EST comparison indicates 38% of human mRNAs contain possible alternative splice forms

David Brett[a,*], Jens Hanke[a], Gerrit Lehmann[a], Sabine Haase[a], Sebastian Delbrück[d], Steffen Krueger[b], Jens Reich[a], Peer Bork[a,c]

[a] *Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Strasse 10, Berlin-Buch 13125, Germany*
[b] *AGOWA GmbH, Gleinicker Weg 185, D-12489 Berlin, Germany*
[c] *EMBL, Meyerhofstr. 1, 69012 Heidelberg, Germany*
[d] *GenProfile AG, Robert-Rössle-Str. 10, 13125 Berlin, Germany*

Received 10 March 2000

Edited by Takashi Gojobori

Abstract Expressed sequence tag (EST) databases represent a large volume of information on expressed genes including tissue type, expression profile and exon structure. In this study we create an extensive data set of human alternative splicing. We report the analysis of 7867 non-redundant mRNAs, 3011 of which contained alternative splice forms (38% of all mRNAs analysed). From a total of 12 572 ESTs 4560 different possible alternative splice forms were detected. Interestingly, 70% of the alternative splice forms correspond to exon deletion events with only 30% exonic insertions. We experimentally verified 19 different splice forms from 16 genes in a total subset of 20 studied; all of the respective genes are of medical relevance.

Key words: Alternative splicing; Disease gene; Database

1. Introduction

There are over 1.5 million human expressed sequence tags (ESTs) in the public databases representing a wide variety of different tissue types, including diseased, normal and immortalised cell lines. In addition many different time points in human development are covered (from embryos to late adulthood). This variety of ESTs provides an ideal resource to examine a number of biological questions concerning gene expression and structure. One such question is that of alternative splicing (AS). AS is an important mechanism allowing tissue-specific or temporal expression of a novel gene form. AS allows one pre-mRNA to be processed into many different mature forms within a cell, each of which can have distinct functions. AS allows tissues to express a single splice variant or a ratio of several forms. Estimates of AS range from 5 to 30% for specific tissue types [1,2]. AS has also been shown to be specifically associated with disease phenotypes [3,4]. The purpose of this study was threefold. Firstly, to develop a computational method by which large numbers of human AS forms could be detected by comparison with ESTs. Secondly, to create a database of novel candidate AS forms. This could act as a primary resource to examine the alternative splicing of specific genes of interest or as diagnostic markers of disease tissue. Finally, to experimentally validate the predictions on a subset of medically relevant genes. This not only enhances experimental identification of splice sites but also hints at the accuracy of the method.

2. Materials and methods

2.1. Computational approach

Twelve thousand and thirty-three radiation hybrid mapped mRNA sequences were downloaded from the current release (20.1.2000) of the NCBI Genemap (ftp://ncbi.nlm.nih.gov/repository/genemap/Mar1999/) (see Fig. 1). To remove redundant sequences with identical sequences but different annotation, the script 'nrdb' was used from the same NCBI site. This reduced the total to 10 226. To remove partial sequences and published truncated forms, each sequence was matched against the EST database with BLASTN [5]. If an EST matched two or more sequences with an expect value greater than 1e−30, only the best scoring sequence was retained in the data set; this reduced the total to 7867 mRNAs. To extract candidate alternative splice forms, in-house written software compared end positions of ESTs with those of the remaining genes from GenBank. The BLASTN program was used with an Expect value 1e−30 without gaps. To prevent frameshifts in the ESTs from producing false alignments BLASTN parameter $X = 1$ was used. To reduce the risk of contamination of our set with pseudogenes or paralogous sequences, stringent parameters were used. For a possible AS form to be considered the combined alignment of each piece had to have an identity of over 97%. Filter programs to extract possible AS forms and remove false ones were written in Perl and C. The default BLASTN filter DUST was used to exclude simple repeats present in the mRNAs. Further filtering was applied to remove matches representing internally repeated domains within mRNAs. Such situations caused many false positives and indeed, this filter removed 376 candidates.

2.2. Experimental approach

For each of the candidates tested, primers were designed from each end of a candidate EST to allow amplification of possible PCR products from a cDNA multiple tissue panel (MTC) panel 1 (Clontech). In certain cases primers were designed to exons either side of a candidate splice form. For details of PCR conditions see legend to Fig. 2. Complete experimental results including gel pictures and sequences are to be found at the following site: ftp://ftp.bioinf.mdc-berlin.de/pub/database/SPLICE_SITE/method.html.

3. Results

After employing the filters, methods and stringent parameters described in Section 2, we generated 4600 possible candidates. In addition to the expected insertions and deletions of sequence regions, 40 possible rearrangements were detected. However, on closer examination, the majority of these were artificial and were in fact the result of internally repeated domains within the mRNA sequence. For only five possible

*Corresponding author. Fax: (49)-30-9406 2834.
E-mail: dbrett@mdc-berlin.de

PII: S0014-5793(00)01581-7

Step 1. 12033 mRNAs down loaded from the NCBI gene map ftp site (ftp://ncbi.nlm.nih.gov/repository/genemap/Mar1999/).

10228 mRNAs left after filtering of redundant sequences using nrdb script (NCBI)

Step 2. 7867 mRNAs used as the non redundant test set after filtering out gene fragments and truncated forms by selecting sequence with the highest scoring ESTS matches.

4976 candidates AS forms identified prefiltering.

Step 3. 376 errors caused by internal repeats and 40 possible rearrangements removed.

Data base of 4560 candiates from 3011 mRNAs created including 132 disease associated genes

3192 exonic deletions (70%). 1368 exon (30%) insertions (estimated 2.4 % false positive rate within this collection)

Step 4. Sub set of 19 splice forms from 16 genes of medical interest verified experimentally.

Fig. 1. A flow diagram of the bioinformatic method used to extract alternative splice forms found within ESTs from a collection of 12 033 mRNA radiation hybrid mapped sequences.

rearrangements, no obvious explanation was found. As none of these could be experimentally verified, we treated possible rearrangements as false positives. After errors and possible rearrangements were removed this left 4560 candidate splice forms from 3011 human mRNAs. Within this list we identified 1368 inserts and 3192 exon deletions.

The resulting database of AS forms (ftp://ftp.bioinf.mdc-berlin.de/pub/database/SPLICE_SITE/MRNA/humrna.html) contains all candidates ESTs identified and the respective genes. SwissProt disease-associated proteins with their corresponding EST candidate AS forms are labelled in their own table (ftp://ftp.bioinf.mdc-berlin.de/pub/database/SPLICE_SITE/disease475.html).

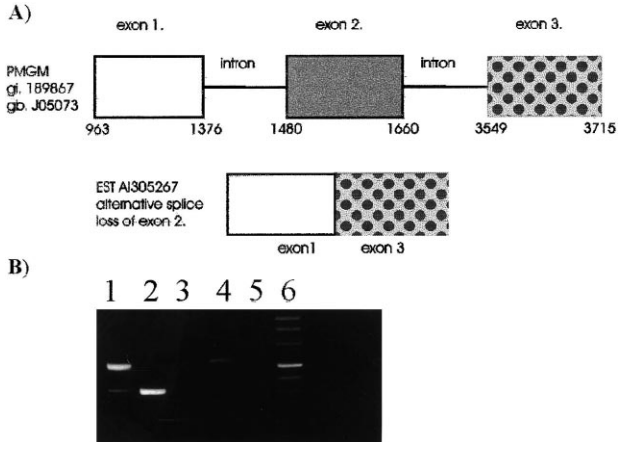

Fig. 2. Human muscle phosphoglycerate mutase deficiency is associated with exercise intolerance, muscle cramps, chronic serum creatine kinase elevation, and recurrent episodes of myoglobinuria. A: The patterned boxes show the normal order of the three exons. The lower of the two graphics shows a complete deletion of exon 2 (178 bp) found within EST AI305267. The numbers are derived from GenBank entry J05073. B: Primers designed to either side of exon 2 produced two bands, a larger approximately 500 bp band plus a lower approximately 320 bp band (lanes 1). The PCR product was concentrated and the bands cut from the gel and run in lane 2 (lower band) and lane 4 (upper band) as a comparison. Lane 6 is the marker lane; the more prominent marker band is approximately 500 bp. Conditions: A 30 cycle PCR was performed at 92°C for 1 min, 58°C for 1 min, 72°C for 1 min with a 72°C extension for 10 min on the last cycle. If these standard conditions failed to give a clear result the 58°C step was varied in a range from 55 to 65°C. Eight different tissues are represented on the panel: from 1 to 8 brain, heart, kidney, liver, lung, pancreas, placenta, skeletal muscle. Products were run out on a 1% agarose gel, excised, cleaned and direct sequencing was attempted.

To get some information on the accuracy of the method, experimental verification was attempted on 20 candidate genes. Approximately half of these genes were chosen from the list in collaboration with a number of different clinical research groups as possible diagnostic markers of disease. The others were chosen as genes of special interest to a particular research group. Despite this directed selection, the 20 genes show no particular bias towards length, chromoso-

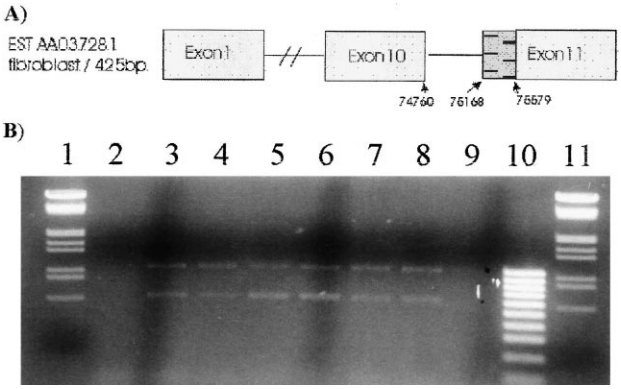

Fig. 3. BIGH3 is a surface recognition protein, inducible by transforming growth factor β [14]. Mutations of the protein have been shown to cause lattice corneal dystrophy [15]. A: In EST AA037281 we identified a novel transcript. After designing primers from exon 10 and 11 the PCR sequence included an extension of the 5′ end of exon 11 by 411 bp. Numbers are derived from genomic sequence (gi 2996635) comparison. B: Both the novel extended sequence (larger band 1064 bp) and the sequence matching the published mRNA (lower band of 653 bp) are seen in seven of eight tissues examined (lanes 3–9). As in Fig. 2B the outer lanes are lambda markers and lane 10 is a 1 kb ladder. Both published sequence (lower band of 653 bp) and extended exon 11 (higher band of 1064 bp) are absent from lane 2 representing brain. Lane 9 representing skeletal muscle produced both bands (very faint).

Table 1
Listed are 16 genes (mRNA or protein entries) with 19 alternative splice forms that differ from the published mRNA sequence

mRNA or protein	Swiss-Prot No. or gi No.	EST	Number of tissues expressed in	Splice form	Position[a]
PMGM	189867	AI305267	8	178 bp deletion	Exon 2
PYR5	J03626	AA331635	7	153 bp deletion	CDS
HLA class I	X13111	AA838034	8	32 bp insert	Exon 5
HPS	U79136	AA700609	8	521 bp deletion	3′ UTR
TM4SF	3152700	AA133048	6	76 bp insert	3′ UTR
Leucophysin	425353	AA005421	8	103 bp insert	5′ UTR
SYK kinase	4507328	AA009756*	6	69 bp insert	CDS
CDC16	603230	AA010055	8	79 bp insert	3′ UTR
GRB7	601890	AA010229	8	219bp insert	Exon 9
Sialyltransferase	3169563	AA015645	8	201 bp insert	5′ UTR
GNB3	M31328	W27169	3	144 bp insert	Exon 10
		AA767991*	3	123 bp deletion	Exon 9
PPAR-γ	P37231	AA053612	5	184 bp deletion	5′ UTR
		AA298089	2	144 bp deletion	5′ UTR
BIGH3	3282161	AA037281	7	411 bp insert	Exon 11
NOS3	266648	AA994247	2	31 bp insert	Exon 23
CROC4	5453625	AA013012	2	96 bp deletion	5′ UTR
			2	252 bp deletion	CDS
SDR-1 homolog	AF035287	AA130752	4	54 bp insert	CDS

Each alternative splice form was experimentally verified in at least two different tissues. A search of the literature for each gene showed that of the 19 splice forms only two were published (EST marked with an *). One of these alternative splice forms detected within the GNB3 protein had previously been studied in connection with hypertension [9]. CDC16 splice form was noted in a direct submission GenBank entry (AF164598) of that gene. For details of particular tissue types, gel pictures and sequence information see WWW page listed in Section 2.
[a]Where CDS is reported as position, no genomic information was available.

mal localisation, function, type, etc., although every candidate has some either direct or indirect influence on a disease phenotype. From the 20 candidate alternatively spliced genes, four produced only the published mRNA sequence (alternative EST sequence not confirmed). From the other 16, we were able to verify 19 splice forms which were alternatives to the published mRNA sequence (see Table 1). All candidate AS forms were present in at least two different tissue types. For an example of experimental verification of deletion or insertion of a sequence within a published exon, see Figs. 2 and 3.

4. Discussion

In our final test set of 7867 mRNAs (Fig. 1), we could extract 4560 candidate AS forms in 3011 human mRNAs. Approximately 38% of the proteins tested seem to be alternatively spliced and on average, 2.75 splice forms exist per spliced gene. In a recent smaller study of 475 proteins, we detected 34% alternative splicing using the same BLAST parameters [6]. Both studies produced a similar ratio of genes exhibiting alternative splicing. However, in the new study the ratio of different ESTs covering a single candidate AS site doubled from on average two per site to four. One possible reason for this is the inclusion of untranslated regions (UTRs) in the new study. The generally increased prevalence of ESTs matching the 3′ UTR of an RNA may give rise to this increase in ratio. It is interesting to note that the inclusion of UTR sequences had little effect on the level of alternative splicing. This is in contrast to single nucleotide polymorphism (SNP) in which selection seems stronger in the coding regions [7]. The rate of AS in exons encoded by UTRs did not significantly differ from coding exons in our study.

The estimate of AS in human mRNAs based on this study is still likely to be an underestimate as published mRNAs evidently exclude intronic information. Novel splice variants could occur in published introns not detectable by this study. In addition the ESTs used as primary source represent on average only four different ESTs per RNA query. Although this is an increase on the previous study, this sampling ratio represents only a small number of the possible tissue types or time points present in vivo. From an SNP study carried out on a similar large number of human mRNAs, we calculated an approximate coverage of 30% of exons within our study set [7]. 30% of the possible splice forms found with ESTs arose through insertion of a sequence. Within this group we estimate 8% of inserts will be false positives (2.4% of all 4560 candidates) due to pre-mRNA or genomic DNA contamination of the EST database. This estimate was derived from an analysis of 137 proteins with their corresponding genomic sequences from the smaller pilot study [6].

The experimental validation of AS in 16 out of 20 genes was performed as a proof of principle of the extraction method (see Table 1). Although this number is too small to make a statistical calculation for the whole set, 80% of predicted splice forms confirm a significant success rate. The remaining 20% (four cases) cannot be taken as a direct reflection of a false positive rate. The failure of the experimental technique to find four of the possible candidates may be a simple function of low abundance of that particular form or its existence in a rare tissue type. In two of the four cases where no alternative splice form was identified (only published mRNA sequence found), the genomic information indicated no contamination with pre-mRNAs or DNA. In the other two cases, no genomic sequence was available. However, it is interesting to note that in all four cases inserts to exons were predicted. In another 10 cases where genomic sequences were available, but no experiment was performed, the candidates were also manually examined for possible pre-mRNA or DNA contamination of the EST sequence. No indication of any of these events was found.

From the 16 candidate genes experimentally verified, 13 have been shown in the literature to have a direct effect in the development of disease phenotypes. In four cases splice forms of the gene are associated with disease. Specific splice

forms of peroxisome proliferator-activated receptor γ (PPAR-γ) have been linked to the development of obesity [8]. GNB3 alternative splicing is associated with a hypertensive phenotype [9]. A novel splice variant of Grb7 is associated with invasive carcinoma [10]. Selective loss of specific HLA splice variants is associated with melanoma progression [11]. In all four of these cases we have found additional novel splice forms. In the other cases (HPS and SYK genes) alternative splicing is known to occur with as yet no direct disease association [12,13]. All these examples show the important role that alternative splicing plays in human disease. In the near future we expect both the number of human mRNAs and ESTs deposited in public access databases to grow considerably. Regular updating of our data set should provide a very useful resource for clinical investigators. Our database differs from a published database of alternative splice forms ASDB (http://cbcg.nersc.gov/asdb) as we used ESTs to find novel human AS forms. ASDB at present uses only key words to extract AS forms that are already published in GenBank or SwissProt. As this article was in preparation a smaller-scale study of alternative splice forms found in ESTs was published. This mirrors our pilot study in that the authors used 392 genes, it also confirms a similar rate of alternative splicing, 38% of genes in this case [16].

In conclusion, this study provides a vast number of human candidate AS forms with a high confidence level. The results are accessible in a database that currently contains 4560 candidate AS forms from 3011 human mRNAs, many of which have a direct role in disease progression.

References

[1] Sharp, P.A. (1994) Cell 77, 805–815.
[2] Sutcliffe, J.G. and Miller, R.J. (1988) Trends Genet. 4, 297–299.
[3] Klamt, B., Koziell, A., Poulat, F., Wieacker, P., Scambler, P., Berta, P. and Gessler, M. (1998) Hum. Mol. Genet. 4, 709–714.
[4] Qi, M. and Byers, P.H. (1998) Hum. Mol. Genet. 7, 465–469.
[5] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, J.D. (1997) Nucleic Acids Res. 25, 3389–3402.
[6] Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbrück, S., Friedrich, L., Reich, J. and Bork, P. (1999) Trends Genet. 15.
[7] Sunyaev, S., Hanke, J., Brett, D., Aydin, A., Zastrow, I., Lathe, W., Bork, P. and Reich, J. (2000) Adv. Protein Chem. 54, 409–437.
[8] Vidal-Puig, A.J., Considine, R.V., Jimenez-Linan, M., Werman, A., Pories, W.J., Caro, J.F. and Flier, J.S. (1997) J. Clin. Invest. 99, 2416–2422.
[9] Siffert, W., Rosskopf, D., Siffert, G., Busch, S., Moritz, A., Erbel, R., Sharma, A.M., Ritz, E., Wichmann, H.E., Jakobs, K.H. and Horsthemke, B. (1998) Nature Genet. 18, 45–48.
[10] Tanaka, S., Mori, M., Akiyoshi, T., Tanaka, Y., Mafune, K., Wands, J.R. and Sugimachi, K. (1998) J. Clin. Invest. 102, 821–827.
[11] Wang, Z., Marincola, F.M., Rivoltini, L., Parmiani, G. and Ferrone, S. (1999) J. Exp. Med. 190, 205–215.
[12] Wildenberg, S.C., Fryer, J.P., Gardner, J.M., Oetting, W.S., Brilliant, M.H. and King, R.A. (1998) J. Invest. Dermatol. 110, 777–781.
[13] Yagi, S., Suzuki, K., Hasegawa, A., Okumura, K. and Ra, C. (1994) Biochem. Biophys. Res. Commun. 200, 28–34.
[14] Skonier, J., Neubauer, M., Madisen, L., Bennett, K., Plowman, G.D. and Purchio, A.F. (1992) DNA Cell Biol. 11, 511–522.
[15] Stewart, H., Black, G.C., Donnai, D., Bonshek, R.E., McCarthy, J., Morgan, S., Dixon, M.J. and Ridgway, A.A. (1999) Ophthalmology 106, 964–970.
[16] Mironov, A., Fickett, J. and Gelfand, M. (1999) Genome Res. 9, 1288–1293.