W72–W76 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki396

SCRATCH: a protein structure and structural feature prediction server

J. Cheng, A. Z. Randall, M. J. Sweredoski and P. Baldi*

Institute for Genomics and Bioinformatics, University of California, Irvine, CA, USA

Received February 13, 2005; Revised and Accepted March 14, 2005

ABSTRACT

SCRATCH is a server for predicting protein tertiary structure and structural features. The SCRATCH software suite includes predictors for secondary structure, relative solvent accessibility, disordered regions, domains, disulfide bridges, single mutation stability, residue contacts versus average, individual residue contacts and tertiary structure. The user simply provides an amino acid sequence and selects the desired predictions, then submits to the server. Results are emailed to the user. The server is available at http://www.igb.uci.edu/servers/psss.html.

INTRODUCTION

Knowledge of a protein's structure provides insight into how it can interact with other proteins, DNA/RNA, and small molecules. It is these interactions which define the protein's function and biological role in an organism. Thus, protein structure and structural feature prediction is a fundamental area of computational biology. Its importance is exacerbated by large amounts of sequence data coming from genomics projects and the fact that experimentally determining protein structures remains expensive and time consuming.

Publicly available bioinformatics web servers allow researchers from around the world to apply tools developed in other laboratories to their own data and fully automated systems provide a framework for high-throughput proteomics and protein engineering projects. We have developed a web server, SCRATCH, to predict protein tertiary structure and structural features.

METHODS

The SCRATCH suite combines machine learning methods, evolutionary information in the form of profiles, fragment libraries extracted from the Protein Data Bank (PDB) (1), and energy functions to predict protein structural features and tertiary structures. See Table 1 for a summary of the specific methods used by each predictor. The suite includes the following main modules:

(i) SSpro (2): three class secondary structure.
(ii) SSpro8 (2): eight class secondary structure.
(iii) ACCpro (3): relative solvent accessibility.
(iv) CONpro (3): contacts with other residues compared to average.
(v) DOMpro: domain boundaries.
(vi) DISpro: disordered regions.
(vii) MUpro: effect of single amino acid mutation on stability.
(viii) DIpro (4): disulfide bridges.
(ix) CMAPpro (5,6) : residue-residue contact maps.
(x) 3Dpro: tertiary structure.

Structural feature predictors

All predictors are trained in a supervised fashion using curated, non-redundant, datasets extracted from the PDB. SSpro, SSpro8, ACCpro, CONpro, DISpro and DOMpro use ensembles of one-dimensional recursive neural network (1D-RNN) architectures (6). CMAPpro and DIpro predictors use ensembles of 2D-RNN architectures (5,6). DIpro also uses support vector machines (SVMs) to discriminate proteins with disulfide bonds from proteins without disulfide bonds, and graph matching algorithms to pair the cysteines. MUpro uses feed-forward neural networks and SVMs.

These RNN architectures are based on the theory of probabilistic graphical models (Bayesian networks) meshed with a neural network parameterization to accelerate belief propagation and learning. These architectures systematically combine standard information contained in a local input window with more distant contextual information extracted by translation-invariant recursive neural networks that are convolved along the entire length of the protein (1D) or of the contact maps (2D) from all possible directions.

All predictors, except 3Dpro and MUpro, directly leverage evolutionary information in the form of input profiles derived using PSI-BLAST (7) to include all homologous proteins (8,9). In addition, for SSpro and ACCpro, very high levels of local homology to known structures are used either directly

*To whom correspondence should be addressed. Tel: +1 949 824 5809; Fax: +1 949 824 4056; Email: pfbaldi@ics.uci.edu

Table 1. Summary of methods used in SCRATCH predictors

	PDB training set	PSI-BLAST profile	Feed forward NN	1D RNN	2D RNN	SVM	Graph matching	Direct homology	Fragment database
SSpro	D	D		D				D	
SSpro8	D	D		D					
ACCpro	D	D		D					
CONpro	D	D		D					
DOMpro	D	D		D					
DISpro	D	D		D					
MUpro	D		D			D			
DIpro	D	D			D	D	D		
CMApro	D	D		I	D				
3Dpro	I	I		I	I			I	D

'D' indicates a method is used directly, and 'I' indicates a method is used in one or more of the predictors that are used as an input to the predictor.

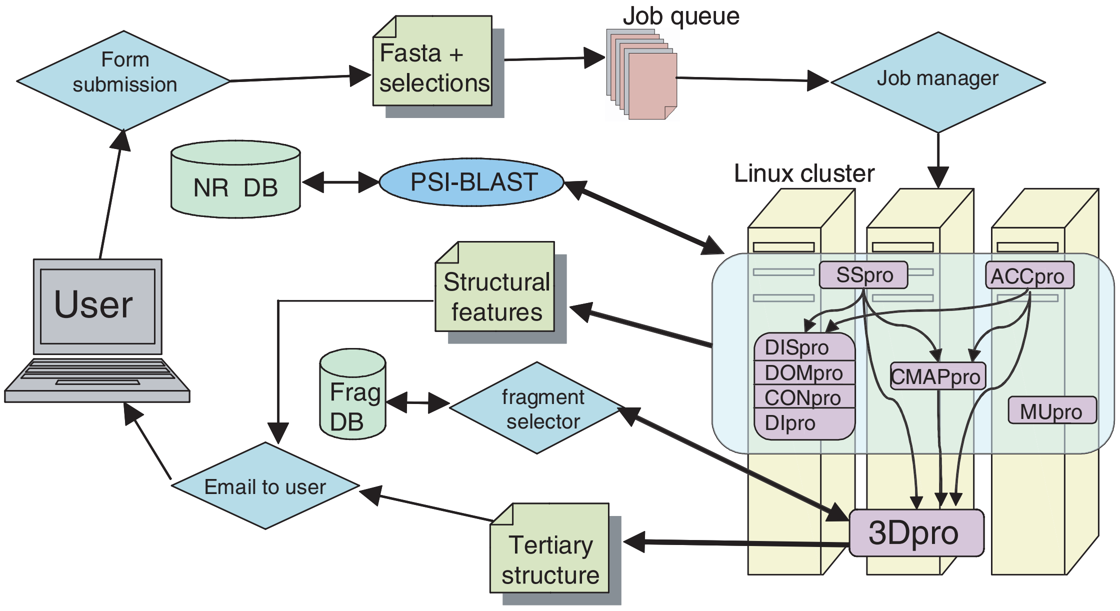

Figure 1. Flow diagram for the SCRATCH server. DISpro, DOMpro, CONpro and DIpro are grouped together because they have the same inputs and their outputs are not used by other predictors; however, they are standalone programs.

or in combination with the prediction output to improve accuracy. Whenever possible and useful, predictors leverage the output of the other predictors (see Figure 1).

DOMpro produces domain predictions in three steps. First, DOMpro predicts whether a residue belongs to a domain boundary region or not. Residues within 20 amino acids from the actual domain boundary in the CATH (10) database are considered to be part of the domain boundary region. Second, a statistical approach is used to infer the domain boundary from the predicted states (boundary/non-boundary) of the individual residues. Finally, the sequence segments separated by domain boundary are assigned to domain numbers.

In addition to the standard 2D-RNN architectures (5,6) to predict the entire contact map in one step, a variant architecture is used to predict contacts from low-sequence separation to high-sequence separation step by step. The predicted contact maps at lower sequence separation are used as inputs for the prediction of contact maps at higher sequence separation. The raw output of CMAPpro is a matrix of contact probabilities for all residue pairs.

Tertiary structure prediction

Our approach to tertiary structure prediction (3Dpro) combines the use of predicted structural features (2,3,5,6), a fragment library (11) and energy terms derived from the PDB statistics. The structural features used are secondary structure, relative solvent accessibility, and a residue level contact map at a distance cut-off of 12 Å. The predicted structural features are used in the energy function. We use a database of protein fragments of length nine, constructed from the structures in the PDB (11).

Two terms in the energy function are based directly on statistics from the PDB, one for residue environments (11,12) and another for bond angles. To encourage the formation of β-strands into sheets, we use a simple, single vector, representation of each strand and penalize unpaired strand vectors.

We include a contact-map energy term (13) based on a binary map derived from the matrix of contact probabilities predicted by CMAPpro. To select the contacts, we use a variable, band-dependent, threshold determined by estimating the total number of contacts in a band from the sum of all the predicted contact probabilities in that band.

The conformational space is searched using a variant of simulated annealing, where the moves we use to modify our models are crankshaft moves (13) on one or more residues and several forms of fragment replacement (11,12). These moves are applied to sequence locations in the model that are selected randomly. We also include a term to encourage the secondary structure of the models to match the predicted secondary structure. During each search, the model with the lowest energy is kept and all the other models are discarded. Many models are produced using different seeds randomly for each search. The single model with the lowest score is returned as the prediction.

New in SCRATCH

SCRATCH is continuously updated; as new methods are developed, existing methods are improved and predictors are retrained on larger datasets. The predictors DOMpro, DISpro, MUpro and 3Dpro are all new. SSpro and ACCpro have been improved by incorporating information from structural templates directly when appropriate. DIpro has been improved by adding secondary structure and relative solvent accessibility as inputs to increase the accuracy of disulfide bonds prediction. Also, the use of SVMs to discriminate proteins having disulfide bonds from proteins without disulfide bonds is new. One change to CMAPpro is the variant architecture discussed in Methods.

INPUT AND OUTPUT FORMAT

Input

The input to the server is provided by the user through a simple HTML form. The user must enter an email address for the results to be sent to and the single letter code for an amino acid sequence. The user may also enter a name for the submission. The user may select multiple predictions for the same submission. MUpro is the exception to this simple input format. Input for MUpro is the single letter code for an amino acid sequence, single mutation site and a new residue to use for replacement. The user may also provide a structure file in the PDB format, but the field is optional. The MUpro prediction results are displayed directly in the browser shortly after submission.

Output

The predictions are returned to the email address provided by the user. The output from SSpro, SSpro8, ACCpro, CONpro, DOMpro, DIpro and DISpro comes in the body of the email with subject: 'SCRATCH structural feature predictions'. The CMAPpro predictions are included as an attachment to the email. The 3Dpro prediction is returned as an attachment in a separate email with subject: 'SCRATCH tertiary structure prediction'. Here we describe the output of the individual predictors.

SSpro: 'H' helix, 'E' strand, 'C' other.

SSpro8: Single letter eight secondary structure class code defined by DSSP (14).

ACCpro: 'e' exposed, '−' buried.

CONpro: '+' more contacts than average, '−' fewer contacts than average.

DISpro: 'O' ordered, 'D' disordered.

DOMpro: First and last residue of each domain.

DIpro: Two class prediction of whether or not the target has disulfide bonds. Predicted bonding state of each cysteine in the protein. Predicted cysteine pairs.

MUpro: A statement of whether the protein stability is predicted to be increased or decreased by the mutation, and a confidence score. A score near 0 means unchanged stability. Score near −1 means high confidence in decreased stability. Score near +1 means high confidence in increased stability.

CMAPpro: The contact map predictions are included as an attachment to the structural features email message. Predictions come as attached raw files, with extension contact_map.8a and contact_map.12a, for thresholds of 8 and 12 Å, respectively. If the query is N amino acids long the files are composed of N lines, each containing N space-separated real numbers. The j-th number on line i-th represents the estimated probability that amino acids i and j are in contact (i.e. of their C-αs being closer than the threshold).

3Dpro: The tertiary structure prediction is a PDB file sent in a separate email message as an attachment, because it takes a significantly longer time to produce than the other predictions. The PDB file contains only the carbon alpha trace. To obtain an all-atom model a user may use other software to add the back bone, such as MaxSprout (15), and side chains, such as MaxSprout and SCWRL (16).

SERVER IMPLEMENTATION AND PERFORMANCE

Statistics

The SCRATCH system has handled ∼175 000 jobs since March 2000, including submissions from more than 90 countries.

Implementation

The basic data flow of the SCRATCH system is portrayed in Figure 1. The user submits the protein sequence through a WWW form. The form data are processed by a Perl script that produces a single Fasta file with the additional information of the user's predictor selections. The Perl script also adds the file to the end of the job queue. The job manager, running on a Sun front-end server, processes a single job by submitting it to a single machine in our Linux cluster and starts another script to run everything necessary to get the results requested by the user. The only software developed outside our laboratory currently used in the pipeline are PSI-BLAST and BLAST (7). PSI-BLAST is used to generate a multiple sequence alignment and consensus sequence from the target sequence. BLAST is used to identify homologs with high-sequence identity in the PDB to improve SSpro and ACCpro predictions.

Appropriate usage

The SCRATCH predictors can be applied to any amino acid sequence; however, in terms of performance versus other methods they are most appropriate to use on targets without high levels of sequence homology to one or more solved structures. There are some exceptions to this caveat. SSpro and ACCpro may be used effectively on any protein sequence because they use template structure information in their predictions when appropriate. MUpro and DISpro are also appropriate to use on any protein sequence.

Structural features prediction performance

The three class per-amino acid accuracy (Q3) of SSpro is ∼77% (2). SSpro version 4.0 has been extensively evaluated on EVA and has been consistently ranked as one of the top secondary structure prediction servers (17). The accuracy of ACCpro is ∼77% at the 25% exposure threshold (3). The prediction accuracies are based on targets where template homology is not used directly, and both systems perform better when template homology can be applied. The eight class per-amino acid accuracy (Q8) accuracy of SSpro8 is ∼63% (2). The accuracy of CONpro is ∼72% (3). DOMpro predicts the correct number of domains ∼69% of the time. The precision and recall of disordered regions of DISpro are 75.4 and 38.8%, respectively. For DIpro, the prediction accuracy of cysteine bonding states is ∼87% (4). The average disulfide bond prediction accuracy of DIpro is 53% (4). The accuracy of mutation stability prediction is ∼86%. On a test set of proteins with length <100 CMAPpro predicted contacts with 49% accuracy and non-contacts with 96% accuracy (6).

Tertiary structure prediction performance

Our current version of 3Dpro is an *ab initio* predictor only. For targets with significant homology to one or more solved structures, comparative modeling (CM) methods consistently produce more reliable models than *ab initio* methods. 3Dpro is most appropriate to use with targets do not have good structural templates. Tertiary structure prediction methods are evaluated by the Critical Assessment of Structure Prediction (CASP) experiments held every two years (18). CASP evaluates predictions in three broad categories: CM, fold recognition (FR) and new fold (NF). The easiest targets to predict are categorized as CM-easy, while the hardest are categorized as NF. There is a continuous spectrum of difficulty and these categories blur at the edges as do the methods that work best on different types of targets. We took part in the most recent experiment, CASP6. For complete results see http://predictioncenter.llnl.gov/. Our tertiary structure predictor 'baldi-group-server' performed well on hard targets (those in the NF and more difficult targets in the FR category) compared with other fully automated predictors. For summarized results for server groups on hard targets follow the link from the SCRATCH home page.

Calculation times

The structural features email is returned in several minutes for most sequences. The contact map prediction is the most time consuming of the structural feature predictions and the time increases quadratically with the length of the sequence. If a contact map is requested for longer sequences, then the structural features email will take a few more minutes to be returned. The tertiary structure email will be returned in <1 h if the sequence is short (length <125). The prediction time increases significantly as the sequence length increases; for this reason we only accept sequences up to length 400 for contact map and tertiary structure prediction through the web interface. The maximum length sequence for the other predictors is 1500. If the user has knowledge of the domains, then it is appropriate to submit each domain separately. We can accommodate longer sequences as well as high-throughput projects with off-line runs.

FUTURE WORK

We are currently developing new methods to increase the utility of SCRATCH. One important addition in development is a CM component, which we will combine with 3Dpro to make a more comprehensive tertiary structure predictor. Another new method in development predicts β-sheet pairings, which will be used in 3Dpro and made available as a standalone predictor.

Our group is also working to improve current methods. The next version of DOMpro will incorporate homology directly, and then we will use DOMpro to automatically break submissions into domains for 3Dpro to predict independently and then combine. We are improving our tertiary structure predictor directly by changing the physical representation from a carbon alpha only representation to an all backbone and side-chain centroid representation, and by adding energy terms for realistic secondary structure packing.

ACKNOWLEDGEMENTS

Work supported by 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, and by the Institute for Genomics and Bioinformatics at UCI. Funding to pay the Open Access publication charges for this article was provided by NIH Biomedical Informatics Training grant (LM-07443-01).

Conflict of interest statement. None declared.

REFERENCES

1. 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 Res*., **28**, 235–242.
2. Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.
3. Pollastri,G., Baldi,P., Fariselli,P. and Casadio,R. (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, 142–153.
4. Baldi,P., Cheng,J. and Vullo,A. (2004) Large-scale prediction of disulphide bond connectivity. *Adv. Neural Inf. Process Syst*., **17**, 97–104.
5. 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.
6. Baldi,P.F. and Pollastri,G. (2003) The principled design of large-scale recursive neural network architectures–DAG-RNNs and the protein structure prediction problem. *J. Mach. Learn. Res*., **4**, 575–602.
7. 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 Res*., **25**, 3389–3402.
8. Pemberton,S.G. and Jones,B. (1999) Ichnology of the pleistocene ironshore formation, Grand Cayman Island, British West-Indies. *J. Paleontol*., **62**, 495.

9. Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins*, **46**, 197–205.
10. Pearl,F.M., Lee,D., Bray,J.E., Sillitoe,I., Todd,A.E., Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282.
11. Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.
12. Simons,K.T., Ruczinski,I., Kooperberg,C., Fox,B.A., Bystroff,C. and Baker,D. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, **34**, 82–95.
13. Vendruscolo,M., Kussell,E. and Domany,E. (1997) Recovery of protein structure from contact maps. *Fold Des.*, **2**, 295–306.
14. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
15. Holm,L. and Sander,C. (1991) Database algorithm for generating protein backbone and side-chain co-ordinates from the C alpha trace application to model building and detection of co-ordinate errors. *J. Mol. Biol.*, **218**, 183–194.
16. Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L.,Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.*, **12**, 2001–2014.
17. Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Fiser,A., Pazos,F., Valencia,A., Sali,A. and Rost,B. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, **17**, 1242–1243.
18. Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **53** (Suppl. 6), 334–339.