choice, resulting in differential transcription start site selection, can produce transcripts with variable 5′ ends.[8,9] Likewise, alternative polyadenylation can produce RNA species with variable 3′ regions.[10–12] To our knowledge, no analysis of the frequency of alternative promoter choice among human genes has been published. Alternative polyadenylation site choice is better understood; a recent manual analysis of 52 human genes located on chromosome 2lq22.3 reports that at least half of these genes exhibit clear evidence of multiple polyadenylation sites.[13]

These mechanisms for generating transcript diversity can affect gene function either at the level of RNA, such as by differential inclusion of structural or sequence-based motifs controlling translation, message localization, or stability, or at the level of the translated protein product when coding regions are affected. In the latter case, alternative transcript structure can affect protein function in at least three ways. First, alternative transcript structure could simply remodel domains present in all forms, such as by mutually exclusive use of cassette exons, each encoding variations on the same general motif or functional domain. Second, functionally important coding sequence can be added or deleted, resulting in proteins with different, even antagonistic functions. For example, the human BclX gene produces at least two distinct variants, a long form which inhibits apoptosis and a shorter form which promotes it.[14] Third, alternative splicing may change the composition of protein motifs that are detected as repeated and/or discontinuous spans across the protein sequence. For example, the number of repeated motif elements, such as WD40 repeats, may differ between variants. Here, we explore the effects of these latter two types of changes using libraries of conserved amino acid motifs.

Numerous methods for classifying proteins according to their homologies to known amino acid patterns or motif signatures have been developed. For example, the BLOCKS library of protein family profiles is based on conserved, ungapped segments ("BLOCKS") that typically occur in clusters within related proteins.[15] The InterPro database organizes protein sequence profiles from several different motif databases into a single resource and so provides a comprehensive description of the known protein universe.[16] Profiles which detect similar patterns have been grouped into single InterPro entries, and many of these InterPro "meta-motifs" have been assigned biological function using the Gene Ontology Consortium's controlled vocabulary.[17] Using the program InterproScan, one can easily search for homologies to protein motifs described by the PRINTS, ProDom, PROSITE, Pfam, and other InterPro member databases.

These and other protein classification systems have been widely used to predict function and family affiliation for newly discovered proteins deduced from the conceptual translation of genomic DNA and cDNA sequences. Given the sensitivity and wide scope of these methods, it is possible that they could also be used to study how alternative splicing impacts protein function. By comparing the pattern of functional motifs detected in different proteins produced by the same gene, it should be possible to construct hypotheses describing how alternative transcript structure may affect gene function. To determine whether this approach is feasible, we developed a test set of human genes with solved genomic structure and used BLOCKS and InterPro member libraries to examine their diverse protein products.

2. Methods

2.1. *Gene structures*

Human cDNA sequences were downloaded from GenBank. From these, a subset was selected that were annotated as encoding a full-length protein product. The selected sequences were aligned to working draft human genomic sequence (April, 2001 release) using pslayout, a fast cDNA-to-genomic sequence alignment program.[18] Gene structures were defined from cDNA-to-genomic sequence alignments as follows: Regions of continuous alignment between cDNA and genomic sequence were used to define exons. Gaps exceeding 20 bases in the cDNA sequence relative to genomic were used to define introns. Exons separated by fewer than 20 bases were merged to form a single exon.

2.2. *Selecting high-quality alignments*

Coding sequence annotations from the original Genbank records were mapped onto genomic sequence and the translation frame for coding-region exons was deduced. A conceptual translation of genomic sequence was then attempted for each gene structure inferred from cDNA-to-genomic alignments. Poor quality alignments were identified and eliminated by aligning the conceptual translations of genomic sequence to the corresponding protein sequence from GenPept using the pairwise alignment program bl2seq, a BLAST algorithm implementation available from N.C.B.I. Sequences that failed to align with 95 percent identity or better across the full length of both alignment partners were discarded. In this way, sequences with unreliable exon-intron structures within the coding region were identified and excluded from subsequent analyses.

2.3. *Gene classification*

Using mappings of coding regions onto genomic sequence as a guide, transcripts sharing at least 50 bases (15 amino acids) of same-frame, continuous coding region sequence were assigned to the same gene. Using this “shared coding region” rule, transcripts mapping entirely to introns within other transcripts were automatically assigned to different genes. Similarly, transcripts which overlapped only within their non-coding regions, such as when the 5' UTR of one transcript overlapped the 3' UTR of an upstream gene, were also assigned to different genes. To simplify subsequent analyses, genes containing transcripts that were assigned to more than one gene were discarded.

2.4. *Splice group classification*

Transcripts that aligned to genomic sequence in a similar fashion were assigned to the same transcript structure group. That is, if two transcripts exhibited an identical pattern of alignment to genomic sequence across all inferred, internal splice boundaries, then they were assigned to the same splice group and were treated as the same variant in the DiffMotif analysis described below. Conversely, if an inferred intron in one transcript alignment overlapped an inferred exon in another, then the two transcripts were assigned to different splice groups.

2.5. *Protein sequence analysis*

Conceptual translations of genomic sequence were searched against the BLOCKS database and InterPro member libraries using default parameter settings. The InterPro and BLOCKS databases used were the versions available in November, 2001.

Matches between individual profiles and protein query sequences were modeled as collections of one or more spans within the query protein sequence. For example, a match to a profile that detects a single, continuous stretch of conserved amino acids was represented as a single span with start and end indices indicating the matched region within the query protein. Similarly, matches to profiles that detect repeated or discontinuous stretches of conserved amino acids (such as with BLOCKS) were represented as sets of non-overlapping spans within the matched protein sequence.

2.6. *DiffMotifs*

Results from the transcript classification and protein annotation steps described above were loaded into a database and then mined using a Perl program developed for this study. This program compares profiles of matched motifs between splice groups and notes differences between the types of motifs or number of matched spans per motif profile found in different splice groups. Such differences are defined as "diff motifs," conserved regions detected by BLOCKS or InterPro that differ between splice groups.

2.7. *Visualization*

The Java-based, Neomorphic Genome Software Development Kit was used to create a prototype viewer (ProtAnnot) that displays protein annotations together with gene structures along the genomic sequence axis. Following the identification of genes encoding "diff motifs," ProtAnnot was used to view the exon-intron structure of the transcripts alongside the protein annotations associated with each.

2.8. *Availability*

Data and visualization software produced in this study are available upon request.

3. Results

3.1. *Transcript classification*

To carry out a large-scale survey of alternative splicing, it was necessary to build a collection of alternatively spliced genes for which the exon-intron structure is known. To create this collection, 17 811 transcripts with high-quality alignments to genomic sequence were grouped into genes by comparing the frame and relative locations of their exons along the genomic sequence axis. Here, a gene is defined as a collection of transcripts in which all members have protein-coding sequence in common and therefore encode variant forms of the same protein. Using this criterion, the 17 811 transcripts were sorted into 8317 genes. In 187 cases, the same transcript was assigned to more than one gene. To simplify subsequent analysis, these genes were discarded, leaving 8223 genes and 17 486 transcripts.

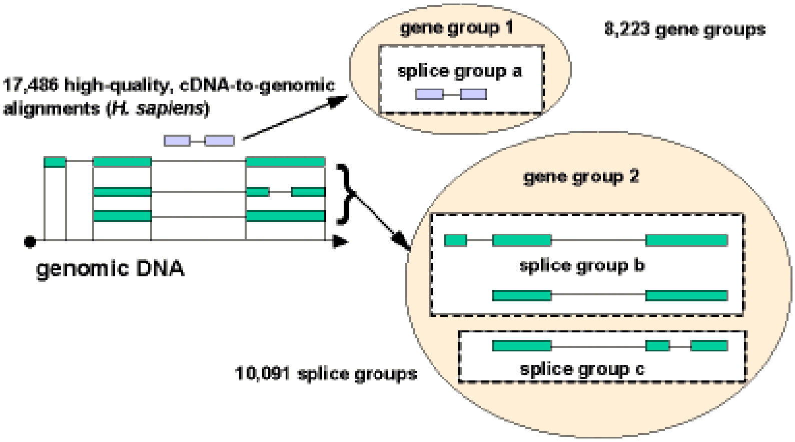

Fig. 1. Transcript classification protocol. Alignments between mRNA and genomic sequence were used to group transcripts into distinct gene groups. In this example, the alignments of four transcripts to genomic sequence are shown. One transcript is located entirely within an intron defined by the other three, and so it is sorted into gene group 1. The other three overlap in-frame along the genomic sequence axis and therefore are assigned to gene group 2. And within gene group 2, two transcripts are spliced according to the same pattern and are placed in the same splice group. The third transcript contains an intron that overlaps with exonic sequence in the other two, and so is assigned to a different splice group within the same gene.

In an additional classification step, transcripts were sorted into splice groups according to the following simple rule: if an inferred intron in one transcript alignment overlapped an inferred exon in another, then the two transcripts were assigned to different splice groups. Using this method, the 17 486 transcripts were sorted into 10 091 distinct splice groups. Thus, the 8223 genes produced 10 091 distinct transcript variants, and 1378 genes (17 percent) were found to produce more than one variant form. See Fig. 1.

3.2. *Protein sequence analysis*

The genome-derived conceptual translations were searched against the BLOCKS and InterPro databases of conserved amino acid sequence profiles. Although the profiles contained in each database are often redundant in that many are based on the same protein families, both methods were used in order to maximize coverage. BLOCKS detected conserved motifs in 5389 (66 percent) of these genes, while the InterPro method recognized 6017 (73 percent). Table 1 presents a list of the 20 most frequently observed InterPro and BLOCKS motifs found among the full set of 8223 genes. We find that genes containing motifs involved in transcriptional regulation, signal transduction, and the immune system are highly represented in the human genome, a result that has been observed previously.[19]

Next, we assessed the frequency with which different protein forms produced by the same gene contained a different pattern of conserved motifs. To do this, we developed a simple data-mining method (DiffMotif) that compares protein annotations associated with transcripts from different splice groups belonging to the same gene. First, the method finds cases where the type of motifs detected differ between variants. For example, if one variant lacks a motif, such as a homeobox domain, that is present in another variant from a different splice group, then the DiffMotif method reports this. The method also finds cases where variants contain the same motif type, but disagree on the number of motif segments or spans. See Fig. 2. For example, a protein variant containing a different number of repeated, WD40 motifs than are found in alternative protein products from the same gene would be reported. However, this method does not detect instances where motifs differ only in length between variants.

Applying the DiffMotif method to the BLOCKS protein sequence annotations identified 340 genes that contained transcript variants with differing patterns of motifs. The InterPro analysis identified 366 genes where the motifs differed between splice groups. Together, both methods recognized a differing pattern of motifs in 475 genes. Thus, alternative transcript structure was associated with detectable changes in motif structure for 34 percent of the 1378 genes containing two or more distinct transcript variants. The remainder either encoded the same protein because alternative transcript structure affected the non-coding regions of the gene or because the number of conserved spans for the motifs contained in different protein products was not changed.

Table 1a. Most frequently detected InterPro motifs. 1504 different InterPro motifs were detected in 6017 genes.

	Description	Accession	# genes
1	Proline-rich region	IPR000694	457
2	Zinc finger, C2H2 type	IPR000822	239
3	Rhodopsin-like GPCR superfamily	IPR000276	198
4	Eukaryotic protein kinase	IPR000719	188
5	Serine/Threonine protein kinase	IPR002290	186
6	Tyrosine protein kinase	IPR001245	176
7	immunoglobulin and major histocompatibility complex domain	IPR003006	166
8	Immunoglobulin subtype	IPR003599	123
9	Homeobox domain	IPR001356	107
10	G-protein beta WD-40 repeats	IPR001680	101
11	RING finger	IPR001841	96
12	EF-hand	IPR002048	90
13	RAS small GTPases, Rab subfamily	IPR003579	90
14	RNA-binding region RNP-1 (RNA recognition motif)	IPR000504	89
15	RAS small GTPases, Ras subfamily	IPR003577	80
16	RAS small GTPases, Rho subfamily	IPR003578	75
17	Ras GTPase superfamily	IPR001806	74
18	KRAB box	IPR001909	73
19	EGF-like domain	IPR000561	70
20	Src homology 3 (SH3) domain	IPR001452	69

Table 1b. Most frequently detected BLOCKS motifs. 1480 different BLOCKS motifs were detected in 5289 genes.

	Description	Accession	# genes
1	Zinc finger, C2H2 type	IPB000822	219
2	Rhodopsin-like GPCR superfamily	IPB000276	189
3	Tyrosine kinase catalytic domain	IPB001245	184
4	Protein kinase C-terminal domain	IPB000961	180
5	MAP kinase	IPB003527	172
6	Kinase associated domain 1	IPB001772	171
7	PAK-box /P21-Rho-binding	IPB000095	169
8	PKN/rhophilin/rhotekin rho-binding repeat	IPB000861	158
9	POLO box duplicated region	IPB000959	147
10	High mobility group proteins HMG1 and HMG2	IPB000135	142
11	Epidermal growth-factor receptor (EGFR), L domain	IPB000494	142
12	DM DNA binding domain	IPB001275	139
13	Transforming protein P21 RAS signature	PR00449	116
14	Synapsin	IPB001359	113
15	Homeobox domain	IPB001356	104
16	Immunoglobulin and major histocompatibility complex domain	IPB003006	103
17	Proline rich extensin signature	PR01217	98
18	Lambda and other repressor helix-turn-helix signature	PR00031	89
19	CUT domain	IPB003350	87
20	Wilm's tumour protein signature	PR00049	84

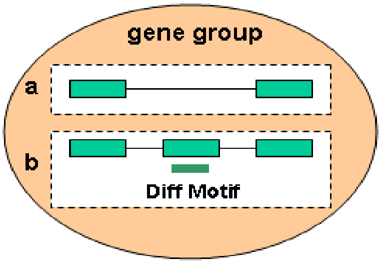

Fig. 2. DiffMotifs protocol. Splice groups associated with each gene were compared as shown in this example. In this example, a single gene group contains two splice groups, containing one variant each. Here, exon skipping in splice group (a) deletes a conserved motif that is present in splice group (b). This difference in motif profiles between the two splice groups is recorded as a Diff Motif for this gene.

3.3. *Gene-by-gene analysis*

Gene-level analysis was then done for the 475 genes identified by the DiffMotif technique. To facilitate this analysis, an interactive, Java-based visualization tool (ProtAnnot) was developed which displays protein motifs superimposed on alignments between transcripts and genomic sequence. Using this tool, hypotheses regarding the function of alternative transcript structure were developed for some of the genes in which the functional significance of the affected motif is known, and examples of these are presented below.

Although this tool is briefly described elsewhere,[20] we describe it in detail here to aid readers in interpreting the following figures. As with other genome browser-like applications, ProtAnnot presents gene structures as rectangles indicating exons linked by line segments indicating introns. Translated regions are shown as filled rectangles, while untranslated 3′ and 5′ regions are shown as unfilled rectangles. The fill color for coding exons indicates the frame of translation relative to the first base of the genomic sequence axis. Thus, same-color exons that overlap along the genomic sequence axis are translated in the same genomic frame and contribute the same sequence of amino acids to their respective protein products. Amino acid motifs are shown as linked spans beneath the exons that encode them.

ProtAnnot summarizes the number of transcripts which contain exonic sequence at each base position as a series of blocks of varying height shown at the bottom of the display. We have found that this feature is extremely useful when attempting to interpret a complex locus with many variants. As observed by Edward Tufte,[21] the human eye is “naturally practiced in detecting deviations from the horizon,” such as the relative differences in exon summary block height that become apparent when scanning ProtAnnot figures from left to right. By taking advantage of this aspect of human visual perception, the exon summary makes often complex splicing patterns easier to interpret.

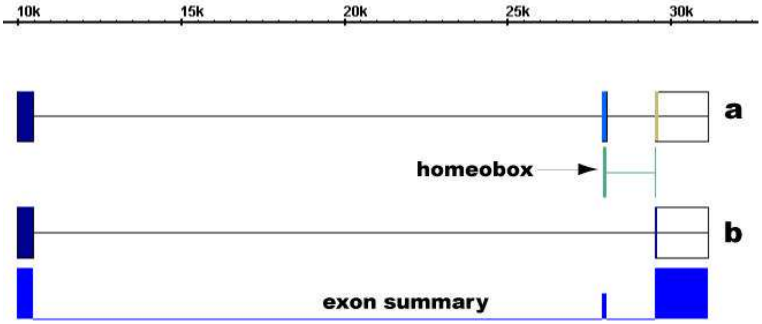

Fig. 3. Alternative splicing at the MEOX1 locus deletes a conserved, homeobox motif. An annotated screen capture from the ProtAnnot visualization tool is shown here. One transcript (a) contains a 3′ homeobox-like motif while the other (b) does not, a result of alternative splicing. The 5′-most exon of each transcript is translated in the same frame, but the final exon on the 3′ end is translated in different frames in the two different variants shown here. The two MEOX1 variants share a common amino terminus but differ in the C-terminal regions. Transcripts shown are: (a) NM_004527.1 and (b) NM_013999.1. A match to homeobox motif (InterPro accession IPR001356, Pfam accession PF00046) is shown as two linked spans underneath the exons encoding it.

3.3.1. *MEOX1*

MEOX1 (mesenchyme homeobox 1, also called MOX1) is located in the BRCA1 region on chromosome 17 and is the human homolog of the mouse Mox1 homeobox-containing gene.[22] MEOX1 interacts with homeobox-containing transcription factor Pax and is thought to be involved in axial skeleton development.[23] MEOX1 is reported to encode at least two distinct proteins, a longer form Fig. 3(a) recognized by BLOCKS and InterPro homeobox profiles and a shorter form Fig. 3(b) that was found to contain no recognizable protein motifs. The shorter form lacks an internal exon (173 bp) that contains part of the homeobox motif. Exon-skipping in this form introduces a frameshift in the final exon, indicated by the different fill colors for this exon in the two different forms shown in the figure. Thus, both proteins share a common N-terminal region but differ in their C-termini.

Homeobox-containing proteins bind to DNA via the homeodomain often as hetero- or homodimers in complex with other homeodomain-containing proteins.[24] Thus the longer MEOX1 form appears competent to interact with DNA and therefore is probably capable of regulating transcription, while the shorter form most likely is not since it lacks a homeodomain. However, if the N-terminal region of MEOX1 permits the protein to interact with other proteins involved in regulating MEOX1, then the shorter form which lacks a homeobox but which retains the amino-terminal region may serve as a negative regulator of MEOX1 function. Alternatively, the short form of MEOX1 may represent an aberrant splicing event or other artifact and may have no function.

To investigate this further, we used the U.C.S.C. genome browser Web site to inspect the MEOX1 locus.[25] We used its human spliced EST, non-human EST, and non-human mRNA tracks to search for MEOX1 sequences which also omit this internal exon. We found none, suggesting that the exon-skipped variant is either extremely rare or that it is indeed an artifact of some sort. To distinguish these possibilities, further sequencing and functional analysis of MEOX1 cDNAs would need to be done to determine whether the shorter form is a true variant and whether it is competent to interact with known MEOX1 protein partners.

3.3.2. *AIRE*

The human AIRE (autoimmune regulator) locus encodes a DNA-binding protein implicated in immune cell transcriptional regulation.[26] Mutations in this gene are responsible for autoimmune polyendocrino-pathy candidiasis ectodermal dystrophy (APECED), a multi-systemic autoimmune disorder. As shown in Fig. 4, the AIRE gene encodes at least three variants due to alternative promoter use and optional splicing of an internal intron. Removal of this intron introduces a frame-shift in the downstream coding region, resulting in deletion of one copy of repeated PHD finger motif. The other two variants, which retain this intron, contain two copies of the PHD motif. In addition, the longer form contains an N-terminal SAND domain not present in the other two. Examination of the two shorter forms revealed that both contain stop codons in all three reading frames upstream of the annotated start codon, suggesting that these shorter variants encode full-length proteins. In

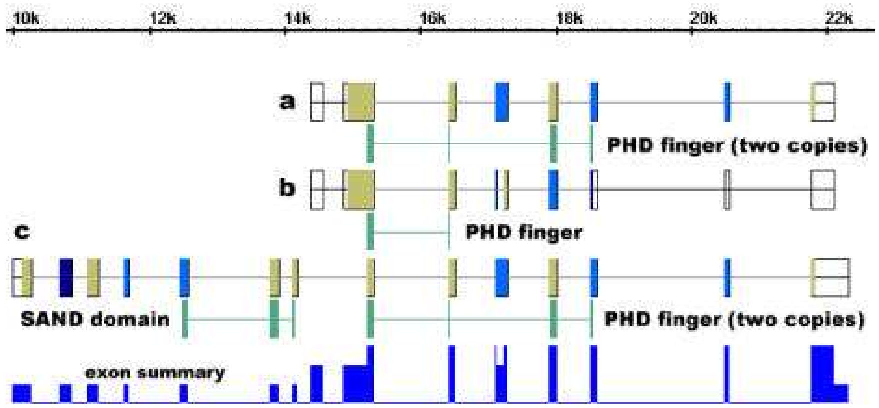

Fig. 4. Alternative transcript structure at the AIRE locus involves alternative promoter choice and optional splicing of a cassette intron. Use of an alternative upstream promoter produces a variant containing a SAND domain (InterPro accession IPR000770, Pfam accession PF01342) that is absent from the other two forms. All three forms contain one or two repeated PHD motifs (IPR001965, PF00628), but removal of a cassette intron in one of the shorter forms introduces a frame-shift that eliminates the second, downstream PHD motif. Variants shown are (a) NM_000658.1 (b) NM_000659.1 and (c) NM_000383.1.

addition, the boundaries of the optionally spliced, cassette intron conform to the expected GT-(intron)-AG splice site consensus sequences.

As described for MEOX1, we examined the AIRE locus using the U.C.S.C. genome browser, but found no spliced ESTs covering the regions affected by alternative splicing or alternative promoter choice. Given the sparseness of EST annotations in this region, we can not with any confidence assess the relative expression levels of the different AIRE variants using purely computational means.

Transfection studies have revealed that the AIRE protein is localized both to speckled domains in the nucleus as well as to cytoskeletal filaments in the cytoplasm.[27] In addition, it has also been shown that the AIRE protein can form homodimers and homotetramers *in vitro* and that these multimeric forms can be detected in thymic extracts, suggesting that these multimeric AIRE protein complexes are also present *in vivo*.[26] A recent deletion analysis revealed that the SAND domain is essential for nuclear targeting.[28] Deletion of the PHD motifs did not disrupt nuclear targeting but did cause changes in the speckled pattern of AIRE protein complexes in the nucleus. These results suggest that the shorter variants which lack the SAND domain are probably not capable of entering the nucleus on their own. If they do enter the nucleus, they may do so in association with the longer form which contains the SAND domain. Similarly, the variant in which the PHD motif is modified may play a role in altering the ability of the protein to bind to its as-yet unknown nuclear targets. Thus, the AIRE gene may represent an example of how alternative splicing may play a role in regulating gene function. We suggest that researchers studying this gene analyze the relative expression levels of these three variants and attempt to deduce how these three forms may interact to regulate gene expression.

3.3.3. *PLAT*

PLAT encodes tissue-type plasminogen activator, an extracellular serine protease that converts inactive plasminogen to plasmin, another protease that attacks fibrin, the fibrous component of blood clots. Thus PLAT, along with plasmin, is a component of a proteolytic cascade that results in clearing of blood clots.[29] PLAT has also been implicated in metastasis through its interaction with the tumor suppressor protein mastin, a serine protease inhibiter.[30] As shown in Fig. 5, this gene encodes at least two distinct variants. Both forms contain conserved serine protease, Kringle, and epidermal growth factor homology regions, but one form (5b) contains a fibronectin type I motif not present in the other. The variant shown here which lacks this motif (5a) was isolated from a cDNA library (NIH_MGC_20) made from melanotic melanoma, a form of skin cancer.

As before, we used the U.C.S.C. genome browser to investigate this locus further. Specifically, we searched for other sequences that exhibit an identical splicing pattern to the exon-skipped variant (5a). The spliced EST track in this browser revealed several such sequences, all of which were derived from the same melanotic

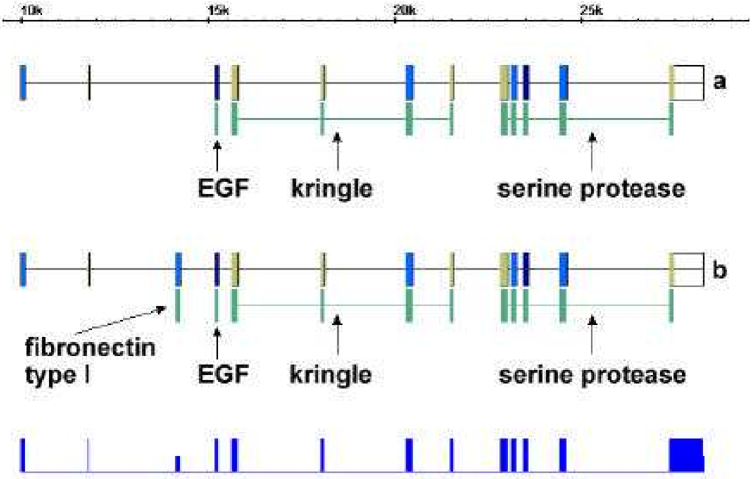

Fig. 5. Exon skipping at the PLAT locus (Tissue type plasminogen activator) deletes a conserved Fibronectin type I Motif. Coding region exons are shown for two variants produced by the PLAT locus. An additional 5′ non-coding region exon 15 kb upstream is not shown. Each variant contains EGF-like motifs (InterPro accession IPR000561, Pfam accession PF00008), kringle motifs (IPR000001, PF000521), and serine protease motifs (IPR001254, PF00089), but one form (a) lacks a type I fibronectin motif (IPR000083, PF00039) which is present in the other. Variants shown here are Genbank accessions (a) BC002795.1 and (b) NM_000930.1.

melanoma library. The human mRNA track contained another just one sequence that showed an identical splicing pattern to the exon-skipped form. This sequence (accession X02901) was originally isolated from human Detroit 562 cells, a pharynx carcinoma cell line.[31] We found no example of this particular variant in any libraries made from normal tissues, however. An on-line blastn search of Genbank confirmed this. This analysis raises the possibility that this particular variant, which appears to be especially abundant in melanotic melanoma, may play a role in the pathology of cancer cells.

3.3.4. *CD79b*

CD79b is a component of the B-cell receptor complex. In combination with membrane-bound antibody and the CD79a protein, it participates in B-cell activation upon binding of the B-cell membrane-bound antibody to its specific antigen. As shown in Fig. 6, this gene encodes two variants which both contain a C-terminal region ITAM motif, which is involved in signal transduction.[32] One variant, however, lacks an amino terminal region Ig_MHC (Immunoglobulin Major Histocompatibility) motif that is present in the other. This motif is thought to be involved in protein-protein interactions, suggesting that the variant which lacks it is unable to interact with some of the receptor's normal protein partners. In fact, transfection studies have revealed that the variant that lacks the Ig_MHC motif

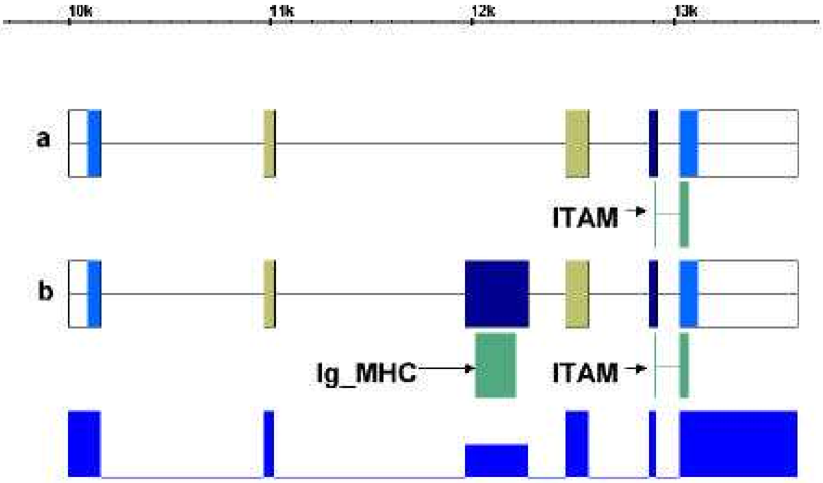

Fig. 6. Exon skipping at the CD79b locus deletes an Ig_MHC motif. Two variants produced at the CD79b locus contain C-terminal region ITAM motifs (immunoreceptor tyrosine-based activation motif; InterPro accession IPR003110; Pfam accession PF02189) but one variant (a) lacks an upstream Ig_MHC motif (Immunoglobulin major histocompatibility domain; InterPro accession IPR003006; Pfam accession PF00047). The two variants shown are (a) NM_021602.1 and (b) NM_000626.1.

is incapable of forming a functional B-cell receptor complex.[33] Interestingly, this variant's expression increases upon B-cell receptor stimulation.[33] If its abundance relative to the other form is also increased, then it is possible that exon-skipping in variant Fig. 6(a) may be regulated and that its differential expression upon B-cell activation may serve to modulate the normal function of the receptor.

4. Conclusions and Discussion

This report demonstrates that amino acid profile libraries such as BLOCKS and InterPro are useful tools for analyzing how alternative transcript structure affects conserved regions in the encoded proteins. Using a simple data-mining approach that examines the differential pattern of conserved motifs present in distinct variants produced by the same gene, it was found that for 475 out of 1378 genes (34 percent), changes in transcript structure coincided with changes in conserved regions in the encoded proteins.

One limitation of this approach is that our method does not analyze the precise regions that are changed between variants. In at least ten genes that were flagged by the DiffMotif method, we found that the difference in conserved motifs was due to relative truncations in the 5′ coding regions of transcripts that were assigned to different splice groups because of differences in splicing further downstream. Although all the transcripts analyzed here were thought by their submitters to encode full-length proteins, it is possible that the submitters were incorrect and the full-length cDNAs for some variants have yet to be isolated. Because we are not certain whether these transcripts are truly complete, we do not include them our final estimate of the percentage of genes in which conserved regions are affected by alternative splicing. Also, we included one commonly occurring motif (InterPro accession IPR000694, Proline-rich region) in our analysis that is reported by the InterPro Web site to lack specificity, and so we cannot be sure that every proline-rich region it detected in our test set should be considered. Matches to this particular profile generated a "diff motif" call for 17 genes. Taking these limitations into account, we estimate that 30 percent of the alternatively spliced genes in our test set contained different profiles of motifs as a result of alternative transcript structure.

Like all such estimates, this is only a rough appraisal of the true impact of alternative splicing on gene function. For example, the method employed here does not detect cases where single spans of conserved amino acids are shortened as a result of alternative splicing. Another limitation is that in some cases, the affected motif was detected using profiles that were originally built using alignments that included members of our test set. However, we are able to say with confidence that the regions affected by alternative splicing are indeed conserved, since these profiles were typically developed using homologs of these same proteins from multiple species.

Using a collection of over 17 000 high-quality, cDNA-to-genomic sequence alignments to define genes and transcript variants, we found that approximately 17 percent of our test set of 8223 genes produced multiple transcript variants. This number includes transcript variation due to alternative splicing, alternative transcriptional promoter choice, and alternative polyadenylation. However, this number is lower than other estimates, which typically have used ESTs in addition to full-length cDNAs to calculate transcript variation frequencies. Since the goal of this study was to examine the impact of alternative transcript structure on protein sequences, it was necessary to restrict the analysis to high-quality alignments. Thus, our observation that 17 percent of human genes product multiple variants represents a potentially useful lower bound for estimating alternative transcript frequency in the human genome.

A protein domain visualization tool that displays conserved motifs in the context of genomic sequence was used to investigate specific examples of how alternative transcript structure deletes conserved motifs in the encoded proteins. In the case of PLAT and MEOX1, in which exon-skipping deletes conserved motifs, EST data presented in the U.C.S.C. genome browser suggests that the exon-skipped form is a minor species compared to the variants that retain the deleted motif. Similarly, the exon-skipped variant of CD79b is reported to be a minor species compared to the non-exon-skipped form.[33] These observations, although anecdotal, are consistent with a previous report showing that among genes which exhibit exon-skipping events, the non-skipped variant typically is more abundant.[34]

It is possible that for some genes, these minor species may function to modulate or dampen the activity of the more abundant form. That is, they may function as weakly dominant negatives, perhaps by sequestering other proteins that interact with both forms. This possibility seems especially likely in the case of the exon-skipped form of CD79b, since its abundance increases upon B-cell activation. In the case of the PLAT variant that lacks the upstream type 1 fibronectin motif, this particular variant may be the result of aberrant splicing events associated with melanotic melanoma, the source tissue for the cDNA library from which this variant was isolated and in which it appears to be most abundantly expressed. However, it should be noted that bioinformatic analyses such as ours can only suggest hypotheses, not prove them. In light of this, we suggest that groups interested in melanoma investigate whether or not the exon-skipped variant of PLAT we describe here is abundant in normal tissues.

It is attractive to speculate that alternative transcript structure affecting coding regions generates biologically meaningful functional diversity in the encoded protein products. However, it is possible that in some cases, alternative splicing and other forms of alternative transcript production that affect the coding region actually exert their effects at the level of RNA, such as by differential inclusion of RNA sequences involved in message localization and stability. Although this study does not resolve this issue, the high incidence of differential motif structure coinciding with alternative transcript structure provide support for the first possibility, since such motifs represent conserved regions and therefore are likely to impact protein function.

We have shown that analyzing variant protein forms with respect to protein motifs is feasible and can be informative whenever the functional or structural significance of the motifs is known. Furthermore, detailed examination of genes producing proteins with distinct motif profiles can suggest hypotheses regarding the functional significance of alternative transcript structure, as described for MEOX1, AIRE, PLAT, and CD79b. However, when the functional or structural significance of the affected protein sequence motif is not known, biological interpretation becomes difficult or impossible. In light of this, future efforts will incorporate other types of protein annotations, such as structure-based annotations, in an effort to gain a deeper understanding of how alternative transcript structure affects gene function.

Acknowledgments

Ray Wheeler, David Kulp, and Alan Williams wrote the genome analysis software that produced pslayout alignments and conceptual translations of genomic sequence. Gangwu Mei provided InterPro analysis results. We are especially grateful to the BioPerl and Postgresql open source communities for programming libraries and software used in this study.

References

1. E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature* **409**, 860–921 (2001).
2. A. A. Mironov *et al.*, "Frequent alternative splicing of human genes," *Genome Res.* **9**, 1288–1293 (1999).
3. J. Xie and D. L. Black, "A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels," *Nature* **410**, 936–939 (2001).
4. E. Stickeler *et al.*, "Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis," *Oncogene* **18**, 3574–3582 (1999).
5. R. C. Chan and D. L. Black, "Conserved intron elements repress splicing of a neuron-specific c-src exon *in vitro*," *Mol. Cell. Biol.* **15**, 6377–6385 (1995).
6. E. I. Rogaev *et al.*, "Analysis of the 5′ sequence, genomic structure, and alternative splicing of the presenilin-1 gene (PSEN1) associated with early onset Alzheimer disease," *Genomics* **40**, 415–424 (1997).
7. J. M. Verdi *et al.*, "Mammalian NUMB is an evolutionarily conserved signaling adapter protein that specifies cell fate," *Curr. Biol.* **6**, 1134–1145 (1996).
8. S. Rensen *et al.*, "Expression of the smoothelin gene is mediated by alternative promoters," *Cardiovasc. Res.* **55**, 850–863 (2002).
9. T. Chen *et al.*, "A novel Dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with active de novo methylation," *J. Biol. Chem.* **277**, 38746–38754 (2002).
10. E. Beaudoing and D. Gautheret, "Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data," *Genome Res.* **11**, 1520–1526 (2001).
11. E. Beaudoing *et al.*, "Patterns of variant polyadenylation signal usage in human genes," *Genome Res.* **10**, 1001–1010 (2000).
12. D. Gautheret *et al.*, "Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering," *Genome Res.* **8**, 524–530 (1998).
13. C. Iseli *et al.*, "Long-Range Heterogeneity at the 3′ Ends of Human mRNAs," *Genome Res.* **12**, 1068–1074 (2002).
14. J. K. Taylor *et al.*, "Inhibition of Bcl-xL expression sensitizes normal human keratinocytes and epithelial cells to apoptotic stimuli," *Oncogene* **18**, 4495–4504 (1999).
15. J. G. Henikoff *et al.*, "Blocks-based methods for detecting protein homology," *Electrophoresis* **21**, 1700–1706 (2000).
16. R. Apweiler *et al.*, "The InterPro database, an integrated documentation resource for protein families, domains and functional sites," *Nucleic Acids Res.* **29**, 37–40 (2001).
17. M. Ashburner *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat. Genet.* **25**, 25–29 (2000).
18. W. J. Kent and D. Haussler, "Assembly of the working draft of the human genome with gigassembler," *Genome Res.* **11**, 1541–1548 (2001).
19. M. Cline *et al.*, in *Pac. Symp. Biocomput.*, ed. R. B. Altman *et al.* (World Scientific, Kauai, Hawaii), pp. 127–138 (2002)
20. A. Loraine and G. Helt, "Visualizing the genome: techniques for displaying human genome data," *BMC Bioinformatics* **3**, 19–27. (2002).
21. E. R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, Connecticut, Graphics Press, 1986).
22. P. A. Futreal *et al.*, "Isolation of a diverged homeobox gene, MOX1, from the BRCA1 region on 17q21 by solution hybrid capture," *Hum. Mol. Genet.* **3**, 1359–1364 (1994).

23. D. Stamataki *et al.*, "Homeodomain proteins Mox1 and Mox2 associate with Pax1 and Pax3 transcription factors," *FEBS Lett.* **499**, 274–278 (2001).
24. S. Khorasanizadeh and F. Rastinejad, "Transcription factors: the right combination for the DNA lock," *Curr. Biol.* **9**, R456–R458 (1999).
25. W. J. Kent *et al.*, "The human genome browser at UCSC," *Genome Res.* **12**, 996–1006 (2002).
26. P. G. Kumar *et al.*, "The autoimmune regulator (AIRE) is a DNA-binding protein," *J. Biol. Chem.* **276**, 41357–41364 (2001).
27. C. Rinderle *et al.*, "AIRE encodes a nuclear protein co-localizing with cytoskeletal filaments: altered sub-cellular distribution of mutants lacking the PHD zinc fingers," *Hum. Mol. Genet.* **8**, 277–290 (1999).
28. C. Ramsey *et al.*, "Systematic mutagenesis of the functional domains of AIRE reveals their role in intracellular targeting," *Hum. Mol. Genet.* **11**, 3299-3308 (2002).
29. L. Stryer, *Biochemistry* (New York, W. H. Freeman and Company, 1988).
30. S. Sheng *et al.*, "Tissue-type plasminogen activator is a target of the tumor suppressor gene maspin," *Proc. Natl. Acad. Sci. USA* **95**, 499–504 (1998).
31. H. Kagitani *et al.*, "Expression in E. coli of finger-domain lacking tissue-type plasminogen activator with high fibrin affinity," *FEBS Lett.* **189**, 145–149 (1985).
32. S. Cassard *et al.*, "Regulation of ITAM signaling by specific sequences in Ig-beta B cell antigen receptor subunit," *J. Biol. Chem.* **271**, 23786–23791 (1996).
33. M. Koyama *et al.*, "The novel variants of mb-1 and B29 transcripts generated by alternative mRNA splicing," *Immunol. Lett.* **47**, 151–156 (1995).
34. W. A. Hide *et al.*, "The contribution of exon-skipping events on chromosome 22 to protein coding diversity," *Genome Res.* **11**, 1848–1853 (2001).

Ann E. Loraine received her Ph.D. degree in Molecular and Cell Biology from the University of California at Berkeley in 1996, supervised by Prof. Wilhelm Gruissem. Following this, she did postdoctoral work in bioinformatics at the Berkeley Drosophila Genome Project and FlyBase. She has worked at Affymetrix as a Staff Bioinformatics Scientist since December, 2000.

Gregg A. Helt received his Ph.D. degree in Molecular and Cell Biology from the University of California at Berkeley in 1997, supervised by Prof. Gerald Rubin. Gregg Helt served as Chief Technology Officer for Neomorphic Software, which he co-founded in 1996. Following Neomorphic's acquisition by Affymetrix in 2000, Gregg Helt re-joined the new company as Principal Scientist. Since then, he has developed several visualization applications for genomics, including ProtAnnot, described above, and Unibrow, a general purpose genome viewer.

Melissa Cline is a member of the Gene Characterization group at Affymetrix. She received her Ph.D. in Bioinformatics from University of California, Santa Cruz, where her research involved protein structure prediction using hidden Markov models, work she applied successfully in three Critical Assessment of Techniques for Protein Structure Prediction (CASP) contests. Her dissertation was on the prediction of accurate regions in sequence alignments. Prior to that, Melissa spent seven years in industry as a software engineer.

Michael A. Siani-Rose is currently the Manager of the Gene Characterization Group at Affymetrix Corporation, responsible for annotating genomes for Affymetrix microarray products. His bioinformatics interests lie in developing new protein-level annotations for the elucidation of the roles of proteins in disease. Prior to his bioinformatics work at Affymetrix, he worked as a computational chemist in drug discovery efforts at Kosan Biosciences, Gryphon Therapeutics, and Chiron. His graduate work was in artificial intelligence (U.C. San Diego) and undergraduate work in chemical engineering (University of Rochester).