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

Floris, Matteo; Raimondo, Domenico; Leoni, Guido; Orsini, Massimiliano; Marcatili, Paolo; Tramontano, Anna

Published in:
Bioinformatics (Online)

Link to article, DOI:
10.1093/bioinformatics/btr198

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
MAISTAS: a tool for automatic structural evaluation of alternative splicing products

Matteo Floris1,†, Domenico Raimondo2,†, Guido Leoni2, Massimiliano Orsini1, Paolo Marcatili2 and Anna Tramontano3,4,∗

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

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, and therefore that most of them lacked an active site, key trans-membrane segments, essential signalling regions and post-translationally 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 Moul and co-workers (Melamud and Moulh, 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 (Leonii 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.iit.it/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).
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/Accembly/) (Thierry-Mieg and Thierry-Mieg, 2006) or ASPicDB (http://aspicdb.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., 2009), EMBL ID(s) (Leinonen et al., 2011), EntrezGene ID(s) (Maglott et al., 2011), GO ID(s) (Ashburner et al., 2000), HGNIC automatic gene name, HGNIC curated gene name (Seal et al., 2011), UniProt/IEBM 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^{-3}, 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 Modeller9v9 (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 Å².

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/APPD_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.
MAISTAS: modelling of alternative splicing isoforms

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.
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: ENSG0000078668), 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, ENSP00000428029, 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 unambiguously identify specific isoforms of the gene since they fall in the exons present in all of them.

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.

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 ENSP00000428845, ENSP00000428977, ENSP00000429006 and
We thank Loredana Le Pera, Andrea Sbardellati, Alejandro Giorgetti and Francesca Camilli for valuable feedback. We also thank Gianmauro Cuccuru, Michele Muggigi and Carlo Poddà 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-1-012-43); Fondazione Roma and the Italian Ministry of Health (contract no.snc_ord 25/07, FIRB ITALBIONET and PROTEOMICA).

**Conflict of Interest:** none declared.

**REFERENCES**


**MAISTAS: modelling of alternative splicing isoforms**