BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 12 2011, pages 1625–1629
doi:10.1093/bioinformatics/btr198

Structural bioinformatics

Advance Access publication April 15, 2011

MAISTAS: a tool for automatic structural evaluation of alternative splicing products

Matteo Floris[1,†], Domenico Raimondo[2,†], Guido Leoni[2], Massimiliano Orsini[1], Paolo Marcatili[2] and Anna Tramontano[3,4,*]

[1]CRS4-Bioinformatics Laboratory, c/o Sardegna Ricerche Scientific Park, Pula, 09010 Cagliari, [2]Department of Biochemical Sciences, [3]Department of Physics and [4]Istituto Pasteur Fondazione Cenci Bolognetti, Sapienza University of Rome, P le Aldo Moro 5, 00185 Rome, Italy

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Analysis of the human genome revealed that the amount of transcribed sequence is an order of magnitude greater than the number of predicted and well-characterized genes. A sizeable fraction of these transcripts is related to alternatively spliced forms of known protein coding genes. Inspection of the alternatively spliced transcripts identified in the pilot phase of the ENCODE project has clearly shown that often their structure might substantially differ from that of other isoforms of the same gene, and therefore that they might perform unrelated functions, or that they might even not correspond to a functional protein. Identifying these cases is obviously relevant for the functional assignment of gene products and for the interpretation of the effect of variations in the corresponding proteins.

Results: Here we describe a publicly available tool that, given a gene or a protein, retrieves and analyses all its annotated isoforms, provides users with three-dimensional models of the isoform(s) of his/her interest whenever possible and automatically assesses whether homology derived structural models correspond to plausible structures. This information is clearly relevant. When the homology model of some isoforms of a gene does not seem structurally plausible, the implications are that either they assume a structure unrelated to that of the other isoforms of the same gene with presumably significant functional differences, or do not correspond to functional products. We provide indications that the second hypothesis is likely to be true for a substantial fraction of the cases.

Availability: http://maistas.bioinformatica.crs4.it/.

Contact: anna.tramontano@uniromal.it

Received on October 26, 2010; revised on March 17, 2011; accepted on March 22, 2011

1 INTRODUCTION

Determining the identity and function of all the sequence elements in human DNA is a daunting challenge. The large scale pilot phase of the ENCODE project (Birney *et al.*, 2007) provided an exhaustive identification and verification of functional sequence elements in a limited region of 1% of the human genome. The computational analysis of the data revealed several unexpected features of the genome (Tress *et al.*, 2007). Perhaps the most surprising one was that many transcribed elements could be neutral elements that serve as a reservoir for natural selection. Many of these transcripts derive from alternative splicing events. Their putative products were manually analysed by the BioSapiens European Consortium (Tress *et al.*, 2007). The analysis led to the striking conclusion that more than 50% of them might not give rise to proteins structurally and/or functionally related to the other isoforms of the same genes or be the result of aberrant splicing events giving rise to non-functional proteins (Tress *et al.*, 2007).

Indeed, comparison of the putative proteins encoded by the alternatively spliced transcripts with the main isoform showed that most of them lacked an active site, key trans-membrane segments, essential signalling regions and post-transcriptionally modified sites. Most importantly, models of their putative three-dimensional structures did not seem to correspond to plausible folds (Tress *et al.*, 2007).

This observation was confirmed by Moult and co-workers (Melamud and Moult, 2009a, b) who, using a completely different dataset of alternative splicing variants, found that the vast majority of them resulted in putatively unstable protein conformations.

Recently, some of us manually analysed the putative structures of isoforms of the human genome, the existence of which had been confirmed by mass-spectrometry and of isoforms of the same genes for which no evidence exists in proteomic databases reaching essentially the same conclusions (Leoni *et al.*, 2011).

Altogether these observations suggest that we might be observing the effects of noisy selection of splice sites by the splicing machinery and/or that alternatively spliced products of a gene might assume unrelated conformations.

These findings raise several interesting questions, but also a few practical issues. First of all, the careful manual analysis performed by the BioSapiens consortium on 1% of the genome needs to be scaled up to the whole genome and therefore automated. Secondly, analysis tools should be available to biologists performing experiments in a user-friendly manner.

At present, there are a few systems that partially satisfy this need. For example, the ProSas database (Birzele *et al.*, 2008) (http://www.bio.ifi.lmu.de//forschung/structural-bioinformatics/prosas) stores structures and models (provided the target proteins shares at least 40% sequence identity with a known template) for the alternative isoforms annotated in Ensembl (Hubbard *et al.*, 2002)

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

and Swiss-Prot (Bairoch *et al.*, 2004) and allows the visualization of the exon boundaries in the context of the three-dimensional structures, but there is no provision for automatic analysis of the plausibility or completeness of the resulting structures and models. The same is true for AS-ALPS (Shionyu *et al.*, 2009) (http://as-alps.nagahama-i-bio.ac.jp/), a server that provides information about the putative effect of alternative splicing on human and mouse proteins, provided that at least one of the isoforms has an experimentally solved structure.

Here, we describe a system named Modelling and Assessment of ISoforms Through Automated Server (MAISTAS) that, given the accession codes of one or more genes or proteins, collects all their putative spliced isoforms annotated in the Ensembl genome database (Hubbard *et al.*, 2002), builds, whenever possible, comparative models for their structures, analyses their features and provides an estimate of the likelihood that the isoforms correspond to potentially stable and structurally plausible proteins in the absence of major conformational rearrangements.

Alternative splicing isoforms can also be uploaded in the FASTA format in order to allow the user to analyse data from more comprehensive and specialized databases such as Aceview (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/) (Thierry-Mieg and Thierry-Mieg, 2006) or ASPicDB (http://t.caspur.it/ASPicDB/) (Martelli *et al.*, 2010).

Model assessment is performed by analysing the quality of the packing in the core of the structure and/or model, the extent of exposed hydrophobic surface and the putative effect of deletions and insertions. These properties are compared to those observed in known protein structures and in the closest homologs of the known structure. The system is freely available as a Web server.

2 METHODS

The input data can be a set of sequences in the FASTA format or one or more of the following codes: Ensembl Gene ID(s), Ensembl Transcript ID(s), Ensembl protein ID(s) (Flicek *et al.*), EMBL ID(s) (Leinonen *et al.*, 2011), EntrezGene ID(s) (Maglott *et al.*, 2011), GO ID(s) (Ashburner *et al.*, 2000), HGNC automatic gene name, HGNC curated gene name (Seal *et al.*, 2011), UniProt/TrEMBL Accession(s), UniProt/Swissprot ID(s), UniProt/Swissprot Accession(s) (The Uniprot Consortium, 2008), VEGA transcript ID(s), HAVANA transcript ID(s) (Wilming *et al.*, 2008).

The collection of all putative splicing isoforms corresponding to the input gene (or to the gene encoding for the protein when a protein accession code is used) is achieved by taking advantage of a locally stored version of the Ensembl database (release 58) (Flicek *et al.*, 2011). Users can select accession codes for more than 30 different organisms.

The HHsearch 1.1.5 (Söding, 2005) is used to search for possible structural templates (*E*-value lower than 10^{-5}, sequence coverage of at least 90%, global alignment mode, all other parameters set at their default values) and for obtaining the sequence alignment between the target and its templates. Model building is performed using a local version of Modeller9v8 (Sali and Blundell, 1993) (default parameters).

The selected parameters ensure that the quality of the produced models is sufficiently high to be able to reliably measure properties described below as demonstrated by the last CASP experiment (http://predictioncenter.org/CASP9).

POPS (Cavallo *et al.*, 2003) is used to calculate the accessibility to the solvent of each residue of the models. The OS software (Pattabiraman *et al.*, 1995; Fleming and Richards, 2000) is used for computing infrequent environment of residues. Finally, the 'packing-eff' method from the NUCPROT package (Voss and Gerstein, 2005) is used for estimating how well packed the protein is.

The thresholds for POPS, Packing-eff and OS tools were derived by running the programs on 7908 monomeric proteins solved by X-ray crystallography at a resolution better than 2.5 Å. The chosen thresholds, 20.1 for POPS values, 17.8% for Packing-eff values and 0.54 for OS values, correspond to two standard deviations from the average (data not shown).

Residues are considered exposed if their mean solvent accessibility—calculated considering three residues on each side of them—is larger than 5 $Å^2$.

The average response time for a typical request (three to four isoforms, a few hundreds amino acid long) is <1 h, the time limiting factor being the construction of the HMMs and of the corresponding models. The entire pipeline was built using python scripts and the interface is PHP based.

In order to verify that the system can be applied to a substantial fraction of cases and that is able to recognize translated proteins, we ran it on protein isoforms whose existence is unambiguously identified by mass spectrometry. We used the May 2010 human build (http://www.peptideatlas.org/builds/human/201005/APD_Hs_all.fasta) containing 72 396 different peptides ranging in size from 6 to 66 (mean 17) (Deutsch *et al.*, 2008). Of these, 19 513 could be unambiguously mapped to 2972 isoform products annotated in Ensembl (release 58). We also compared the results of MAISTAS with those obtained by a manual analysis of human transcript products described in Leoni *et al.* (2011).

3 RESULTS

The automatic analysis performed by MAISTAS requires that the user inputs one or more protein/gene accession codes from the common public databases (see Section 2) or a set of sequences in the FASTA format. In all but the last case, the sequence(s) corresponding to the user query is retrieved and mapped back to the appropriate genome database by using a local installation of the BioMart database (Durinck *et al.*, 2005). The peptide sequences of all isoforms of the target gene, as annotated in Ensembl, are retrieved.

If the input is a set of amino acid sequences in the FASTA format, they are assumed to be different isoforms of the same gene.

The user can supply an email address (optional) to which the results will be sent or bookmark the result page. The initial query page of MAISTAS provides a link to an example result page, which allows the user to inspect a typical output (Fig. 1).

In the first step, the tool evaluates whether a structure exists for any of the isoforms or, lacking this, whether a comparative model can be built. In the latter case, the template is identified using the HHsearch program, which builds a Hidden Markov Model (HMM) of the target protein family and compares it to the HMMs representing a set of non-redundant families of proteins of known structure (sequence identity between any pair below 70%). This strategy has been shown in blind tests to be one of the most sensitive for finding structural templates (Battey *et al.*, 2007).

The target sequence, the template(s) and the alignment obtained by the HHsearch are automatically analysed. Only models based on template structures solved by X-ray crystallography or an NMR are considered. They are inspected to detect any possible gaps in the coordinate set (for example, because of the absence of electron density in X-ray structures). If these regions are present at the N- or C-terminus of the protein they are trimmed, otherwise a warning is issued. A warning is also issued if the alignment includes insertions larger than 50 residues that might correspond to an inserted domain or deletions larger than 20 residues.

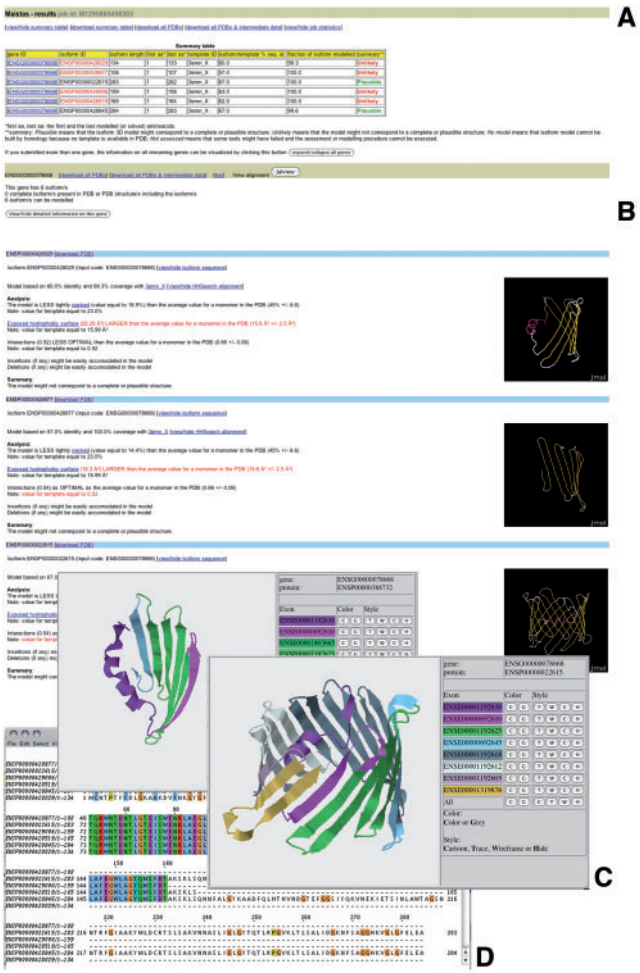

Fig. 1. Snapshots of the MAISTAS output page. (**A**) Summary table for the modelled isoforms. The following data are shown: gene ID (gene identification code), isoform ID (isoform identification code), isoform length (number of residues of each isoform), first aa, last aa (the first and last modelled or solved amino acid), template ID (the PDB code of the template protein used for modelling or the PDB code of the known isoform structure), isoform/template % seq. ID (sequence identity between the splicing isoform and the sequence of the selected template), fraction of isoform modelled (percentage of the splicing isoform sequence modelled), summary (assessment of the plausibility of the structure). (**B**) Snapshot of the isoform section showing results of the analysis for each isoform, its final assessment and the modelled structure in a small Jmol window. Different links in the section allow the user to download the coordinates of the model, view their 3D structure with regions corresponding to exons in different colours, view the amino acid sequence and the isoform/template alignment generated by the HHsearch. (**C**) Alternative spliced isoform three-dimensional structures are displayed in separate windows allowing their simultaneous analysis and comparison. On the right side of each Jmol window, the user can choose which exons should be displayed and select different representation modes. By default, all exons are mapped on the protein structure, each in a different colour. (**D**) Multiple sequence alignment of the isoforms displayed via the JALVIEW applet.

The alignment is used to build the model using a local installation of Modeller (Sali and Blundell, 1993). Once the model has been built, the system computes the model hydrophobic solvent accessible area and packing efficiency.

If the modelled isoform presents deletions with respect to the template, the Euclidean distance between the Cα residues before and after the deletion(s) is recorded. If insertions are present, the surface exposed to the solvent of the amino acids surrounding them and the number of inserted amino acids is computed.

The tool informs the user that the model might not correspond to a complete or plausible structure if the distance between the two residues on either side of a deletion is >15 Å and/or if there are more than three residues inserted in the core of the protein and/or if the hydrophobic solvent accessible area of the model is larger than a set threshold (see Section 2). In assessing the results, the system takes into account the corresponding values for the template used for modelling.

The output of MAISTAS is shown in Figure 1 and includes a summary table, where all the data regarding the modelled isoforms are reported. These can also be downloaded as a csv file. The user can download the coordinates of all the models and, if desired, all the intermediate data used in the procedure. The next section of the output page describes the detailed results for each modelled isoform and reports (see Section 2 for details):

- The sequence identity and coverage of the template and its PDB code.
- The packing efficiency of the model and of its template together with their comparison with the expected value.
- The extent of the exposed hydrophobic area of the model and of its template together with their comparison with the expected value.
- The packing environment of residues in the model and the template together with their comparison with the expected value.
- The assessment of whether insertions and deletions (if any) can be easily accommodated into the model.
- The modelled or experimental structure in a Jmol window.
- The option to inspect the multiple sequence alignment via a JALVIEW applet (Waterhouse *et al.*, 2009).
- The option to visualize and analyse the models via a Jmol applet (http://www.jmol.org/).
- A final remark about the plausibility/completeness of the predicted structure.

MAISTAS depends on the availability of structural templates to predict the three-dimensional structure of the isoforms by comparative modelling. If no structural templates are available, a 'No template satisfying all parameters' warning is issued. When MAISTAS is unable to provide a reasonable structural model (e.g. when very large insertions are present) the system will return the message 'Maistas is having trouble modelling or assessing this isoform'.

The online result pages are accessible via the URL sent either by e-mail or via the 'Retrieve results by job identifier or by email' window, using the provided job identification code or the e-mail address.

Produced models and the results of their analysis are stored in a local database unless the user requests them to be kept private. This implies that a user might be able to immediately retrieve the results on the gene(s) of interest if they were already been produced in a previous run of the system. The entries of the database are time stamped and presented to the user together with an option to repeat the analysis, which is advisable if major updates of the genome or structure database have taken place since the previous analysis was performed.

We ran the system on all human alternatively spliced isoform whose existence at the protein level could be unambiguously verified by mass spectrometry, i.e. of those protein isoforms for which a peptide that unambiguously identifies them has been detected with high reliability by mass spectrometry.

The server was able to produce and analyse models in 30% of the cases (890 out of 2972). In 2082 of them (70%), the model could not be built because there is no template satisfying all parameters. This had to be expected since we use rather stringent parameters to select the template (E-value better than 10^{-5}, template coverage >90%, X-ray resolution <2.5 Å or solved by the NMR). Out of the modelled isoforms, 712 (80%) were assessed as structurally plausible (see http://www.bioinformatica.crs4.org/maistas/pub/dataset.xls). In the majority of the remaining cases, (160 out of 178) the model showed a large hydrophobic surface exposed to the solvent. In these cases, the protein might indeed represent an incomplete and therefore not plausible structure, but also simply be a subunit of a larger complex.

We compared the results obtained by MAISTAS with those derived from a manual analysis of the isoforms of genes for which at least one isoform had been detected in mass-spectrometry experiments [and unambiguously identified by the presence of a peptide in the PeptideAtlas database (Deutsch *et al.*, 2008) and at least one had not (Leoni *et al.*, 2011)]. The results obtained automatically using MAISTAS are consistent with those reported in Leoni *et al.* (2011). In particular, MAISTAS was able to model 30% of the 555 proteins for which there is an evidence of translation (to be compared with the 26.4% obtained in the manual analysis), 85% of which were assessed as structurally plausible. The difference in coverage between the manual and automatic analyses is due to the increased size of the protein sequence and structure databases. Models were also produced for 181 out of 555 isoforms for which there is no evidence of translation in PeptideAtlas. Only 44% of these isoforms were reported as complete and plausible by the automatic pipeline. The corresponding numbers for manual analysis are 145 isoforms (26%) modelled and 48% classified as structurally consistent.

3.1 Application example

As an example of the use of MAISTAS, we describe the results obtained using the gene coding as input for the voltage-dependent anion channel 3 (VDAC3) (Ensembl gene identification code: ENSG00000078668), a protein that forms a channel through the mitochondrial outer membrane allowing diffusion of small hydrophilic molecules. Six splice variants are present in the Ensembl database for the gene encoding the protein, identified by the following Ensembl peptide codes: ENSP00000428845, ENSP00000022615, ENSP00000428519, ENSP00000428977, ENSP00000429006 and ENSP00000428029.

The UniProt database entry of VDAC3 (Q9Y277) describes only two of these isoforms (ENSP00000388732 and ENSP00000022615). Although four peptides mapping to the putative products are present in the PeptideAtlas database (PeptideAtlas IDs: PAp00006999; PAp00007806; PAp00077146; and PAp00423732), they cannot be used to unambiguous identify specific isoforms of the gene since they fall in the exons present in all of them.

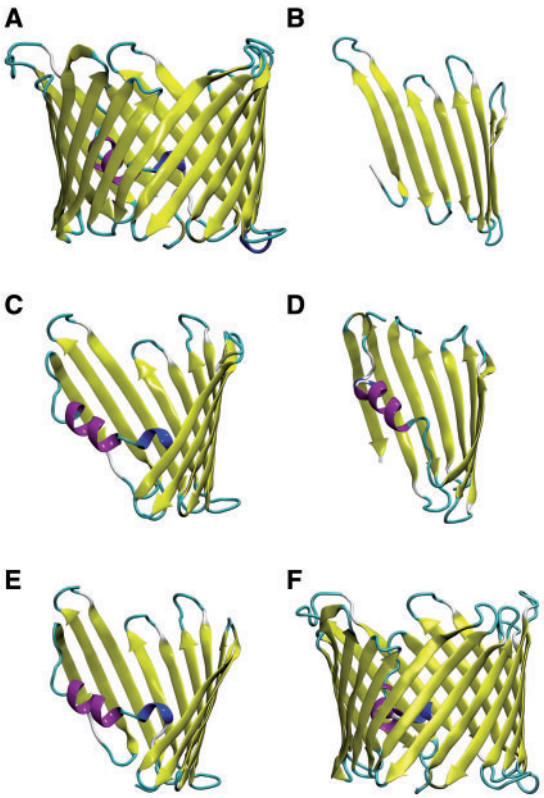

Fig. 2. Three-dimensional models of the VDAC3 protein isoforms. (**A**) ENSP00000428845. (**B**) ENSP00000428977. (**C**) ENSP00000428519. (**D**) ENSP00000428029. (**E**) ENSP00000429006. (**F**) ENSP00000422615.

Decker *et al.* (Decker and Craigen, 2000) used specific anti-VDAC3 antibody and demonstrated the existence of the ENSP00000428845 and ENSP00000022615 isoforms. The only difference between these two alternatively spliced isoforms is the insertion of a single methionine at position 39 of the ENSP00000428845 sequence.

ENSP00000022615 is also annotated in the CCDS database, a resource that centralizes the identification of well-supported, consistently annotated, protein-coding regions (Pruitt *et al.*, 2009). MAISTAS was able to provide a plausible structural model for isoforms ENSP00000428845 and ENSP00000022615 (Fig. 2A and F), while models of ENSP00000428519, ENSP00000428977, ENSP00000429006 and ENSP00000428029 were considered unlikely or incomplete (Fig. 2B–E). Inspection of the HHpred alignment used for building the ENSP00000428519, ENSP00000428977, ENSP00000429006 and ENSP00000428029 isoform models does not highlight any specific problem with the alignment (data not shown); however, the VDAC3 beta-barrel domain architecture is completely disrupted in the models of ENSP00000428519, ENSP00000428977, ENSP00000429006 and

ENSP00000428029 (Fig. 2B–E). All these isoforms show a large exposed hydrophobic surface, (around 22 Å^2, compared with the expected value of 15.6 Å^2 and with the value observed for the template of 15.9 Å^2). This dramatic architecture variation might imply that the isoforms are non-functional or that they perform a completely different function.

4 CONCLUSION

The more detailed is the analysis of the genomes of higher eukaryotes, the more complex they are revealed to be. For example, it is becoming clear that alternative splicing events do not simply result in a modulation of the function of the gene products, for example, by removing or adding structurally compact domains, or by modifying the sequence of specific regions of the encoded protein, but that they can either have a profound effect on the structure and function of the products of the same gene or give raise to non-functional products (Melamud and Moult, 2009a, b; Tress *et al*., 2007).

The latter can nevertheless have a relevant biological function. For example, Poliseno *et al*. demonstrated that transcripts may also function by competing for microRNA binding, a biological activity independent of the translation of the protein they encode (Poliseno *et al*., 2010). It is impossible for any currently available method, including ours, to assess which is the case.

The method described here is able to correctly classify as plausible a large fraction of the experimentally characterized isoforms, and to highlight dubious cases. Our aim is to provide easy access to a computational tool able to draw the attention of the life science community to them. Consequently, we took special care to convey the results of the analysis, although based on rather sophisticated tools, in an easy and understandable fashion. MAISTAS provides access to all the intermediate data used to generate the results, but it describes them in a human readable form. We believe that MAISTAS represents a step in the direction of using the knowledge accumulated in structural bioinformatics as well as the maturity of the tools available for applications related to the interpretation of genomic data and that it can be effectively used as a first step in characterizing novel proteins as well as a support for selecting interesting and intriguing cases for structural and functional studies.

ACKNOWLEDGEMENTS

We thank Loredana Le Pera, Andrea Sbardellati, Alejandro Giorgetti and Francesca Camilli for valuable feedback. We also thank Gianmauro Cuccuru, Michele Muggiri and Carlo Podda of the CRS4 High Performance Computing Group for their technical advice. We thank all the groups that kindly provided us with databases and binaries or source codes of the software installed and interfaced in this pipeline.

Funding: King Abdullah University of Science and Technology (KAUST; Award No. KUK-I1-012-43); Fondazione Roma and the Italian Ministry of Health (contract no.onc_ord 25/07, FIRB ITALBIONET and PROTEOMICA).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Bairoch,A. *et al.* (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinformatics*, **5**, 39–55.

Battey,J.N. *et al.* (2007) Automated server predictions in CASP7. *Proteins*, **69** (Suppl. 8), 68–82.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Birzele,F. *et al.* (2008) ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res*., **36**, D63–D68.

Cavallo,L. *et al.* (2003) POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res*., **31**, 3364–3366.

Decker,W.K. and Craigen,W.J. (2000) The tissue-specific, alternatively spliced single ATG exon of the type 3 voltage-dependent anion channel gene does not create a truncated protein isoform *in vivo*. *Mol. Genet. Metab*., **70**, 69–74.

Deutsch,E.W. *et al.* (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep*., **9**, 429–434.

Durinck,S. *et al.* (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.

Fleming,P.J. and Richards,F.M. (2000) Protein packing: dependence on protein size, secondary structure and amino acid composition. *J. Mol. Biol*., **299**, 487–498.

Flicek,P. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res*., **39**, D800–D806.

Hubbard,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res*., **30**, 38–41.

Leinonen,R. *et al.* (2011) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.*, **38**, D39–D45.

Leoni,G. *et al.* (2011) Coding potential of the products of alternative splicing in human. *Genome Biol*., **12**, R9.

Maglott,D. *et al.* (2011) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res*., **39**, D52–D57.

Martelli,P.L. *et al.* (2010) ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. *Nucleic Acids Res*., **39**, D80–D85.

Melamud,E. and Moult,J. (2009a) Stochastic noise in splicing machinery. *Nucleic Acids Res.*, **37**, 4873–4886.

Melamud,E. and Moult,J. (2009b) Structural implication of splicing stochastics. *Nucleic Acids Res.*, **37**, 4862–4872.

Pattabiraman,N. *et al.* (1995) Occluded molecular surface: analysis of protein packing. *J. Mol. Recognit*., **8**, 334–344.

Poliseno,L. *et al.* (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, **465**, 1033–1038.

Pruitt,K.D. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*., **19**, 1316–1323.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*., **234**, 779–815.

Seal,R.L. *et al.* (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res*., **39**, D514–D519.

Shionyu,M. *et al.* (2009) AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic Acids Res*., **37**, D305–D309.

Soding,J. (2005) Protein homology detection by HMM–HMM comparison. *Bioinformatics*, **21**, 951–960.

The Uniprot Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res*., **36**, D190–D195.

Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol*., **7** (Suppl. 1), s12: 1–14.

Tress,M.L. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.

Voss,N.R. and Gerstein,M. (2005) Calculation of standard atomic volumes for RNA and comparison with proteins: RNA is packed more tightly. *J. Mol. Biol*., **346**, 477–492.

Waterhouse,A.M. *et al.* (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

Wilming,L.G. *et al.* (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res*., **36**, D753–D760.