BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 12 2005, pages 2917–2920
doi:10.1093/bioinformatics/bti445

Structural bioinformatics

Meta-DP: domain prediction meta-server

Harpreet Kaur Saini[1,*] and Daniel Fischer[1,2]

[1]Center of Excellence in Bioinformatics and Department of Computer Science and Engineering, University at Buffalo, 901 Washington Street, Suite 300, Buffalo, NY 14203, USA and [2]Department of Bioinformatics and Computer Science, Ben Gurion University, Beer-Sheva 84015, Israel

Received on February 17, 2005; revised on April 6, 2005; accepted on April 7, 2005
Advance Access publication April 19, 2005

ABSTRACT

Summary: Meta-DP, a domain prediction meta-server provides a simple interface to predict domains in a given protein sequence using a number of domain prediction methods. The Meta-DP is a convenient resource because through accessing a single site, users automatically obtain the results of the various domain prediction methods along with a consensus prediction. The Meta-DP is currently coupled to 10 domain prediction servers and can be extended to include any number of methods. Meta-DP can thus become a centralized repository of available methods. Meta-DP was also used to evaluate the performance of 13 domain prediction methods in the context of CAFASP-DP.

Availability: The Meta-DP server is freely available at http://meta-dp.bioinformatics.buffalo.edu and the CAFASP-DP evaluation results are available at http://cafasp4.bioinformatics.buffalo.edu/dp/update.html

Contact: hkaur@bioinformatics.buffalo.edu

Supplementary information: Available at http://cafasp4.bioinformatics.buffalo.edu/dp/update.html

INTRODUCTION

A domain is defined as a polypeptide chain or a part of a polypeptide chain that can fold independently and still exhibit the biological activity even if excised from the chain (Brändén and Tooze, 1991). A protein may be comprised of a single domain or several domains. On average, the number of domains in a protein increases with the complexity of the organism. Thus, eukaryotic proteins usually contain more domains than archaeal or bacterial proteins do (Brändén and Tooze, 1991). Multidomain proteins very often consist of a mixture of globular domains interspersed with non-globular regions. Domains in proteins can participate in crucial functions, such as protein interactions, DNA binding, enzyme activity and other important cellular processes. Thus, knowing the domain composition of a protein is essential for a detailed understanding of its function. Identification of domains in proteins can also be useful in structure determination. Moreover, database searching, phylogeny and protein modeling often perform better on single domains, rather than on a complete multidomain protein.

Prediction of domains from protein sequence has become an intensely researched area. The most useful and straightforward way to predict domains is by sequence homology where the target sequence is scanned against databases of protein domains and families. Adda (Heger and Holm, 2003), Biozon (Nagarajan and Yona, 2004), Dopro (von Ohsen *et al*., 2004), Mateo (Lexa and Valle, 2003) and Ginzu (Chivian *et al*., 2003) are examples of homology-based domain prediction methods which make use of domain databases to identify domains.

There are a number of well-known domain databases which store annotated multiple sequence alignments (MSAs) [in the form of position specific scoring matrices (PSSMs) or hidden Markov models (HMMs)] of protein domains, which can be used to identify conserved domains in the target sequence. Pfam (Bateman *et al*., 2000) and SMART (Schultz *et al*., 2000) databases are the largest collections of the manually curated protein domains. Apart from Pfam and SMART, there are a number of other databases that identify domains based on evolutionary relationships, such as Prodom (Corpet *et al*., 2000), TIGRFAMS (Haft *et al*., 2001), Superfamily (Gough *et al*., 2001), PROSITE (Falquet *et al*., 2002) and others. The MSAs as well as the annotations of each of these databases vary. Therefore, it is reasonable to use more than one method and infer domain boundaries from the analysis of all results. InterPro (Apweiler *et al*., 2001) is such an integrated database which combines a number of databases into one powerful resource. Its sequence search package, InterProScan (Zdobnov and Apweiler, 2001) combines the search methods from each of the databases and provides an output containing domain composition with consensus domain boundaries from all matches within each entry. In the presence of homology, InterProScan is among the best choices for domain prediction. However, there are a number of limitations of InterProScan. The output of InterProScan is often ambiguous and consists of overlaps which make it difficult to infer the domain boundaries clearly. Therefore, in order to predict domains, a user needs to analyze the InterProScan output visually. Although InterProScan provides valuable information about domains for proteins with similar sequences, it fails in the absence of homologous domains. Examples of such ambiguous and incomplete InterProScan predictions can be found at http://cafasp4.bioinformatics.buffalo.edu/dp/status.html. Thus, in the absence of clear homology, additional methods need to be applied.

In the absence of homologous sequences, homology-based domain prediction methods produce no prediction. However, a target protein may be structurally similar to one of the proteins of known 3D structures, even if there is no significant sequence similarity. In such cases, domains can be predicted using fold recognition or threading techniques, where the target sequence is aligned into a given structure or fold (Fischer *et al*., 1996). Dompred (Marsden *et al*., 2002)

*To whom correspondence should be addressed.

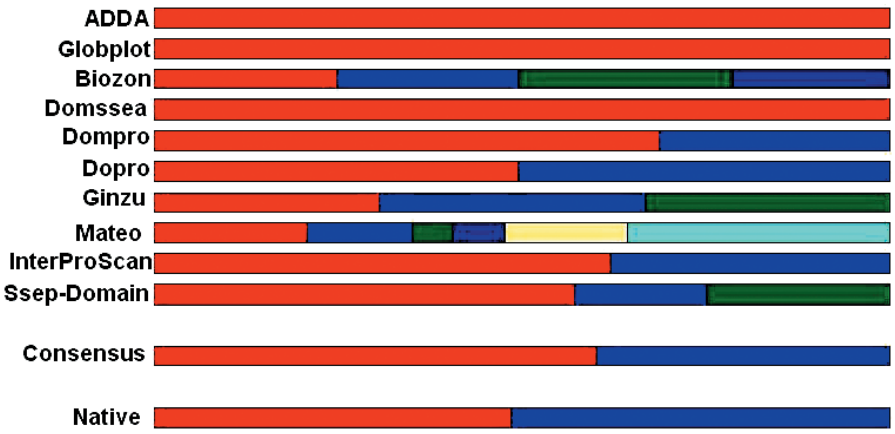

Fig. 1. Sample representation of Meta-DP output. Predicted domains are represented by colored horizontal bars. The consensus prediction computed from the individual predictions (see text) and the real domain boundaries from the experimental structure (native) are shown at the bottom. The real output of Meta-DP represents the domains numerically as 1, 2 and so on (see http://meta-dp.bioinformatics.buffalo.edu/help/sample.html for a sample output).

and Ssep-domain (http://www.bio.ifi.lmu.de/cafasp/) are examples of fold-recognition based domain prediction methods.

There are also other, non-homology based methods, such as Dompro (http://www.ics.uci.edu/~baldig/dompro.html) and Globplot (Linding *et al*., 2003). Dompro uses neural networks to predict domains using MSAs and predicted secondary structure information. Many proteins [Src, P53, IRS1 (Tompa, 2002)], especially in higher eukaryotes are multidomain in nature and have an architecture where the domains are spread over the sequence often with disordered or unstructured regions between them. Domain prediction methods are thus also related to methods for defining protein disorder. One such method is Globplot (Linding *et al*., 2003) which identifies domains by delineating the disordered, non-globular regions from the globular ones based on amino acid preferences. It is widely used by structural genomics and structural biologists to determine cloning constructs.

Up-to-date, users wishing to predict the domains of a particular sequence, need to know which methods are available and what their strengths and weaknesses are. Users wishing to use more than one method need to compile the individual results one by one. This paper describes a meta-server, Meta-DP, which provides a single platform for predicting domains using a number of available prediction servers. It should be noted that Meta-DP predicts only globular domains. To provide some insights on the performance of the various methods, we include the results of the evaluation of 13 servers on CAFASP4 prediction targets (http://bioinformatics.buffalo.edu/cafasp4/dp/status.html). Most of these servers have also participated in the CAFASP-DP (http://www.cs.bgu.ac.il/~dfischer/CAFASP4) experiment.

DESCRIPTION OF THE META-SERVER

The user is asked to submit the query sequence, the name of the query and his or her email address. Sequences submitted to Meta-DP must be in one-letter amino acid format. A sequence can either be pasted into the submission form or uploaded from a file. Users can select any of the available domain prediction servers to run, or use all by default. Meta-DP validates the input (sequence format and email address) and places the request into a processing queue which is further coupled to SQLite database engine. The targets in the database are monitored with their respective job numbers. The input is then converted into a format consisting of two lines, where the first line is the name of the query and the job number and the second line is the query sequence in one-letter code. The input is submitted by email to each domain prediction server. To avoid overloading the prediction servers, Meta-DP submits a new request to each server only if the previous requests have been completed.

The results of all servers are converted into a uniform format where the predicted domains are marked numerically as 1, 2 and so on (Fig. 1; see also http://meta-dp.bioinformatics.buffalo.edu/help/sample.html for a sample output). The unpredicted residues are marked as '-'. There are links to the output of each individual server. The query sequence is also compared against the conserved domain database (CDD) (Marchler-Bauer *et al*., 2003). Finally, a consensus prediction is also computed. The final HTML output is then emailed back to the user.

Currently, Meta-DP is coupled to ten different domain prediction servers, briefly described below:

Adda (Heger and Holm, 2003). ADDA first decomposes sequences into domains using BLASTP alignments derived from all-against-all sequence comparisons of nrdb40 database. After domain decomposition, it clusters domains into families using a profile–profile comparison between domains.

Biozon (Nagarajan and Yona, 2004). The method is based on the analysis of MSAs that are derived from a database search on NR. Multiple measures are defined to quantify the domain information content of each position along the sequence and are combined into a single predictor using a neural network. The output is further smoothed and post-processed using a probabilistic model to predict the domain boundaries.

Dompred–Domssea (Marsden et al., 2002). Domssea aligns the secondary structure predicted for a query protein against a database of domains assigned from 3D structures (CATH) (Orengo *et al*., 1997) and simply derives the domain boundaries from the known domain with the most similar secondary structure.

Dompro (http://www.ics.uci.edu/~baldig/dompro.html). Dompro predicts protein domain boundaries based on bidirectional

recurrent neural network and statistical methods. The input to the neural network is the PSI-BLAST generated PSSMs obtained by searching the target sequence against NR database and predicted secondary structure and solvent accessibility.

Dopro (von Ohsen et al., 2004). The method starts by constructing a set of subsequences from the query sequence, each subsequence representing a hypothesis for a possible protein domain. This is done by scanning against the InterPro database (Apweiler *et al*., 2001) and using secondary structure prediction from PSIPRED (Jones, 1999). Each of the potential domains is then subjected to five different fold-recognition methods from the Arby server (http://www.hnbioinfo.de, (von Ohsen *et al*., 2004)) and the highest scoring template in the database is computed. The search results are assessed using confidence measures. Finally, a set of non-overlapping annotations along the sequence is selected by performing combinatorial optimization of a heuristic score based on the confidence values. For each of these selected annotations, a separate protein domain is predicted.

Globplot (Linding et al., 2003). Globplot finds the putative domains by identifying the globular and non-globular regions within protein sequence based on the amino acid propensities for random coil (disordered) or secondary structure (ordered) regions.

InterProScan (Zdobnov and Apweiler, 2001). InterProScan output consists of matches obtained from different databases and is not designed to provide a single domain definition. Thus, in order to incorporate InterProScan into Meta-DP as a fully automated domain prediction server, its output needs to be analyzed. To obtain clear domain definitions and boundaries from InterProScan, its XML output is parsed and the information about the start and end position of the matches are extracted along with their respective *E*-values and status of the match. From the output, the true (*T*) matches of 'domain', 'family' and 'noIPR' entries are extracted and checked for overlaps. If there is no overlap, the matches are reported as such, otherwise, if there are overlaps from different databases, then the match is selected based on the following order: HMMPfam, HMMTigr, FPrintScan, HMMSmart, ProfileScan, Superfamily, BlastProDom and HMMPIR. If the overlap is from the same database, then the match is selected based on the better *E*-value. For identical *E*-values, the longest match is reported. This protocol for parsing the XML output and the order of reliability of databases was devised in consultation with the EBI support team.

Mateo (Lexa and Valle, 2003). Mateo is based on three scoring functions. The first serves as a safety component, where the sequence in question is compared with the SCOP database of folds and CDD domain database by reverse position specific blast (RPS-BLAST). Another component is the DomCut (Suyama and Ohara, 2003) score for the sequence (to evaluate a likelihood of a given word to form a domain boundary) and a Peptimex (Lexa and Valle, 2003) correlation score, which gives the likelihood of a given word to be surrounded by segments belonging to the same domain.

Ssep-domain (http://www.bio.ifi.lmu.de/cafasp/). Ssep-domain predicts domains in three steps. The first step consists of finding potential domain boundaries by aligning the target sequence against the template library of domains. In the second step, the similarity score for the domain region is calculated using a combination of secondary structure and log average profile–profile alignment on both sequence and secondary structure profiles. The third step consists of ranking all the valid combinations of non-overlapping domain regions according to a simple combination score.

Robetta–Ginzu (Chivian et al., 2003). Robetta–Ginzu scans the protein sequence with successively less confident methods of detection to determine any homolog with experimentally determined structures, starting with BLAST, and followed by the more remote fold-recognition methods FFAS03 (Jaroszewski *et al*., 2000; Rychlewski *et al*., 2000) and 3D-Jury (Ginalski and Rychlewski, 2003; Ginalski *et al*., 2003). After a homolog is identified, a search of remaining regions is done with HMMER (Eddy, 1998) against the Pfam-A (Bateman *et al*., 2000) protein family database. Finally, the PSI-BLAST (Altschul *et al*., 1997) MSA is used to assign regions of increased likelihood of having a contiguous domain based on sequence clusters. The final step consists of selecting cut-points between the domains using the PSI-BLAST MSA.

In addition to the individual results of the domain prediction servers, Meta-DP computes and reports a consensus prediction using a 'majority vote' (Fig. 1) (http://meta-dp.bioinformatics.buffalo.edu/help/sample.html). In case of tie, the decision is made using a weighting scheme that gives preference to the best performing methods according to our CAFASP-DP evaluation.

EVALUATION OF DOMAIN PREDICTION METHODS

Meta-DP was used to evaluate the performance of different domain prediction methods on the CAFASP4 set of 58 predictions targets. The 10 domain prediction methods listed above were evaluated along with other few methods that participated in CAFASP4 but are no longer available. A number of performance measures were used: absolute number of correctly predicted domains, sensitivity–specificity and overlap score (Jones *et al*., 1998) plots, separate analysis for single-domain, two-domain and all targets and separate evaluation on homology modeling and fold-recognition targets. Overall, the performance of most of the servers is better for single-domain targets than for two-domain targets. Most of the servers perform better than a random predictor. The best performers in the various measures were: rosettadom, consensus, ginzu, dopro, interproscan and ssep-domain. The detailed evaluation parameters and results are available in the Supplementary section.

CONCLUSION

We have described the availability of Meta-DP, a new meta-server for prediction of domains in proteins. Meta-DP is extensible and new methods will be incorporated as they become available. Meta-DP is a convenient centralized site of fully automated methods for domain prediction. Insights regarding the performance of the various methods have already been obtained during the CAFASP-DP experiment.

ACKNOWLEDGEMENT

We thank Dr Maria Krestyaninova at EBI for her help in understanding the InterProScan output. This work was partially supported by the GENEFUN European Union Grant.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Apweiler,R. *et al*. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.

Bateman,A. *et al*. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Brändén,C.-I. and Tooze,J. (1991) *Introduction to Protein Structure*. Garland Publishing, New York.

Chivian,D. *et al*. (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins*, **53** (Suppl. 6), 524–533.

Corpet,F. *et al*. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Falquet,L. *et al*. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.

Fischer,D. *et al*. (1996) Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J.*, **10**, 126–136.

Ginalski,K. and Rychlewski,L. (2003) Detection of reliable and unexpected protein fold predictions using 3D-Jury. *Nucleic Acids Res.*, **31**, 3291–3292.

Ginalski,K. *et al*. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.

Gough,J. *et al*. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.

Haft,D.H. *et al*. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.*, **29**, 41–43.

Heger,A. and Holm,L. (2003) Exhaustive enumeration of protein domain families. *J. Mol. Biol.*, **328**, 749–767.

Jaroszewski,L. *et al*. (2000) Improving the quality of twilight-zone alignments. *Protein Sci.*, **9**, 1487–1496.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Jones,S. *et al*. (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.*, **7**, 233–242.

Lexa,M. and Valle,G. (2003) PRIMEX: rapid identification of oligonucleotide matches in whole genomes. *Bioinformatics*, **19**, 2486–2488.

Linding,R. *et al*. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.

Marchler-Bauer,A. *et al*. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.

Marsden,R.L. (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.*, **11**, 2814–2824.

Nagarajan,N. and Yona,G. (2004) Automatic prediction of protein domains from sequence information using a hybrid learning system. *Bioinformatics*, **20**, 1335–1360.

Orengo,C.A. *et al*. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Rychlewski,L. *et al*. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.

Schultz,J. *et al*. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, **28**, 231–234.

Suyama,M. and Ohara,O. (2003) DomCut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics*, **19**, 673–674.

Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.

von Ohsen,N. *et al*. (2004) Arby: automatic protein structure prediction using profile–profile alignment and confidence measures. *Bioinformatics*, **20**, 2228–2235.

Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.