ABSTRACT

The UCL Bioinformatics Group web portal offers several high quality protein structure prediction and function annotation algorithms including PSIPRED, pGenTHREADER, pDomTHREADER, MEMSAT, MetSite, DISOPRED2, DomPred and FFPred for the prediction of secondary structure, protein fold, protein structural domain, transmembrane helix topology, metal binding sites, regions of protein disorder, protein domain boundaries and protein function, respectively. We also now offer a fully automated 3D modelling pipeline: BioSerf, which performed well in CASP8 and uses a fragment-assembly approach which placed it in the top five servers in the de novo modelling category. The servers are available via the group web site at http://bioinf.cs.ucl.ac.uk/.

INTRODUCTION

As the rate of deposition of new protein sequences outstrips the rate at which new protein structures are deposited in the public databases, there will be a continued need for accurate protein structure prediction. It is also arguable that the need for novel function annotation tools is even more pressing. The UCL Bioinformatics Group web portal aggregates a range of methods, developed and maintained at UCL, which predict key structural features of proteins from either their primary structure (sequence) or tertiary structure. We have recently revamped all of our web tools and have taken the opportunity to port all of our services to a new unified server architecture. The services are now presented with a user-friendly common look and feel and we have greatly increased our capacity to serve requests to the community.

ALGORITHMS

This section gives an overview of the algorithms and services available via the UCL Bioinformatics Group web server (http://bioinf.cs.ucl.ac.uk/web_servers/).

PSIPRED is a secondary structure prediction method (1) which uses a two-stage neural network to predict secondary structure using PSI-BLAST output (2). PSIPRED V3.0 currently offers the highest available three-state accuracy (Q3) for secondary structure prediction of 81.4% (+0.6%). This year, in addition to improvements to the underlying algorithm, we have also made some minor updates to the PSIPRED web output, where we now provide coloured alignment schematic diagrams alongside the original secondary structure 'cartoons'.

MEMSAT3 and MEMSAT-SVM

Transmembrane topology prediction is provided by our MEMSAT methods. Our server now supports two algorithms: MEMSAT3 (3) and MEMSAT-SVM (4). Predictions made with MEMSAT3 and MEMSAT-SVM begin with the generation of a PSI-BLAST profile (PSSM). Both algorithms employ dynamic programming to enumerate all possible transmembrane topologies, but differ in the method used to score the different topologies. MEMSAT3 uses a neural network for scoring, and MEMSAT-SVM makes use of a support vector machine (SVM) along with additional target features such as signal peptides and re-entrant helices. Benchmarked with full cross-validation on a data set composed solely of sequences with crystal structures available to validate their topologies, MEMSAT3 and MEMSAT-SVM achieved 76 and 89% accuracy, respectively. Additionally, MEMSAT-SVM was able to predict both signal peptides (93% accuracy) and re-entrant helices (44% accuracy). When run via the portal, both MEMSAT3 and MEMSAT-SVM are run simultaneously and the results
produced allow the user to compare both predictions. Results may be returned by email and the web presentation displays both the schematic and cartoon diagrams of the predictions (Figure 1).

DISOPRED2 allows the user to predict disordered regions of proteins (5). Disordered regions are known to play important roles in protein–protein interactions, cell signalling, enzyme kinetics, signalling and a range of other processes. Disordered regions are characterized as those regions which do not have static structure and are believed to move continually through different configurations when the protein is in its functional configuration. DISOPRED2 generates a PSSM matrix and then analyses this with an appropriately trained SVM. DISOPRED accuracy has previously been shown to be ~93.1%. Our updated results presentation on the web now features a simple schematic which clearly displays the location of the disordered residues.
BioSerf

BioSerf is our new fully automated de novo and homology modelling pipeline available via the World Wide Web or as an XML-driven web service (manuscript in preparation). BioSerf uses a selection of algorithms including PSI-BLAST, PSIPRED and pGenTHREADER to attempt to intelligently select appropriate template structures for homology modelling. Where a suitable template or templates can be found, BioSerf uses MODELLER (13) to build a model. Where no suitable template structure or insufficient coverage can be found, FRAGFOLD (14) is used to create an ab initio template (Figure 2). As this service uses MODELLER, a valid MODELLER licence key is required to submit sequences to the service. Users receive a PDB file as the primary result along with the intermediate results used to generate the model. These intermediate results and the model itself can be inspected via a web page.

FFPred

The FFPred server (16) offers a method for predicting protein function using a machine learning approach. FFPred attempts to assign reliable Gene Ontology (GO) classes (17) to queries using a series of SVM classifiers. Query sequences are annotated with a large set of possible features including such data as amino acid content, transmembrane regions and disordered regions. Every GO class is represented by five SVM classifiers and the sequence is scored for each class. Final GO class assignment is determined by majority rule wherever at least three out of the five classifiers agree. The method achieves a specificity of >90% at sensitivities >30%. Results are
then returned to the user who can also explore the results via a temporary dynamically created web page.

Use case scenario

In this section, we describe a typical use case scenario for our methods which uses a range of our new servers. Typically we envisage that researchers with a single or a small number of uncharacterized or partially characterized proteins will be interested in making one or more predictions of structural features using our methods. We would advise users wishing to perform high-throughput studies to download the software which we make available via our web site (http://bioinf.cs.ucl.ac.uk/software_downloads/).

Structural annotation

We selected a previously uncharacterized, putative membrane protein from UniProt, A8MYE5 (currently available via http://www.uniprot.org/uniprot/A8MYE5.fasta). This is a 552 residue protein derived from human gene *MERTK*. The annotation via the protein domain databases available at UniProt suggests this may be a putative protein tyrosine kinase. Initially we submitted the sequence to our FFpred server. FFpred confirms this tentative functional annotation, identifying a range of GO terms the most significant of which are: GO:0008152, GO:0004672, GO:0006468 for ‘metabolic process’, ‘protein kinase activity’ and ‘protein amino acid phosphorylation’, respectively. Structural annotation began by submitting the sequence for analysis with the MEMSAT-SVM service (Figure 3b). The MEMSAT-SVM prediction indicates the possible presence of two membrane spanning helices towards the C-terminus of the sequence, between positions 115 and 130, and 146 and 165. Presence of these helices could indicate that the putative description assigned by Uniprot is correct. However were this a receptor tyrosine kinase, we might expect there to be only a single helix such that a ligand binding domain might be presented to the extracellular space.
The next analysis performed was to submit the sequence to the DomPred server (Figure 3c). The output shows two strong peaks around residue 150 and 234, with the domain boundary prediction made at residue 234 as the first peak likely corresponds to the transmembrane region we had already predicted. As such, this annotation appears to be in reasonable agreement with the initial MEMSAT-SVM annotation.

Based on the MEMSAT-SVM and DomPred annotation, the sequence was then divided into three segments. The initial break point was at residue 234, as suggested by the DomPred analysis, gave two sub-sequences: an N-terminal sequence containing both the initial domain alongside the transmembrane region and a larger C-terminal region containing the second domain. The smaller N-terminal sequence was then divided again at residue 130 to cleave the putative initial domain from the predicted transmembrane region. Both the putative N-terminal domain and the larger C-terminal domain were then submitted to the BioSerf homology modelling server, (Figure 3d). This generated two good, compact homology models for each domain.

The smaller N-terminal domain appears to be a simple two-layer α–β sandwich but the larger domain appears to be more complex. The smaller domain has a large exposed beta sheet, this type of conformation is not common in extracellular domains and this lends some support to the MEMSAT-SVM annotation that predicts two transmembrane helices, which in turn places the N-terminal domain in the cytosolic region. The homology model of the larger C-terminal domain shows an initial region comprised of a well-resolved beta sheet with an attendant layer of alpha helices forming a two-layer α–β sandwich. The other portion of the model appears to form a further discrete region that contains a clear four helix bundle. These two discrete compact regions may indicate that the larger C-terminal domain is in fact comprised of two smaller domains. The homology model for the C-terminal domain has a similar topology to PDB entry 1apm. Searching the CATH database for 1apm reveals that this domain has a similar topology to PDB entry 1apm.

DISCUSSION

The UCL Bioinformatics Group server offers some of the best performing algorithms of their kind. They can be accessed via the web (http://bioinf.cs.ucl.ac.uk/web_servers/) or most of the programs themselves can be downloaded for use on a local machine (http://bioinf.cs.ucl.ac.uk/software_downloads/). These services will be useful to any bioinformatician or biologist looking to functionally and structurally characterize proteins.

FUNDING

Funding for open access charge: Biotechnology and Biological Sciences Research Council, UK.

Conflict of interest statement. None declared.

REFERENCES

1. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol., 292, 195–202.
2. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–402.
3. Jones, D.T. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics, 23, 538–544.
4. Nugent, T. and Jones, D.T. (2009) Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics, 10, 159.
5. Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F. and Jones, D.T. (2004) The DISOPRED server for the prediction of protein disorder. Bioinformatics, 20, 2138–2139.
6. Bryson, K., Cazzetta, D. and Jones, D.T. (2007) Computer-assisted protein domain boundary prediction using the DomPred server. Curr. Protein Pept. Sci., 8, 181–188.
7. Lobley, A., Sadowski, M.I. and Jones, D.T. (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics, 25, 1761–1767.
8. Laskowski, R. A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.
9. Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J. P., Chothia, C. and Murzin, A. G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
10. Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A. et al. (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.*, **35**, D291–D297.
11. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. and Barton, G. J. (2009) Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
12. Yeats, C., Lees, J., Reid, A., Kellam, P., Martin, N., Liu, X. and Orengo, C. (2008) Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res.*, **36**, D414–D418.
13. Eswar, N., Eramian, D., Webb, B., Shen, M. Y. and Sali, A. (2008) Protein structure modeling with MODELLER. *Methods Mol. Biol.*, **426**, 145–159.
14. Jones, D. T. and McGuffin, L. J. (2003) Assembling novel protein folds from super-secondary structural fragments. *Proteins*, **53**(Suppl. 6), 480–485.
15. Sodhi, J. S., Bryson, K., McGuffin, L. J., Ward, J. J., Wernisch, L. and Jones, D. T. (2004) Predicting metal-binding site residues in low-resolution structural models. *J. Mol. Biol.*, **342**, 307–320.
16. Lobley, A. E., Nugent, T., Orengo, C. A. and Jones, D. T. (2008) FFPred: an integrated feature-based function prediction server for vertebrate proteomes. *Nucleic Acids Res.*, **36**, W207–W302.
17. The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
18. Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydrophathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.