DOMpro: Protein Domain Prediction Using Profiles, Secondary Structure, Relative Solvent Accessibility, and Recursive Neural Networks

Jianlin Cheng jianlinc@ics.uci.edu
Michael J. Sweredoski msweredo@ics.uci.edu
Pierre Baldi pfbaldi@ics.uci.edu
Institute for Genomics and Bioinformatics, School of Information and Computer Sciences, University of California Irvine, Irvine, CA 92697, USA

Abstract. Protein domains are the structural and functional units of proteins. The ability to parse protein chains into different domains is important for protein classification and for understanding protein structure, function, and evolution. Here we use machine learning algorithms, in the form of recursive neural networks, to develop a protein domain predictor called DOMpro. DOMpro predicts protein domains using a combination of evolutionary information in the form of profiles, predicted secondary structure, and predicted relative solvent accessibility. DOMpro is trained and tested on a curated dataset derived from the CATH database. DOMpro correctly predicts the number of domains for 69% of the combined dataset of single and multi-domain chains. DOMpro achieves a sensitivity of 76% and specificity of 85% with respect to the single-domain proteins and sensitivity of 59% and specificity of 38% with respect to the two-domain proteins. DOMpro also achieved a sensitivity and specificity of 71% and 71% respectively in the Critical Assessment of Fully Automated Structure Prediction 4 (CAFASP-4) (Fischer *et al.*, 1999; Saini and Fischer, 2005) and was ranked among the top *ab initio* domain predictors. The DOMpro server, software, and dataset are available at *http://www.igb.uci.edu/servers/psss.html.*

Keywords: protein structure prediction, domain, recursive neural networks

1. Introduction

Domains are considered the structural and functional units of proteins. They can be defined using multiple criteria, or combinations of criteria, including evolutionary conservation, discrete functionality, and the ability to fold independently (Holm and Sander, 1994). A domain can span an entire polypeptide chain or be a subunit of a polypeptide chain that can fold into a stable tertiary structure independently of any other domain (Levitt and Chothia, 1976). While typical domains consist of a single continuous polypeptide segment, some domains may be comprised of several discontinuous segments.

The identification of domains is an important step for protein classification and for the study and prediction of protein structure, function, and evolution. The topology of secondary structure elements in a domain is used by human experts or automated systems in structural classification databases such as FSSP-Dali Domain Dictionary (Holm and Sander, 1998*a*; Holm and Sander, 1998*b*), SCOP (Murzin *et al.*, 1995), and CATH (Orengo *et al.*, 2002). The prediction of protein tertiary structure, especially *ab initio* prediction, can be improved by segmenting the protein using the putative domain boundaries and predicting each domain independently (Chivian *et al.*, 2003). However, the identification of protein domains based on sequence alone remains a challenging problem.

A number of methods have been developed to identify protein domains starting from their primary sequence. These methods can be roughly classified into three categories: template based methods (Chivian *et al.*, 2003; Heger and Holm, 2003; Marsden *et al.*, 2002; von Ohsen *et al.*, 2004; Zdobnov and Apweiler, 2001; Gewehr *et al.*, 2005), non-template based (*ab initio*) methods (Bryson *et al.*, 2005; George and Heringa, 2002; Lexa and Valle, 2003; Linding *et al.*, 2003; Liu and Rost, 2004; Nagarajan and Yona, 2004; Wheelan *et al.*, 2000), and meta domain prediction methods (Saini and Fischer, 2005). Some template-based methods use a sequence alignment approach where domains are identified by aligning the target sequence against sequences in a domain classification database (Marchler-Bauer *et al.*, 2003). Other methods use alignments of secondary structures (Marsden *et al.*, 2002). In these methods, domains are assigned by aligning the predicted secondary structure of a target sequence against the secondary structure of chains in CATH, which have known domain boundaries.

Some *ab initio* methods, such as tertiary structure folding approaches, average several hundred predictions obtained from coarse *ab initio* simulations of protein folding to assign domain boundaries to a given sequence (George and Heringa, 2002). One drawback of these approaches is that they are computationally intensive. Other *ab initio* use a statistical approach, such as Domain Guess by Size (Wheelan *et al.*, 2000), to predict the likelihood of domain boundaries within a given sequence based on the distributions of chain and domain lengths.

The *ab initio* prediction of domains using machine learning techniques is aided by the availability of large, high quality, domain classification databases such as CATH, SCOP and FSSP-Dali Domain Dictionary. Two recently published algorithms attempt to predict domain boundaries using neural networks (Nagarajan and Yona, 2004; Liu and Rost, 2004). The networks used by Nagarajan and Yona (2004) incorporate the position specific physio-chemical properties of amino acid and predicted secondary structure. Liu and Rost (2004) use neural networks with amino acid composition, positional evolutionary conservation, as well as predicted secondary structure and solvent accessibility.

Here we describe DOMpro, an *ab initio* machine learning approach for predicting domains, which uses profiles along with predicted secondary structure and solvent accessibility in a 1D-recursive neural network (1D-RNN). These networks are also used for the prediction of secondary structure and solvent accessibility (Pollastri *et al.*, 2001; Pollastri *et al.*, 2002) in the SCRATCH suite of servers (Baldi and Pollastri, 2003; Cheng *et al.*, 2005*a*). Unlike previous neural network-based approaches (Liu and Rost, 2004; Nagarajan and Yona, 2004), the direct use of profiles in DOMpro is based on the assumption that sequence motifs and their level of conservation in the boundary regions are different from those found in the rest of the protein. The final assignment of protein domains is the result of post-processing and statistical inference on the output of the 1D-RNN.

2. Methods

2.1. Data

DOMpro is trained and tested on a curated dataset derived from the annotated domains in the CATH domain database, version 2.5.1. Because the CATH database contains only the sequences of domain regions, sequences from the Protein Data Bank (PDB)(Berman *et al.*, 2000) must be incorporated to reconstruct entire chains. Once the chains are reconstructed, short sequences (< 40 residues) are filtered out.

UniqueProt (Mika and Rost, 2003) is then used to reduce sequence redundancy in the dataset by ensuring that no pair of sequences have a HSSP value greater than 5. The HSSP value between two sequences is a measure of their similarity and takes into account both sequence identity and sequence length. A HSSP value of 5 corresponds roughly to a sequence identity of 25% in a global alignment of length 250.

Finally, the secondary structure and relative solvent accessibility are predicted for each chain using SSpro and ACCpro (Baldi and Pollastri, 2003; Pollastri *et al.*, 2001; Pollastri *et al.*, 2002). Using predicted secondary structure and solvent accessibility values rather than the true values, which can be easily obtained using the DSSP program (Kabsch and Sander, 1983), gives us a more realistic and objective evaluation since the actual secondary structure and solvent accessibility are not known during the prediction phase. To leverage evolutionary information, PSI-BLAST (Altschul *et al.*, 1997) is used to generate profiles by aligning all chains against the Non-Redundant (NR) database, as in other methods (Jones, 1999; Przybylski and Rost, 2002; Pollastri *et al.*, 2001).

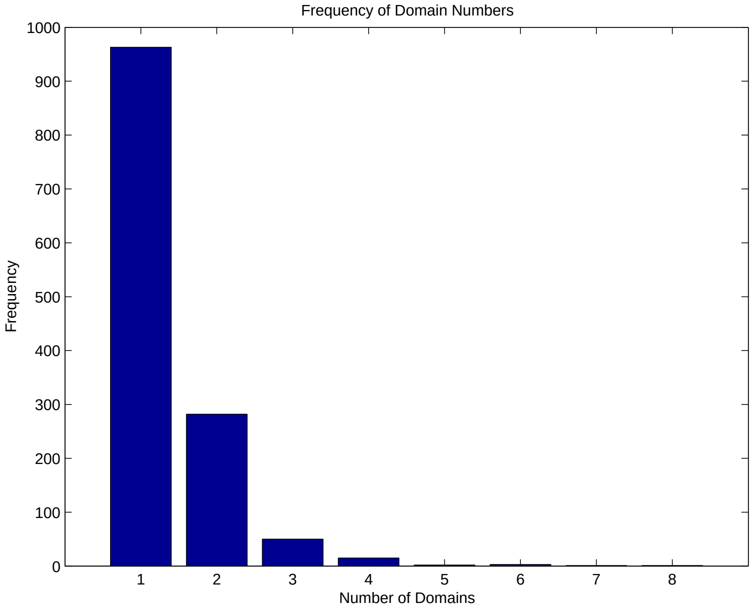


Figure 1. Frequency of single and multi-domain chains in the redundancy-reduced dataset.

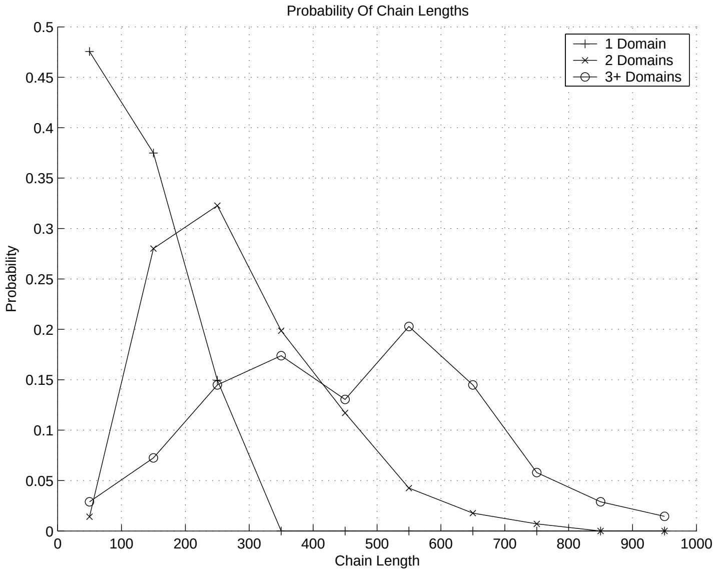


Figure 2. Distributions of the lengths of single and multi-domain chains in the redundancy-reduced dataset.

After redundancy reduction, our curated dataset contained 354 multi-domain chains and 963 single-domain chains. The ratio of single to multi-domain chains reflects the skewed distribution of single-domain chains in the PDB. Figure 1 shows the frequency of single and multi-domain chains in the redundancy-reduced dataset. Figure 2 shows the distribution of chain lengths among single and multi-domain chains.

Because the recursive neural networks are trained to recognize domain boundaries, only multi-domain proteins are used during the training process. During the training and testing of the neural networks on multi-domain proteins, ten fold cross-validation is used. Additional testing is performed on single-domain proteins using models trained with multi-domain proteins.

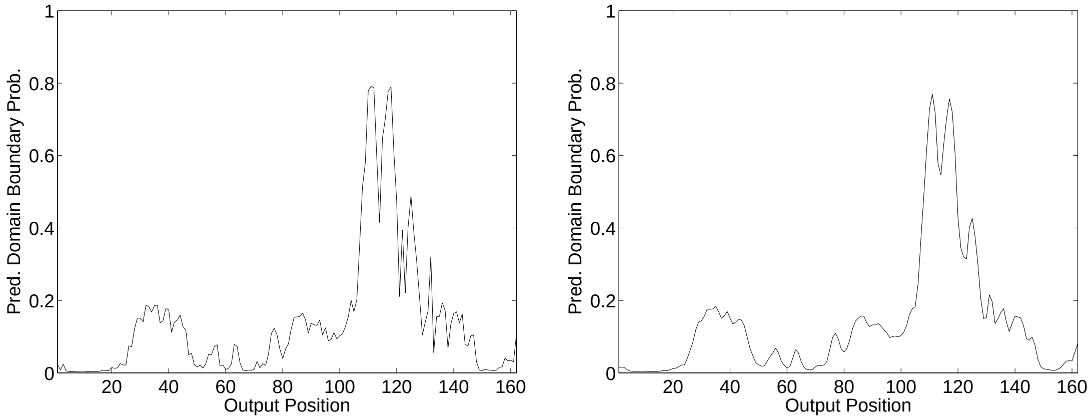
2.2. The Inputs and Outputs of the One Dimensional Recursive Neural Network

The problem of predicting domain boundaries can be viewed as a binary classification problem for each residue along a one-dimensional protein chain. Each residue is labeled as being either a domain boundary residue or not.

Specifically, the target class for each residue is defined as follows. Following the conventions used in prior domain boundary prediction papers (Liu and Rost, 2004; Marsden *et al.*, 2002), residues within 20 amino acids of a domain boundary are considered domain boundary residues and all other residues are considered non-boundary residues. A variety of machine learning methods can be applied to this classification problem, such as probabilistic graphical models, kernel methods, and neural networks. DOMpro employs 1-D recursive neural networks (1D-RNNs) (Baldi and Pollastri, 2003), which have been applied successfully in the prediction of secondary structure, solvent accessibility, and disordered regions (Cheng *et al.*, 2005*b*; Pollastri *et al.*, 2001; Pollastri *et al.*, 2002). For each chain, the input is the array I, where the length of I is equal to the number of residues in the chain. Each element I_i is a vector with 25 components, which encodes the profile as well as secondary structure and relative solvent accessibility at position i. Twenty components of the vector I_i are real numbers corresponding to the amino acid profile probabilities. The other five components are binary: three correspond to the predicted secondary structure class of the residue (Helix, Strand, or Coil) and two correspond to the predicted relative solvent accessibility of the residue (i.e., under or over 25% exposed).

The training target for each chain is the 1-D binary array T, where each T_i equals 1 or 0 depending on whether or not the residue at position i is within a boundary region. Neural networks (and most other machine learning methods) can be trained on the dataset to learn a mapping from the input array I onto an output array O, where O_i is the predicted probability that the residue at position i is within a domain boundary region. The goal is to make the output O as close as possible to the target T.

2.3. Post-Processing of the 1D-RNN Output



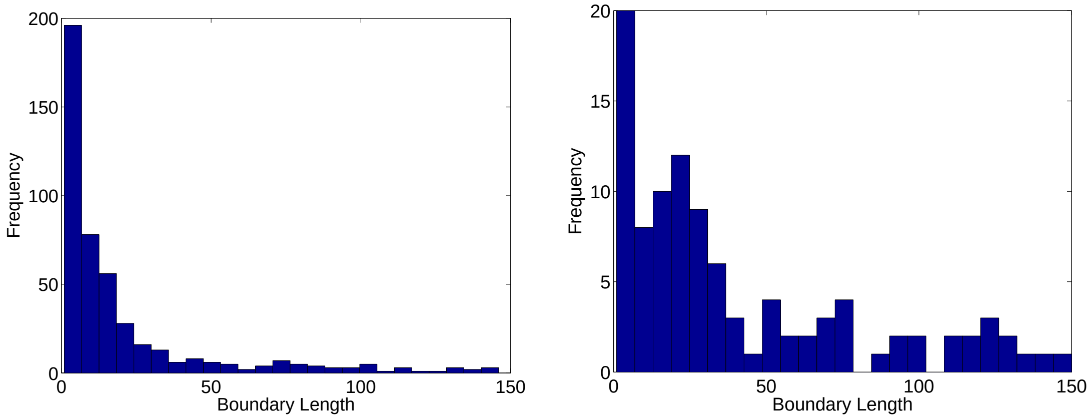
Raw output from 1D-RNN

Smoothed output from 1D-RNN with window of width 3

Figure 3. Example of smoothing applied to the raw output from the 1D-RNN

The raw output from the 1D-RNN is quite noisy (see Figure 3). DOMpro uses smoothing to help correct for the random noise that is the result of false positive hits. The smoothing is accomplished by averaging over a window of length three around each position. Figure 3 shows how this smoothing technique helps to reduce the noise found in the raw output of the 1D-RNN. After smoothing, a domain state (boundary/not boundary) is assigned to each residue by thresholding the network's output at 0.5.

While smoothing the neural network's output helps correct for random spikes, it does not necessarily create the long, continuous segments of boundary residues that are required for domain assignment. Therefore, further inference on the output is required.



False positive boundaries

True boundaries

Figure 4. Histograms of length distributions for false positive and true positive boundary regions

DOMpro infers the domain boundary regions from the residues predicted as domain boundaries by pattern matching on the discretized output. Any section of the output that matches the regular expression pattern $((B+N\{0,m\})+B+)$ is considered a domain boundary region, where B is a predicted boundary residue, N is a predicted non-boundary residue and m is the maximum separation between two boundary residues that should be merged into one region.

Once DOMpro has inferred all possible domain boundary regions, it needs to identify false positive domain boundary regions. DOMpro considers the boundary region's length a measure of its signal strength. Figure 4 shows that there is a clear difference between the length distributions of true domain boundary regions and false domain boundary regions. Based on these statistics, domain boundary regions shorter than three residues are considered false positive hits and are ignored. The target sequence is then cut into domain segments at the middle residue of each boundary region. A target sequence with no predicted domain boundaries is classified as a single-domain chain. The final step of DOMpro is to assign domain numbers to each predicted domain segment. Our method simply assigns each domain segment to a separate domain, ignoring at this time the relatively rare problem of non-contiguous domains.

3. Results

The evaluation and comparison of domain predictors is complicated by the existence of several domain datasets/databases that sometimes conflict with each other (Liu and Rost, 2004). Thus, the performance of a predictor on a dataset other than its training dataset is limited by the percentage of agreement between the training and testing datasets. With this caveat in mind, we observe that DOMpro correctly predicts the number of domains for 69% of the combined dataset of single and multi-domain proteins. DOMpro achieves a sensitivity of 76% and specificity of 85% with respect to the single-domain proteins and sensitivity of 59% and specificity of 38% with respect to the two-domain proteins.

Table I. CAFASP-4 Evaluation Results

Predictor	1-D Sen.	1-D Spec.	2-D Sen.	2-D Spec.	All Sen.	All Spec.
DOMpro	0.85	0.76	0.35	0.50	0.71	0.71
ADDA (Heger and Holm, 2003) †‡	0.85	0.73	0.18	0.33	0.66	0.67
Armadillo †	0.10	1.00	0.24	0.18	0.14	0.31
Biozon (Nagarajan and Yona, 2004) †	0.10	1.00	0.35	0.19	0.17	0.29
Dompred-Domssea (Marsden *et al.*, 2002) ‡	0.80	0.75	0.29	0.63	0.66	0.73
Dompred-DPS (Bryson *et al.*, 2005) †	0.68	0.78	0.47	0.50	0.62	0.69
Dopro (von Ohsen *et al.*, 2004) ‡	0.85	0.88	0.53	0.64	0.76	0.81
Globplot (Linding *et al.*, 2003) †	0.83	0.71	0.18	0.60	0.64	0.70
InterProScan (Zdobnov and Apweiler, 2001) ‡	0.93	0.75	0.24	0.67	0.72	0.74
Mateo (Lexa and Valle, 2003) †	0.51	0.78	0.12	0.15	0.40	0.58
SSEP-Domain (Gewehr *et al.*, 2005) ‡	0.93	0.84	0.47	0.73	0.79	0.82
Robetta-Ginzu (Chivian *et al.*, 2003) ‡	0.80	0.92	0.53	0.69	0.72	0.86
Robetta-Rosettadom ‡	0.83	0.94	0.71	0.75	0.79	0.88

† had lower sensitivity and specificity averaged over all targets compared to DOMpro

‡ template based methods

The precise prediction of domain boundaries for multi-domain proteins is more difficult than the prediction of the number of domains (domain number). DOMpro is able to correctly predict the domain number and boundary for 25% of the two-domain proteins in our dataset derived from CATH. Additionally, DOMpro is able to correctly predict both the domain number and domain boundary location for 20% of the multi-domain chains. For the evaluation of multi-domain chains, we consider that a domain boundary has been correctly identified if the predicted domain boundary is within 20 residues of the true domain boundary as annotated in the CATH database. This definition is consistent with previous work (Marsden *et al.*, 2002).

DOMpro was independently evaluated along with 12 other predictors in the Critical Assessment of Fully Automated Structure Prediction 4 (CAFASP-4)(Fischer *et al.*, 1999; Saini and Fischer, 2005). The results, kindly provided by Dr. Saini, are available at *http://cafasp4.bioinformatics.buffalo.edu/dp/update.html*. The evaluation set consisted of 41 single-domain CASP6 targets and 17 two-domain CASP6 targets (58 targets in total). Since this evaluation set contains only comparative modeling and fold recognition targets (no new fold targets), predictors based on templates have an advantage in this evaluation. DOMpro achieved a higher sensitivity and specificity than one method that uses homologous information and all other *ab initio* predictors averaged over all of the targets (See Table I and Figure 5). However, the performance of the top three *ab initio* predictors (DOMpro, Globplot, and Dompred-DPS) is close. The specificity and sensitivity of DOMpro is 4-5% higher than the template-based method ADDA, similar to Dompred-Domssea, and lower than other template-based methods such as Dopro, SSEP-Domain, and Robetta-Ginzu.

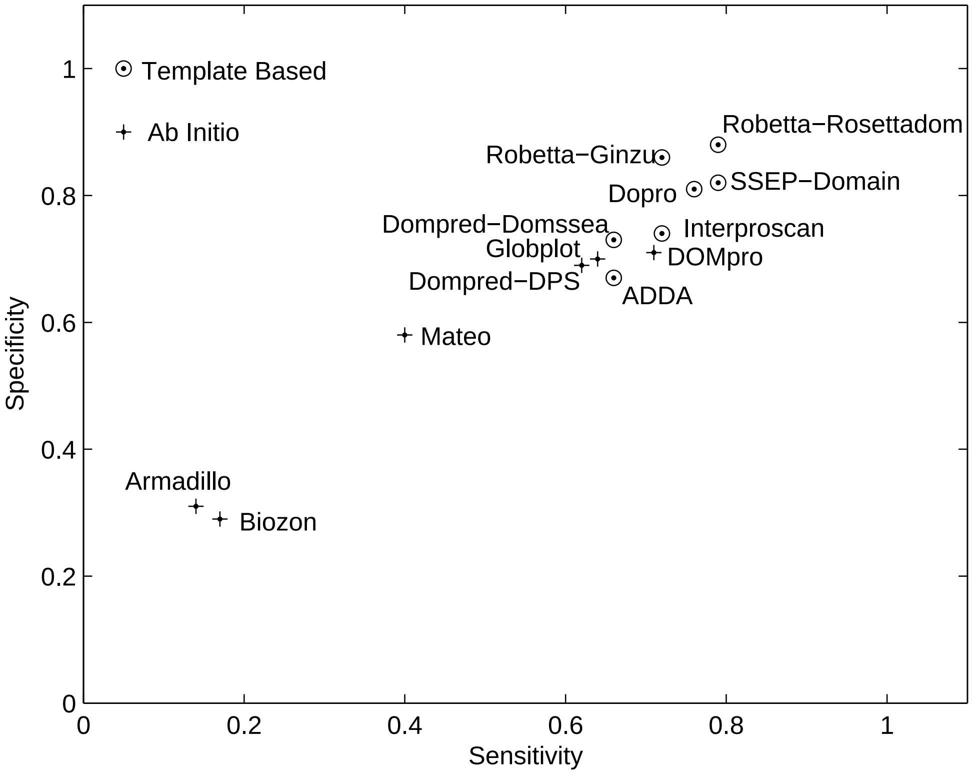


Figure 5. Sensitivity vs specificity in CAFASP-4

4. Conclusions

We have created DOMpro, an *ab initio* predictor of protein domains using a recursive neural network that leverages evolutionary information in the form of profiles and predicted secondary structure and relative solvent accessibility. The raw output of the 1D-RNN in DOMpro goes through a post-processing procedure to produce the final domain segmentation and assignment. In the CAFASP-4 evaluation, DOMpro was ranked among the top *ab initio* domain predictors.

Despite recent advances, domain prediction remains a challenge. A 25% accuracy on the prediction of two-domain proteins is encouraging but not sufficient for most applications and clearly there is room for improvement. We are currently adding a module to DOMpro which would incorporate known domain assignments for proteins that are homologous to structures in the PDB and CATH databases. We are also training ensembles of predictors, although preliminary experiments so far have not lead to significant improvements. In addition, we are focusing on the prediction/classification of discontinuous domains. To overcome the current limitations of DOMpro and the naive assignment of domain numbers, we are experimenting with the use of predicted contact maps, as well as domain length statistics, in the assignment of domain boundaries and the creation of domains consisting of multiple non-adjacent domain segments. The contact maps are predicted using 2D-RNNs (Baldi and Pollastri, 2003; Pollastri and Baldi, 2002). The basic idea is that domains should be associated with a relatively higher density of contacts. Following this logic, two discontinuous segments having the proper length statistics and a sufficient number of inter-segment residue-residue contacts would be predicted to be in the same domain.

Acknowledgments

This work is supported by the Institute for Genomics and Bioinformatics at UCI, a Laurel Wilkening Faculty Innovation award, an NIH Biomedical Informatics Training grant (LM-07443-01), an NSF MRI grant (EIA-0321390), a Sun Microsystems award, a grant from the University of California Systemwide Biotechnology Research and Education Program (UC BREP) to PB.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* **25** (17), 3389–3402.

Baldi,P. and Pollastri,G. (2003) The principled design of large-scale recursive neural network architectures-DAG-RNNs and the protein structure prediction problem. *Journal of Machine Learning Research,* **4**, 575–602.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Research,* **28**, 235–242.

Bryson,K., McGuffin,L.J., Marsden,R.L., Ward,J.J., Sodhi,J.S. and Jones,D.T. (2005) Protein structure prediction servers at University College London. *Nucleic Acids Research,* **33**, w36–38.

Cheng,J., Randall,A.Z., Sweredoski,M.J. and Baldi,P. (2005*a*) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research,* **33**, w72–76.

Cheng,J., Sweredoski,M.J. and Baldi,P. (2005*b*) Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery,* (In press).

Chivian,D., Kim,D.E., Malmstrom,L., Bradley,P., Robertson,T., Murphy,P., Strauss,C.E., Bonneau,R., Rohl,C.A. and Baker,D. (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins,* **53** (S6), 524–533.

Fischer,D., Barret,C., Bryson,K., Elofsson,A., Godzik,A., Jones,D., Karplus,K.J., Kelley,L.A., MacCallum,R.M., Pawowski,K., Rost,B., Rychlewski,L. and Sternberg,M. (1999) CAFASP-1: Critical assessment of fully automated structure prediction methods. *Proteins,* **Suppl 3**, 209–217.

George,R.A. and Heringa,J. (2002) SnapDRAGON: a method to delineate protein structural domains from sequence data. *Journal of Molecular Biology,* **316**, 839–851.

Gewehr,J.E. and Zimmer,R. (2005) SSEP-Domain:protein domain prediction by alignment of secondary structure elements and profiles. *Bioinformatics,* , In press.

Heger,A. and Holm,L. (2003) Exhaustive enumeration of protein domain families. *Journal of Molecular Biology,* **328**, 749–767.

Holm,L. and Sander,C. (1994) Parser for protein folding units. *Proteins,* **19**, 256–268.

Holm,L. and Sander,C. (1998*a*) Dictionary of recurrent domains in protein structures. *Proteins,* **33**, 88–96.

Holm,L. and Sander,C. (1998*b*) Touring protein fold space with Dali/FSSP. *Nucleic Acids Research,* **26**, 316–319.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology,* **292**, 195–202.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577–2637.

Levitt,M. and Chothia,C. (1976) Structural patterns in globular proteins. *Nature,* **261** (5561), 552–558.

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

Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research,* **31**, 3701–3708.

Liu,J. and Rost,B. (2004) Sequence-based prediction of protein domains. *Nucleic Acids Research,* **32** (12), 3522–3530.

Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J., Liebert,C.A., Liu,C., Madej,T., Marchler,G.H., Mazumder,R., Nikolskaya,A.N., Panchenko,A.R., Rao,B.S., Shoemaker,B.A., Simonyan,V., Song,J.S., Thiessen,P.A., Vasudevan,S., Wang,Y., Yamashita,R.A., Yin,J.J. and Bryant,S.H. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Research,* **31** (1), 383–387.

Marsden,R.L., McGuffin,L.J. and Jones,D.T. (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Science,* **11**, 2814–2824.

Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Research,* **31** (13), 3789–3791.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology,* **247**, 536–540.

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., Bray,J.E., Buchan,D.W., Harrison,A., Lee,D., Perl,F.M., Sillitoe,I., Todd,A.E. and Thornton,J.M. (2002) The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics,* **2**, 11–21.

Pollastri,G. and Baldi,P. (2002) Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics,* **18** (Suppl 1), S62–S70. Proceeding of the ISMB 2002 Conference.

Pollastri,G., Baldi,P., Fariselli,P. and Casadio,R. (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins,* **47**, 142–153.

Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2001) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins,* **47**, 228–235.

Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins,* **46**, 197–205.

Saini,H.K. and Fischer,D. (2005) Meta-DP: domain prediction meta server. *Bioinformatics,* **21**, 2917–2920.

von Ohsen,N., Sommer,I., Zimmer,R. and Lengauer,T. (2004) Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics,* **20**, 2228–2235.

Wheelan,S.J., Marchler-Bauer,A. and Bryant,S.H. (2000) Domain size distributions can predict domain boundaries. *Bioinformatics,* **16** (7), 613–618.

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