Functional and structural annotation of human protein variants originated from alternative splicing in human

D'Antonio M[1], Martelli PL[2], Castrignanò T[1], Fariselli P[2], Casadio R[2], Zauli A[3], Pesole G[4,5]

Motivation

Alternative splicing has been suggested as a key mechanism for increasing the the functional landscape of the human genome. In order to detect structural and functional features of alternative protein variants originated from the same gene, a pipeline for annotating all the alternative splicing variants included in the ASPicDB database has been implemented, by integrating different state of the art tools for similarity search and for the prediction of structural and functional features of a protein starting from its residue sequence.

Methods

For each one of the 254,195 protein variants coming from 17,142 human genes a first layer of annotation consists in searching with BLAST for similar sequences annotated in the SwissProt data base (rel. 53.0) or endowed with a resolved three dimensional structure in the PDB data base (rel. Apr 09). Moreover, remote homology searches are performed by mapping on the sequence the structural and functional domains collected in the PFAM database (rel. 23.0). To this aim we adopted the pfam_scan.pl program, downloaded from the PFAM ftp site (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/). The second layer of annotation results form a decision tree integrating several predictive tools developed by the Bologna Biocomputing Group. As a first step, N-terminal signal peptides and C-terminal GPI-anchor propeptides are predicted with Spep (Fariselli et al., 2003) and PredGPI (Pierleoni et al., 2008), respectively. When present, they are cleaved from the protein sequence and the presence and localization of alpha-helical transmembrane domains is predicted with ENSEMBLE (Martelli et al., 2003). Secondary structure and cysteine bonding state are predicted with SecPred (Jacoboni et al., 2000) and CysPres (Martelli et al., 2004), respectively. The subcellular localization of globular proteins is predicted with BaCelLo (Pierleoni et al., 2006).

[1] Consorzio per le Applicazioni di Supercalcolo per Università e Ricerca, Rome, Italy [2] Biocomputing Group, University of Bologna, Bologna, Italy [3] BioDec srl, Bologna [4] Istituto Biomembrane e Bioenergetica, Consigli Nazionale delle Ricerche, Bari, Italy [5] Dipartimento di Biochimica e Biologia Molecolare, University of Bari, Bari, Italy

Results

As a result of the first annotation layer, 228,737 protein variants share similarity with a SwissProt sequence with an E-value lower than 10-5. Out of these proteins, 129,828 share also similarity with a sequence included in the PDB database. In these cases, the transfer of the functional and structural annotations by similarity is feasible, at least for the aligned regions. 350,870 PFAM domains are also mapped onto 159,538 protein variants with an E-value lower than 10-5. The second layer of annotation discriminates 28,991 and 1,679 sequences endowed with signal peptides and GPI-anchor propeptides, respectively. 41,594 variants are predicted as transmembrane and, among globular proteins, BaCelLo classifies 69,251 sequences as nuclear, 90,267 as cytoplasmic, 19,514 as mitochondrial and 31,996 as secreted. The structural and functional annotations of the proteins encoded by the transcript variants were added in the ASPicDB database (Castrignanò et al. 2008) and can be browsed by means of a graphical search interface, that also allows to retrieve all the genes whose splicing variants encode proteins with specific structural or functional properties (e.g. PFAM or TM domain, protein type, etc.) (Fig. 1A) or showing differences in specific features (Fig. 1B) . Availability: The ASPicDB, supplemented with the annotations for the protein coding transcripts, is available at http://www.caspur.it /ASPicDB/

Availability

http://www.caspur.it/ASPicDB/

Contact e-mail

graziano.pesole@biologia.uniba.it

Image

Figure 1. Protein Search form in ASPicDB