Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W169–W172***
doi:10.1093/nar/gkl208

PONGO: a web server for multiple predictions of all-alpha transmembrane proteins

Mauro Amico, Michele Finelli, Ivan Rossi, Andrea Zauli, Arne Elofsson[1], Håkan Viklund[1], Gunnar von Heijne[1], David Jones[2], Anders Krogh[3], Piero Fariselli[4], Pier Luigi Martelli[4] and Rita Casadio[4,*]

BioDec Srl, via Calzavecchio 20/2, I-40033 Casalecchio di Reno (BO), Italy, [1]Stockholm Bioinformatics Center, Stockholm University, SE-10691 Stockholm, Sweden, [2]University College of London, Gower Street, WC1E 6BT London, UK, [3]Bioinformatics Center, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark and [4]Biocomputing Group, University of Bologna, via Selmi 3, I-40126, Bologna, Italy

Received February 10, 2006; Revised March 6, 2006; Accepted March 22, 2006

ABSTRACT

The annotation efforts of the BIOSAPIENS European Network of Excellence have generated several distributed annotation systems (DAS) with the aim of integrating Bioinformatics resources and annotating metazoan genomes (http://www.biosapiens.info). In this context, the PONGO DAS server (http://pongo.biocomp.unibo.it) provides the annotation on predictive basis for the all-alpha membrane proteins in the human genome, not only through DAS queries, but also directly using a simple web interface. In order to produce a more comprehensive analysis of the sequence at hand, this annotation is carried out with four selected and high scoring predictors: TMHMM2.0, MEMSAT, PRODIV and ENSEMBLE1.0. The stored and pre-computed predictions for the human proteins can be searched and displayed in a graphical view. However the web service allows the prediction of the topology of any kind of putative membrane proteins, regardless of the organism and more importantly with the same sequence profile for a given sequence when required. Here we present a new web server that incorporates the state-of-the-art topology predictors in a single framework, so that putative users can interactively compare and evaluate four predictions simultaneously for a given sequence. Together with the predicted topology, the server also displays a signal peptide prediction determined with SPEP. The PONGO web server is available at http://pongo.biocomp.unibo.it/pongo.

INTRODUCTION

All-alpha membrane proteins constitute an important part of the cell proteome. Such proteins perform many basic functions including cell signalling, transcription regulation, energy conservation and transformation, and ion exchange. Membrane proteins are difficult to study, since they are inserted into lipid bilayers and expose to the polar outer and inner environments portions of different sizes. It is therefore difficult to purify them in the native, functional form and even more difficult to crystallize them. For such technical reasons, only a small fraction of the Protein Data Bank structures are membrane proteins (<1% of the total number of structures and far less than their estimated abundance in cells) (1).

There are a number of computational methods available to predict the topology of membrane proteins, which consists of two basic features: (i) the location of transmembrane domains along the protein chain and (ii) the location of the N- and C-termini with respect to the lipid membrane. Topological models are sufficient in many instances to design simple experiments in order to prove the location of the N- and C-protein termini, that of the inner and outer loops with respect to the membrane plane, and concomitantly the number of transmembrane segments in the chain. However, the best-performing methods are offered at different servers and are endowed with different graphical interfaces. This hampers the direct comparison of predictions, especially for experimentalists interested in comparing their results with computational methods. Currently, two other web servers of which we are aware (2,3) are available and separately comprise two of the predictors that we implement TMHMM2.0 (4) and MEMSAT (5). The novelty of our web server is to include in the same framework ENSEMBLE (6) and PRODIV (7), two powerful methods that became available only recently. Furthermore the available web servers render only the consensus prediction, without allowing a critical discrimination among different

*To whom correspondence should be addressed. Tel: +39 051 2091284; Fax: +39 051 242576; Email: casadio@alma.unibo.it

predictions that may be useful, considering that different predictors may highlight different properties, depending on the different implementation.

Automatic topology annotation for membrane proteins has been included among the efforts of the BIOSAPIENS European Network of Excellence (http://www.biosapiens.info) with the specific aim of taking into consideration different predictors for annotating membrane proteins in the human genome. The common platform for these efforts is the BIOSAPIENS Distributed Annotation Servers (DAS) (http://www.biosapiens.info).

In this context, the PONGO-DAS server (http://pongo.biocomp.unibo.it) provides topology annotation for the all-alpha membrane proteins of the human genome. This is done using the DAS protocol (www.biodas.org) to answer at DAS queries that can be seen using specific visualizers such as DASTY (8) or ENSEMBL (9). In order to allow users to browse directly the pre-computed transmembrane annotations, PONGO-DAS provides a simple graphical web interface. The annotation is carried out using four selected predictors, namely ENSEMBLE1.0 (6) and PRODIV (7), in order to allow a direct comparison of the topology prediction for the sequence at hand.

The server has also been set up to make it possible to predict the topology of any putative membrane proteins of interest regardless of provenance. The topological models computed by the different predictors can be directly obtained simply by pasting in a box the sequence of interest and looking at the results. Recently developed web technologies (e.g. AJAX) are used to improve the user interface.

MATERIALS AND METHODS

The predictors

The website implements the following predictors:

(i) *MEMSAT* is a new version of the MEMSAT predictor of transmembrane helices in proteins (4). This new version takes advantage of the evolutionary information derived by multiple sequence alignment. This method is based on a dynamic programming approach and statistical parametrization.
(ii) *TMHMM2.0* which is a predictor of transmembrane helices in proteins based on hidden Markov models (5). It has been shown that it performs quite well taking into account that it uses only single sequence information. For this reason it is also very fast.
(iii) *ENSEMBLE1.0* is an ensemble of two hidden Markov models and one neural network (6). ENSEMBLE takes also advantage of the evolutionary information derived by multiple sequence alignment, both for the neural network and the hidden Markov model systems.
(iv) *PRODIV_TMHMM_0.91* is a recent predictor of transmembrane helices in proteins (7) which uses a hidden Markov model similar to TMHMM, but exploits the evolutionary information derived by multiple sequence alignment.
(v) *SPEP* is a signal peptide predictor based on combination of neural networks (10). This predictor has performance similar to the most widely used SignalP (11), and it has been included since it is quite common that signal peptides are mispredicted as transmembrane helices.

Pre-computed transmembrane annotations for the human proteome

In the context of the BIOSAPIENS project, we downloaded the UNIPROT dataset (September 22, 2004) which consists of 33 135 human protein sequences, and for future use also the IPI dataset (August 5, 2004) which consists of 46 782 human proteins. The union of the two datasets comprises 50 600 sequences from the human genome. It is worth noticing that these 50 600 sequences do not include the known splice variants, since those variants are not presently included as unique UNIPROT codes.

The choice of annotating UNIPROT sequences is well justified from the fact that ENSEMBL gene products (in contrast to the UNIPROT entries) are not stable; for instance, only 60% of the sequences are common (and conserved) between the ENSEMBL releases 34 and 35, respectively. Finally, UNIPROT provides extensive functional annotation. We then use four state-of-the-art predictors already described in the literature in order to identify the most probable integral membrane proteins. In this way the union of the predictions obtained with the four methods gives a set of likely membrane proteins, while the intersection contains those chains on whose annotation as membrane proteins all the predictors agree upon.

The pre-computed annotations are filtered with SPEP before being processed by the transmembrane predictors. This is done since it is quite common that signal peptides are mispredicted as transmembrane helices by all the predictors implemented in our web server. In the case of a positive SPEP answer, the system cuts the corresponding predicted segment and processes the remaining part of the sequence with the four transmembrane predictors.

RESULTS

PONGO has two different usage options:

(i) PONGO-DAS accessing the pre-computed annotations using keywords or DAS queries;
(ii) PONGO-PRED that is an enhanced and modern version of a standard 'through-the-web' server application.

In the case of PONGO-DAS we implemented two types of result visualization:

(i) by means of a DAS, which is a client–server system in which a single client integrates information from multiple servers. It allows a single machine to gather up genome annotation information from multiple distant web sites, collate the information, and display it to the user in a single view. Little coordination is needed among the various information providers, and
(ii) a user-friendly interface that allows the search for a specific prediction.

Through the DAS protocol a web client like ENSEMBL can obtain the list of the annotations for each human protein using its UNIPROT code as requested for the DAS

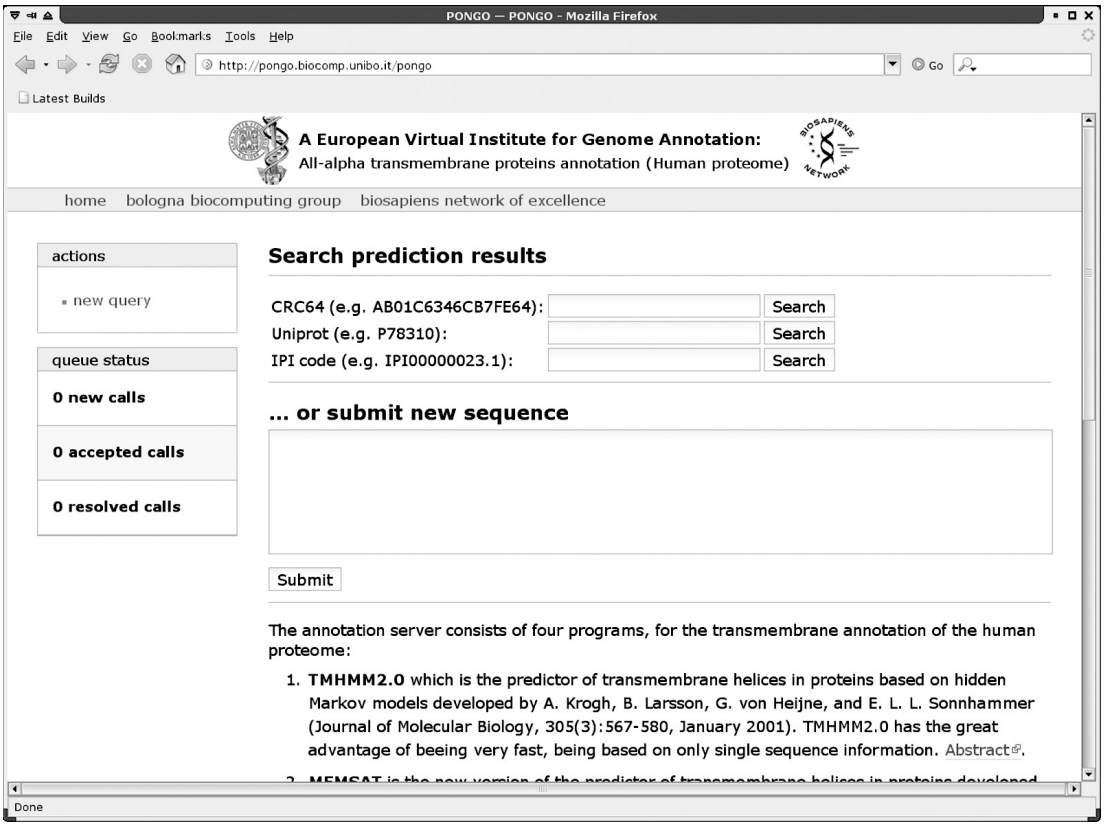

Figure 1. The PONGO homepage. On the left-side it is possible to follow the status of the different queries and the starting of a new action.

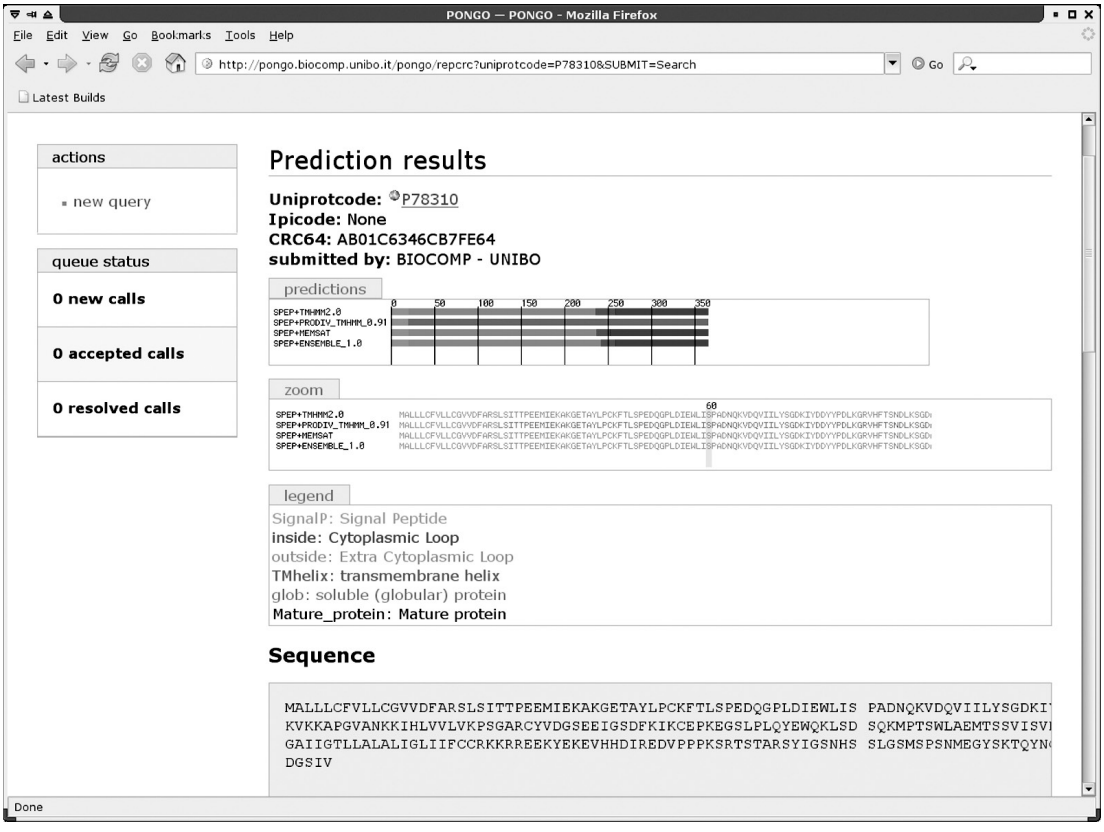

Figure 2. An example of PONGO results for a protein chain endowed with a signal peptide.

queries (http://pongo.biocomp.unibo.it/das/dasdb/features?segment=UNIPROTCODE); in order to check this URL the client needs a DAS client such as the one that will be embedded in ENSEMBL pages at EBI and a UNIPROT code for the sequence. An associated URL is sent to the client for visualization. In the second case a user can directly query the database, without using the DAS infrastructure, using the web interface available at http://pongo.biocomp.unibo.it/pongo (Figure 1) The user can then use a CRC64 (the sequence hash code), or the sequence UNIPROT or IPI code to get the predictions in a graphical view. An example of a sequence filtered with the different predictors is presented together with the detailed sequence annotation (Figure 2). A colour code is used to quickly identify the transmembrane segments.

Conversely, PONGO-PRED provides the user with a unique framework for membrane protein topology and signal peptide predictions. Another interesting feature of PONGO-PRED is the Javascript-enabled portlet that in real time refreshes the queue status (without reloading of the submission page and with very lightweight Xml HTTP Requests). In particular, on the left-side of the page the user can realise whether her/his submission is a new call, whether it is running (in this case the starting date and an absolute link is provided for bookmarking), or whether it has been processed.

ACKNOWLEDGEMENTS

This work was fully supported by the Biosapiens Network of Excellence project. The BioSapiens project is funded by the European Commission within its FP6 Programme, under the thematic area 'Life sciences, genomics and biotechnology for health', contract number LSHG-CT-2003-503265. P.F. acknowledges MIUR for a PNR-2003 grant. R.C. was supported by the following grants: PNR 2001–2003 (FIRB art.8) and PNR 2003 projects (FIRB art.8) on Bioinformatics for Genomics and Proteomics and LIBI-Laboratorio Internazionale di BioInformatica. Funding to pay the Open Access publication charges for this article was provided by the BioSapiens project.

Conflict of interest statement. None declared

REFERENCES

1. Casadio,R., Fariselli,P. and Martelli,P.L. (2003) *In silico* prediction of the structure of membrane proteins: is it feasible? *Brief. Bioinf*., **4**, 341–348.
2. Taylor,P.D., Attwood,T.K. and Flower,D.R. (2003) BPROMPT: a consensus server for membrane protein prediction. *Nucleic Acids Res*., **31**, 3698–3700.
3. Arai,M., Mitsuke,H., Ikeda,M., Xia,J.-X., Kikuchi,T., Satake,M. and Shimizu,T. (2004) ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Res*., **32**, W390–W393.
4. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **15**, 3038–3049.
5. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*., **305**, 567–580.
6. Martelli,P.L., Fariselli,P. and Casadio,R. (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics*, **19**, i205–i211.
7. Viklund,H. and Elofsson,A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci*, **13**, 1908–1917.
8. Jones,P., Vinod,N., Down,T., Hackmann,A., Kahari,A., Kretschmann,E., Quinn,A., Wieser,D., Hermjakob,H. and Apweiler,R. (2005) Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics*, **21**, 3198–3199.
9. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.
10. Fariselli,P., Finocchiaro,G. and Casadio,R. (2003) SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics*, **19**, 2498–2499.
11. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*., **16**, 783–795.