Alternative splicing of human genes
more the rule than the exception?

Jens Hanke*
hanke@mdc-berlin.de

Dave Brett*
dbrett@mdc-berlin.de

Inga Zastrow*
izastrow@mdc-berlin.de

Atakan Aydin°
aydin@fvk-berlin.de

Sebastian Delbrück°
delbruck@mdc-berlin.de

Gerrit Lehmann*
glehmann@mdc-berlin.de

Friedrich Luft°
luft@fvk-berlin.de

Jens Reich*
reich@mdc-berlin.de

Peer Bork*+
bork@embl-heidelberg.de

+EMBL, Meyerhofstr. 1, 69012 Heidelberg, Germany.

*Max-Delbrück-Center (MDC) for Molecular Medicine, Robert-Rössle-Strasse 10, Berlin-Buch, 13125, Germany.

°Franz-Volhard Clinic at the MDC, Virchow Klinikum, Berlin-Buch, Germany.

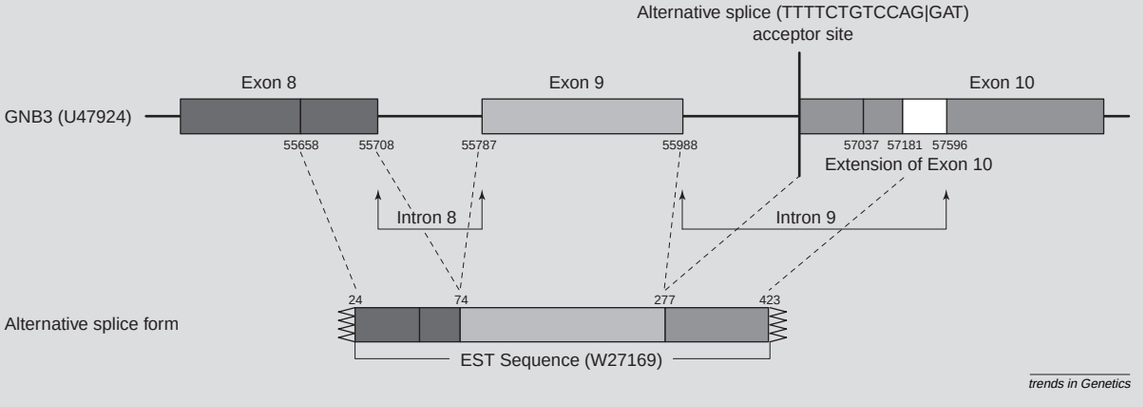
Approximately 5% of all human genes have been estimated to be subjected to alternative splicing (AS)[1,2], although early data indicated an AS level of 30% for genes that are expressed predominantly in the nervous system[3]. AS is an important cellular mechanism that leads to the temporal and tissue-specific expression of unique mRNAs. In some cases, specific splice variants have been associated with human diseases (e.g. Refs 4–6). Databases of human expressed sequence tags (ESTs) derived from more than 600 distinct tissue-specific libraries, and from individuals with large age differences (from embryos to age 75) should, therefore, be a rich source of distinct splice variants.

In order to investigate the level of AS and to search for novel splice variants and candidate targets for diagnostics, we aligned a set of 475 human proteins against six frame translations of about 1.3 million human ESTs using TBLASTN (Ref. 7). The 475 proteins were annotated in the SWISSPROT protein sequence database[8] as disease-associated; they come from all chromosomes and cover all major functional classes, that is, there is no indication of bias in this sample set. To minimize matches of pseudogenes and paralogous sequences, only those ESTs were accepted that shared >97% identity with a protein in a window of 100 amino acids. In order to identify an EST with a possible alternative splice form, a difference in length of match between the query protein sequence and the EST within the TBLASTN alignment was searched for. Successive end positions of each TBLASTN alignment between protein query and translated EST sequence were compared. Small artifactual differences caused by the TBLASTN algorithm were excluded. Where possible, ESTs were compared with genomic sequence. During this comparison, ESTs with insertions that consisted of complete introns were then removed. This avoided detecting possible unprocessed mRNAs. Filters were employed to exclude repeats (REPEAT MASKER, http://ftp.genome.washington.edu/RM/RepeatMasker.html) and vector sequences.

Using these rather stringent criteria, 222 candidate AS sites were predicted (see Fig. 1 for an example) in 162 of the 475 disease-associated proteins. For 137 of the 162 proteins with predicted splice variants, at least partial genomic DNA was available and was used to exclude 18 ESTs that contained complete introns and that might resemble premature mRNAs or unprocessed pseudogenes. As an additional check, we chose ten clones randomly for experimental verification by re-sequencing the original EST clones. To extend these ESTs, tissue-specific RT–PCR was performed using total and polyA mRNA (preparations of mRNA were obtained commercially). In all ten cases, the predicted splice variants were present in at least two distinct tissues from two different individuals (results not shown). Although this number is too low to estimate the general prediction accuracy and the actual protein expression of the respective variants has also to be proven, the correct identification of ten out of ten indicates a high success rate.

The SWISSPROT protein-sequence database provides special feature lines for annotated splice variants that have been previously reported in the literature. Of the 204 predicted splice variants (222 candidates – 18 of them

FIGURE 1. Predicted alternative splicing in GNB3



A splice variant in the gene coding for the G protein GNB3 (Genbank accession no. U47924) has been reported to cause a form of essential hypertension[6]. Next to the known splice variant (in exon 9 of the coding beta 3 subunit) we predict a novel splice variant (in EST W27169, human retina cDNA randomly primed, length 655 bp) that starts in exon 8 and that covers exon 9 completely, ending within 144 bp of intron 9. In GNB3, exons 8 to 10 are shown in different shades of grey, the novel extended exon 10 produced by alternative splicing being the lightest. The EST sequence starts within exon 8, covers all of exon 9 and ends within the novel extended exon 10. A comparison with GNB3 genomic DNA reveals an alternative acceptor site, which provides further evidence of AS. We verified this novel variant experimentally by RT–PCR and confirmed the extension of exon 10 with a 559 bp extension in four tissue types: fetal retina; adipose; colon; and umbilical cord (data not shown).

 PII: S0168-9525(99)01830-2

probably premature mRNAs or pseudogenes), only 24 (12%) were already annotated in SWISSPROT. On the other hand, of the remaining 313 proteins in which no AS was found, 57 (18%) contain previously reported splice sites that are annotated in SWISSPROT, which implies that a considerable number of false negatives exist. Furthermore, ESTs that matched the 475 proteins covered only ~50% of each protein, that is, no conclusion can be made about the other 50%. In addition, each position was covered only by about two distinct EST libraries. This is a very low ratio, in comparison with the large number of human tissues. Thus, we conclude that the level of AS indicated by EST analysis alone (AS occurs in 34% of the proteins studied) might be a significant underestimate. A compilation of the 204 predicted splice variants in 162 disease-associated genes, methodological details and experimental validation of ten candidates, are given at: ftp://ftp.bioinf.mdc-berlin.de/pub/database/SPLICE_SITE/disease475.html.

References

1 Sharp, P.A. (1994) Split genes and RNA splicing. *Cell* 77, 805–815

2 Wolfsberg, T.G. and Landsman, D. (1997) A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* 25, 1626–1632

3 Sutcliffe, J.G. and Milner, R.J. (1988) Alternative mRNA splicing: the Shaker gene. *Trends Genet.* 4, 297–299

4 Stallings-Mann, M.L. *et al.* (1996) Alternative splicing of exon 3 of the human growth hormone receptor is the result of an unusual genetic polymorphism. *Proc. Natl. Acad. Sci. U. S. A.* 93, 12394–12399

5 Liu, W. *et al.* (1997) Silent mutation induces exon skipping of fibrillin-1 gene in Marfan syndrome. *Nat. Genet.* 16, 328–329

6 Siffert, W. *et al.* (1998) Association of a human G-protein beta3 subunit variant with hypertension. *Nat. Genet.* 18, 45–48

7 Altschul, S. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

8 Bairoch, A. and Apweiler, R. (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res.* 27, 310–311

Forcing evolution

Molecular Evolution and Diversity (BBSRC), University of Warwick, UK, 10–11 May 1999

Academic studies of molecular evolution by university-based scientists can have applied relevance to the biotechnology industry, particularly in the area of 'forced evolution'. Communicating that relevance across the sometimes wide divide between evolutionary biologists and practical bench molecular biologists is facilitated by shutting them away together for several days with alcoholic lubrication. The same holds true for facilitating the counterflow of real-world biology into abstract theoretical models. The British Biotechnology and Biological Sciences Research Council (BBSRC) brought together about 80 people from universities and industry to consider this interface in a Molecular Evolution and Diversity Workshop at the University of Warwick.

Life as we know it has not completely explored all of 'sequence space', for the simple reason that life has an evolutionary history, which constrains the possibilities; also, the number of potential sequences is so huge that the universe will not last long enough for natural selection to have looked at them all. We have only begun to scratch the surface of existing natural genetic diversity and to ask how we can improve upon biological molecules for human purposes.

Genetic diversity arises ultimately through mutation, and we began with talks on the mechanisms of mutation and their fates in natural and laboratory populations. Movements of transposable elements are a well known source of mutations. The generation of proteins with new functions by retrotransposon-mediated exon shuffling seemed very plausible after hearing about work on the movements of L1 elements in cultured human cells by John Moran[1] (University of Michigan, USA). It could be that his techniques could be harnessed to build new proteins. Knowledge of the genomic rates of deleterious mutations and the distributions of their effects is crucial to our understanding of basic questions that concern, for example, the evolution of sex and recombination, and more practical problems, such as predicting the response to the selection of laboratory populations. Peter Keightley (University of Edinburgh, UK) reported on experiments in the nematode *Caenorhabditis elegans* to estimate the fitness effects of amino acid-altering point mutations in protein-coding DNA sequences. The vast majority of these sorts of mutations are expected to be deleterious in natural populations, and this is borne out by the observation of strong conservation of such sequences over evolutionary time. He showed that a few mutations had strongly deleterious effects but, surprisingly, a large majority of these deleterious mutations were not detectable in laboratory-fitness assays, suggesting that the distribution of fitness effects is discontinuous.

Many of the workshop's academic participants are engaged in applied research or have links with industry, but Peter Laing of Actinova Ltd (UK) was our only speaker working full-time in industry. He explained Actinova's proprietary covalent display technology for the *in vitro* evolution of protein domains. It became clear to us evolutionists that 'forced evolution' often does not mean evolution as we usually understand it, but rather selection. Much commercial interest is directed at modifying small regions of proteins, such as binding sites, often in directions that natural selection could never go in organisms that are constrained to make a living in the real world. For these

Richard H. Thomas
r.thomas@nhm.ac.uk

Department of Zoology, The Natural History Museum, Cromwell Road, London, UK SW7 5BD.