Published in final edited form as:
Methods Mol Biol. 2011 ; 696: 319–326. doi:10.1007/978-1-60761-987-1_20.

Identification of Alternatively Spliced Transcripts using a Proteomic Informatics Approach

Rajasree Menon[1] and **Gilbert S. Omenn**[1,2]

[1]Center for Computational Medicine and Biology and National Center for Integrative Biomedical Informatics, University of Michigan, Ann Arbor, MI, 48109-2218

[2]Departments of Internal Medicine and Human Genetics, and School of Public Health, University of Michigan, Ann Arbor, MI, 48109-2218

Abstract

We present the protocol for identification of alternatively spliced peptide sequences from tandem mass spectrometry datasets searched using X!Tandem against our modified ECgene resource with all potential translation products and then matched with the Michigan Peptide to Protein Integration scheme. This approach is suitable for human and mouse datasets. Application of the method is illustrated with a study of the Kras activation-Ink4/Arf deletion mouse model of human pancreatic ductal adenocarcinoma.

Keywords

splice variants; splicing; tandem mass spectrometry; X!Tandem; Michigan Peptide to Protein Integration; Ensembl; ECgene; false discovery rate; expect value; pancreatic cancer

Introduction

By means of alternative splicing and post-translational modifications, one gene can generate a variety of proteins. Alternative splice events that affect the protein coding region of the mRNA will give rise to proteins which differ in their sequence and activities. Alternative splicing within the non-coding regions of the RNA can result in changes in regulatory elements such as translation enhancers or RNA stability domains, which may dramatically influence protein expression (1).

Alternative splicing has been associated with such diseases as growth hormone deficiency, Fraser syndrome, cystic fibrosis, spinal muscular atrophy, and myotonic dystrophy (2, 3). In cancers, there are examples of every kind of alternative splicing, including alternative individual splice sites, alternative exons, and alternative introns (4). A number of public alternative splice databases have recently become available, including ASD, HOLLYWOOD, and ASAP II. Each of these repositories contains transcript models that have been constructed from either expression data (ESTs and mRNA) or previous annotations of known proteins. The databases vary in their annotation methods and their

Corresponding Author gomenn@umich.edu.

overall size. One of the larger of these databases is the ECgene database developed by Kim, et al (5). Entries in the database are scored as high, medium, or low confidence reflecting the amount of amassed evidence in support of the existence of a particular alternatively spliced sequence. Evidence is collected from clustering of ESTs, mRNA sequences and gene model predictions.

We have devised a proteomic informatics approach to identify known and novel alternative splice variants. Briefly, we search mass spectrometric data against a custom-built, non-redundant human or mouse database created with translation products using all three reading frames from cDNA sequences taken from ECgene and Ensembl databases (6). The peptide sequences identified are analyzed using Blast and Blat searches and integrated to distinct proteins.

Database of Translated Alternatively Spliced Sequences: The Modified ECgene Database

The target alternative splice variant protein database, the modified ECgene database, was constructed and can be updated by combining the latest Ensembl and ECgene databases (mm8, build 1) for the mouse; the analogous combination generates a modified ECgene database for human studies. Taking alternative splicing events into specific consideration, ECgene combines genome-based EST clustering and the transcript assembly procedure to construct gene models that encompass all alternative splicing events (5). The reliability of each isoform is assessed from the nature of cluster members and from the minimum number of clones required to reconstruct all exons in the transcript.

cDNA sequences from the ECgene database and from the Ensembl database were obtained in FASTA format. Each sequence set was translated separately in three reading frames and the first instance of every protein sequence longer than 14 amino acids was recorded. The cDNA sequences are translated in three reading frames instead of the six frames which are used in translation of genomic double-stranded DNA; cDNAs are made from single-stranded mRNA sequences. Following the three-frame translation, the resulting sequences from each data source were combined and then filtered for redundancy. Preference was given to protein sequences originating from an Ensembl transcript. A collection of common protein contaminant sequences (http://www.thegpm.org/crap) was added to this set. Lastly, all sequences were reversed and appended to the set of forward sequences as an internal control for false identifications. This last step resulted in doubling the total number of entries in the modified ECgene database.

For examples, the mouse ECgene database contains a total of 417,643 splice variants; the Ensembl version 40 database (Each Ensembl release has an integer version number associated with it which is used to identify the correct versions of API, web code and databases that make up that release) has 21,839 mouse genes with 28,110 transcripts, of which there are 10,922 alternative transcripts derived from 4,651 genes. The modified mouse ECgene database contains 10.4 million protein sequences. Similarly, the modified human ECgene database which contains Ensembl 53 version contains 14.2 million protein entries.

Searching Mass Spectral Data Against Alternative Splice Database

The mzXML files containing the spectral information were extracted from mass spectrometric RAW files using ReAdW.exe program (http://tools.proteomecenter.org). The mzXML files were then searched against the modified ECgene database using X!Tandem software **(*7*)**.

Post-Search Analyses

Figure 1 summarizes the analytical work flow of the X!Tandem search results. In brief, the peptides are first integrated to a list of proteins using the MPPI approach described below. Thereafter, all peptides identified from the integrated protein list are searched against the latest protein databases. This step is required due to frequent updates of protein databases. If a match occurs, the peptide is referred as a known peptide; if not, it is considered as a novel peptide. The protein sequences from which these peptides are identified are aligned to the genome to determine the location of the peptides. Next, the known peptides with protein annotations from the latest databases and the novel peptides identified from modified ECgene variants undergo another round of MPPI. A threshold is applied to keep the FDR < 1%. All known protein identifications that were derived from genes with multiple transcripts and the novel identifications from ECgene entries are retained and the other proteins are removed. Hence, the final integrated protein list contains the known and novel alternative splice variants.

Michigan Peptide to Protein Integration (MPPI)

The peptide identifications from the X!Tandem searches were integrated to a final list of proteins using MPPI. Only the peptide identifications within the FDR < 1% limit were used in the MPPI analysis. The MMPI algorithm is as follows:

1. List all peptide matches that fall within an FDR < 1 % (based on X!Tandem expect value)
2. Order peptides by number of spectra matching each peptide.
3. Select peptide with largest number of matching spectra.
4. List all proteins containing this peptide, ranked by decreasing number of total distinct peptides identified, decreasing number of total spectra, increasing expect value, and then increasing protein length.
5. Select the highest ranking protein to be included in the final integrated protein list; if a tie, give preference to Ensembl protein over ECgene protein.
6. Remove all other peptides contained within this protein from the peptide list.
7. Repeat steps 3-6 until no peptides remain in the peptide list.

Sequence Analyses

The peptide identifications from the proteins after the first MPPI analysis were searched against the mouse genome using BLAT (8) and against latest Ensembl and NR databases using NCBI blastp (9). In the case of a novel peptide, the translated splice variant sequence

is aligned against the mouse genome. Thus the location of the peptide within the gene is determined. One can deduce which splice mechanism has generated the novel peptide, including deletion or switch of exons, intron retention, alternate splice site and translation in an alternative reading frame. In rare cases, the ECgene variant sequence from which the novel peptide was identified matches to a conserved chromosomal region with no known genes; if so, the identification is retained only if the novel peptide is from multiple good quality spectra.

Validation of Novel Peptides

If total mRNA of the sample used in the mass spectrometric study is available, an independent validation of the novel splice variant peptides by reverse transcription polymerase chain reaction (RT-PCR) or quantitative RT-PCR can be performed. Specific primers can be designed to amplify precisely the novel peptide sequence using free on-line applications, including Primer3 (http://frodo.wi.mit.edu/primer3/). In a comparative study, for example of tumor versus normal tissue specimens, the qRT-PCR enables assay of differential mRNA expression related to the novel peptide and comparison with evidence of differential expression at the protein level.

Differential Expression of Alternative Splice Variants

In studies where samples under different conditions are analyzed, knowledge of the differential expression of the unique peptides by which an alternative splice variant protein is identified would be very useful. This information would indicate whether the particular variant might be functionally involved in the phenotype associated with the specimen. If the samples are labeled by heavy isotopes or molecular tags, proteomics tools such as LIBRA (10) or XPRESS (10, 11) which are embedded in Trans Proteomic Pipeline (TPP, Institute for Systems Biology, Seattle) (10, 11), can help determine the relative expression of the unique peptide. In addition, spectral counting is a label-free method to estimate protein quantification using peptide identification results from tandem mass spectrometry; no isotopic labeling is required to perform spectral counting.

Annotation of Novel Peptides

To characterize the novel peptides identified, on-line tools including InterProScan or MotifScan can be used. InterProScan combines different protein signature recognition methods from the InterPro consortium member databases into one resource (12). MotifScan scans a sequence against protein profile databases (http://myhits.isb-sib.ch/cgi-bin/motif_scan). The Berkeley Drosophila Genome Project Splice Site Prediction by Neural Network (http://www.fruitfly.org/seq_tools/splice.html) can be used for predicting alternative splice sites which may have generated these novel peptides.

The interactions of the alternative splice variants can be displayed by the Cytoscape MiMI **(*13*)** plugin using parent gene symbols as the input genes. Michigan Molecular Interactions (MiMI) gathers data from well-known protein interaction databases and deep-merges the information. Utilizing an identity function, molecules that may have different identifiers but represent the same real-world object are merged. The Cytoscape MiMI plugin enables one to

connect to the MiMI database and view the interactions (http://portal.ncibi.org/gateway/mimiplugin.html).

Alternative Splice Variant Analysis of a Pancreatic Tumor Dataset

To assess the potential of tumor-associated alternatively spliced gene products as a source of biomarkers in biological fluids, a large dataset of mass spectra derived from the plasma proteome of a mouse model of human pancreatic ductal adenocarcinoma was analyzed **(*14*)**. MS/MS spectra were interrogated for novel splice isoforms using the non-redundant modified ECgene database described above. Among 1278 distinct proteins, this integrated analysis identified 420 distinct splice isoforms, of which 92 did not match any previously annotated mouse protein sequence. Novel variants of muscle pyruvate kinase, malate dehydrogenase 1, glyceraldehyde-3-phosphate dehydrogenase, proteoglycan 4, minichromosome maintenance complex component 9, high mobility group box 2 and hepatocyte growth factor activator are of particular interest for pancreatic cancer. Isotopic labeling of cysteine-containing peptides from tumor-bearing mice and wild-type controls enabled relative quantification of identified proteins. Statistically significant differential expression between tumor-bearing and control mice was noted for peptides from 9 novel alternative splice variant proteins. We validated a subset of seven of the 92 novel peptide sequences, all of which had multiple spectra, with appropriate primers for the corresponding mRNAs, using qRT-PCR of the tissues **(*14*)**.

These results, in this mouse model for pancreatic cancers, show that novel and differentially expressed alternative splice isoforms are detectable in plasma. Such alternatively spliced protein variants may be clues to cancer progression and cancer biology and may become a source of candidate biomarkers.

Notes

The proteomic informatics approach presented here is intended to identify specific alternative splice variants, including novel proteins with differential expression under different conditions. Different organs, tissues, and biofluids may have different splicing propensities and different responses to external or internal stimuli, which will lead to interesting comparative analyses.

A major limitation of the protocol is the large size of the database. Our modified ECgene non-redundant translation product database for the human species contains 14.2 million records; the corresponding database for the mouse species contains 10.4 million records. Searching the mass spectral data against this database takes several days to complete. In addition, when the experimental protocol includes deep fractionation of the proteins (e.g., 162 fractions in the case of the pancreatic cancer-associated plasma sample described above), the computer search time is multiplied to weeks. With very large datasets, sufficient memory is essential and may become apparent only when the server freezes and stops, requiring re-start on the search. Dividing the database into subgroups and searching the mass spectral data against these databases in parallel can reduce the search time appreciably. An alternative is to run the forward and reverse sequences separately.

Another complication is the frequent updating of the Ensembl database. The novel peptide identifications have to be searched against the latest protein sequence database in order to be annotated as novel. Of course, as soon as a novel variant is published and made available to repository users, the novel variants became known splice variants for the next study. We have a series of studies of mouse and human cancers and cell lines in progress using this protocol.

Conclusions

The combined proteomic bioinformatics approach of the modified ECgene database and X! Tandem-MPPI search tools can identify specific known and novel splice variants in tissue and plasma specimens. The study of MS/MS data from the mouse plasma proteome of pancreatic tumor-bearing mice showed many specific known and novel alternative splice variants, some with differential expression between tumor-bearing and wild-type mouse. Differentially-expressed splice variant proteins may influence many yet-to-be-identified cancer-related mechanisms. The data suggest that alternative splice variant proteins are a potentially rich source of candidate biomarkers for cancers and probably for other diseases, as well.

References

1. Bracco L, Kearsey J. The relevance of alternative RNA splicing to pharmacogenomics. Trends in Biotechnology. 2003; 21:346–353. [PubMed: 12902171]
2. Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes & Development. 2003; 17:419–437. [PubMed: 12600935]
3. Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy. Nat Biotech. 2004; 22:535–546.
4. Venables JP. Aberrant and Alternative Splicing in Cancer. Cancer Res. 2004; 64:7647–7654. [PubMed: 15520162]
5. Kim N, Shin S, Lee S. ECgene: Genome-based EST clustering and gene modeling for alternative splicing. Genome Res. 2005; 15:566–576. [PubMed: 15805497]
6. Fermin D, Allen B, Blackwell T, Menon R, Adamski M, Xu Y, Ulintz P, Omenn G, States D. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biology. 2006; 7:R35. [PubMed: 16646984]
7. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004; 20:1466–1467. [PubMed: 14976030]
8. Kent WJ. BLAT---The BLAST-Like Alignment Tool. Genome Res. 2002; 12:656–664. [PubMed: 11932250]
9. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucl. Acids Res. 2004; 32:W20–25. [PubMed: 15215342]
10. Pedrioli PG. Trans-Proteomic Pipeline: A Pipeline for Proteomic Analysis. Proteome Bioinformatics. 2009; 604:213–238.
11. Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotech. 2001; 19:946–951.
12. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. InterProScan: protein domains identifier. Nucl. Acids Res. 2005; 33:W116–120. [PubMed: 15980438]
13. Gao J, Ade AS, Tarcea VG, Weymouth TE, Mirel BR, Jagadish HV, States DJ. Integrating and annotating the interactome using the MiMI plugin for cytoscape. Bioinformatics. 2009; 25:137–138. [PubMed: 18812364]

14. Menon R, Zhang Q, Zhang Y, Fermin D, Bardeesy N, DePinho RA, Lu C, Hanash SM, Omenn GS, States DJ. Identification of Novel Alternative Splice Isoforms of Circulating Proteins in a Mouse Model of Human Pancreatic Cancer. Cancer Res. 2009; 69:300–309. [PubMed: 19118015]

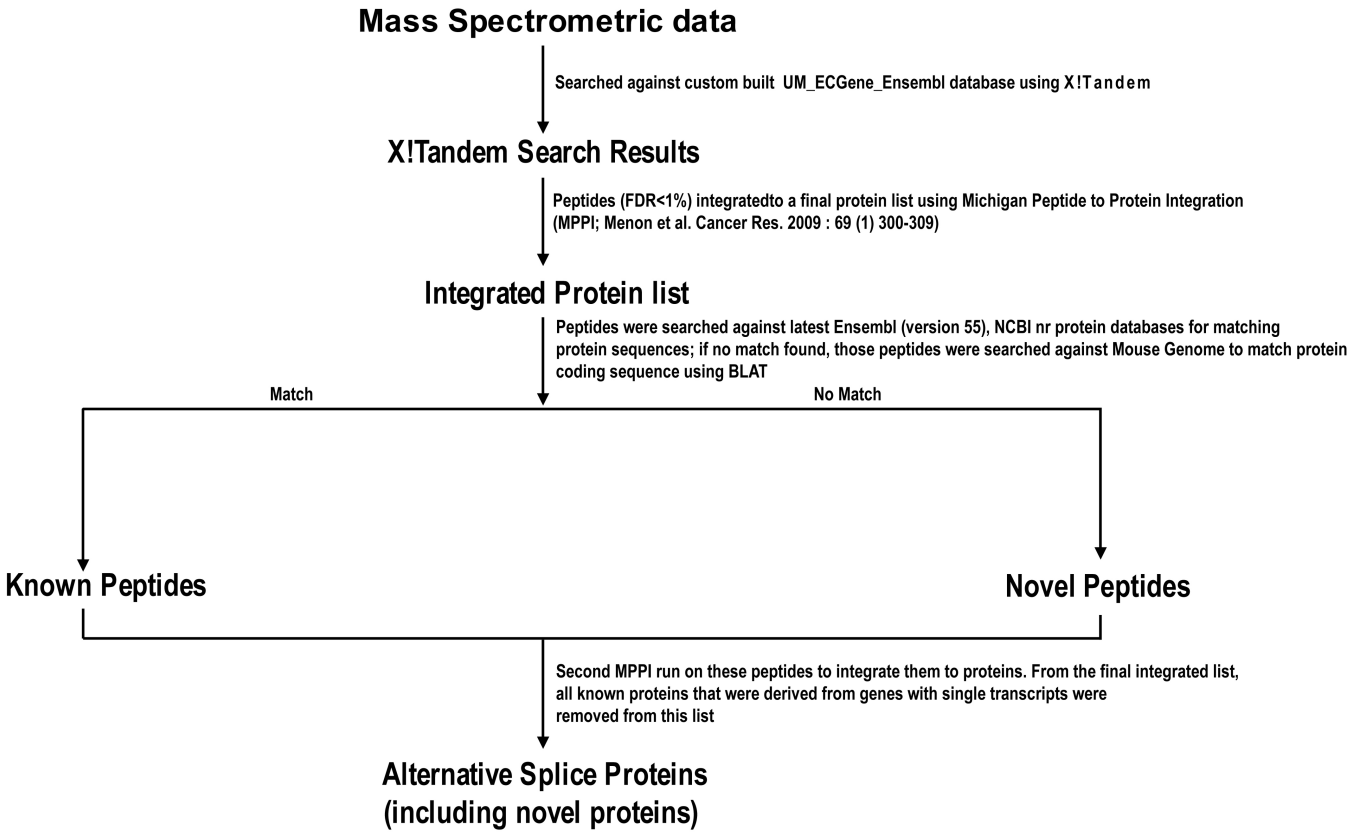

Figure.

The Flow chart displaying the analytical work flow of the X!Tandem search results for the identification of Alternative Splice Variant proteins.