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):

---

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
MAISTAS: a tool for automatic structural evaluation of alternative splicing products

Matteo Floris1, Domenico Raimondo2,†, Guido Leonì2, 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 different 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.

Availability: http://maistas.bioinformatica.crs4.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)
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., 2011), 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/TEMBL Accession(s), UniProt/Swissprot ID(s), UniProt/Swissprot Accession(s) (The UniProt Consortium, 2008), VEGA transcript ID(s), HAVANA transcript ID(s) (Wilmung et al., 2009). 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 genome database (release 58). Users can select accession codes for more than 30 different organisms.

The HHsearch 1.1.5 (Soding, 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 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 (Parabrahman 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.

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

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 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: ENSG00000786680), 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, ENSP00000428215, 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 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 ENSP00000022615 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 ENSP00000428029.
We thank Loredana Le Pera, Andrea Sbardellati, Alejandro Giorgetti
et al.

The more detailed is the analysis of the genomes of higher
organisms, the more exposed hydrophobic surface, (around 22 Å2, compared with
the expected value of 15.6 Å2 and with the value observed for the
template of 15.9 Å2). 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 rise 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: modelling of alternative splicing isoforms

REFERENCES


66–82.

Bimo,Y. et al. (2007) Identification and analysis of functional elements in 1% of the

Breitkreutz,B. et al. (2008) ProSAS: a database for analyzing alternative splicing in the

Cavalli,L. et al. (2003) POPs: a fast algorithm for solvent accessible surface areas at

ATG exon of the type 3 voltage-dependent anion channel gene does not create a


Dunnick,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,P.M. (2000) Protein packing: dependence on protein size,


30, 38–41.

Lennon,R. et al. (2011) Improvements to services at the European Nucleotide Archive.

Genome Biol., 12, R9.

Mogk,O. et al. (2011) Estrogen gene: gene-centered information at NCBI. Nucleic Acids


Prit,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.

Sabah-A and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of

39, D514–D519.


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


supported gene and transcripts annotation. Genome Biol., 7 (Suppl. 1), s12:
1–14.


Vos,N.B. and Greseth,M. (2005) Calculation of standard atomic volumes for RNA and


ACKNOWLEDGEMENTS

We thank Loredana Le Pera, Andrea Shardellati, Alejandro Giorgetti
and Francesca Camilli for valuable feedback. We also thank
Gianmauro Cuccuru, Michele Muggiari and Carlo Podda of the CSRS
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.snc_ord 25/07, FIRB ITALBIOMET and PROTEOMICA).

Conflict of Interest: none declared.