iFC^2: an integrated web-server for the improved prediction of protein fold type, structural class, and secondary structure content

Ke Chen
University of Alberta
Wojciech Stach
University of Alberta
Leila Homaeian
University of Alberta
Lukasz Kurgan
University of Alberta

Method Article

Keywords: protein structure, protein structure prediction, fold type, secondary structure, structural class, fold recognition, secondary structure content

Posted Date: August 13th, 2008

DOI: https://doi.org/10.1038/nprot.2008.162

Abstract

Introduction

Recent research resulted in the development of several 1D protein structure descriptors. They provide an important alternative for analysis/prediction of the protein structure/function. Numerous computational methods that provide accurate prediction of these descriptors from the protein sequence were proposed; they include secondary structure[1-5], secondary structure content[6-9], structural class[10-17], fold type[18-25], relative solvent accessibility[26-32], contact order and number[33-37], and residue depth[38]. Recent work shows that the tertiary structure can be recovered from three 1D descriptors[39]. We developed a server that integrates predictions of several related descriptors including structural class[17], fold type[23], and secondary structure content[9]. The knowledge of these three descriptors was applied in various areas including tertiary structure prediction[40], identification of domain boundaries[41], analysis of protein interactions[42] and prion proteins[43], discrimination of outer membrane proteins[44,45], and in prediction of secondary structure[46], coding and noncoding RNAs[47], folding and unfolding rates[48-52], folding transition-state position[53], DNA-binding sites[54], and enzyme proteins and their class[55,56]. Our server, _iFC^2_ \(Integrated prediction of Fold, Class, and Content), is the first to exploit relations between the three descriptors, which are used to develop a cross-evaluation procedure that improves their predictions. _iFC^2_ predictions provide higher quality than the predictions of the individual methods. The server is located at http://biomine.ece.ualberta.ca/1D/1D.html.

Reagents

A single or multiple \(up to 10) protein sequences to be predicted should be provided in FASTA format.

Equipment

The user needs a computer with access to Internet and a Web browser.

Procedure

1| Enter query sequences. Prepare one or several protein sequences for prediction. The _iFC^2_ server accepts at most 10 protein sequences each time. The input sequences should be in FASTA format. **2**| Use the prediction server. To use _iFC^2_ server, access the prediction page at http://biomine.ece.ualberta.ca/1D/1D.html. Enter the protein sequences into the “Enter protein sequence\(s)” box. The sequence has to be provided in FASTA format and the user is allowed to enter up to 10 sequences at the time. The prediction will be executed sequentially and automatically for all entered sequences. The “Example” button fills the box with an example of FASTA formatted sequence. The “Reset” button clears the contents of the box. **3**| Choose prediction task. There are 4 options, see Figure 1. The user can either perform single prediction task, i.e., secondary structure content prediction

with PSSC-core[9], structural class prediction with SCEC[17], and fold type prediction with PFRES[23], or \(s)he can use the integrated _iFC[2]_ server \(by pressing on the 'all methods' button), which predicts the three targets at the same time. If _iFC[2]_ server is chosen, the cross-evaluation will be performed automatically. After choosing the prediction task, the user should press "Start" button. In the case when sequences entered in the "Enter protein sequence\(s)" box do not adhere to the FASTA format, an error window that describes the problem will be displayed and the user will be asked to correct the formatting. **4**| Obtain the results. After the prediction is done, the user can access the prediction results by pressing the "Show Results" button, or download the results in a comma-separated text format by pressing the "Download CSV File" button. **5**| Interpret the results. The results are displayed using a web page, see Figure 2. The page displays \(from top to bottom) the input sequence, the secondary structure predicted with PSI-PRED[1], the fold type predicted by PFRES[23], the structural class predicted by SCEC[17], the secondary structure contents predicted with PSSC-core[9], and the cross-evaluation results. For the fold type prediction, the output is one the 26 fold types described in [23]. For the structural class prediction, the output is one of the four structural classes \(all-□, all-□, □/□, and □+□). In the case of the secondary structure content prediction, the output is two real values which represent the helix and the strand contents, respectively. The cross-evaluation results include the secondary structure contents of helix and strand re-predicted by _iFC[2]_ server \(which is potentially different from the predictions of PSSC-core[9]), the output label provided by _iFC[2]_ server which flags the prediction of SCEC as potentially correct or incorrect, and the label generated by _iFC[2]_ server that annotates the prediction of PFRES as potentially correct or incorrect.

Timing

The computational time depends on the length of the sequence. Execution of PSSC-core[9] \(for secondary structure content prediction) takes about 10s for a protein sequence consisting of 200 amino acids. Average time to run SCEC[17] \(for structural class prediction), PFRES[23] \(for fold type prediction), and _iFC[2]_ for a sequence of about 200 amino acids is about 2mins for each method.

Troubleshooting

If the server does not accept the input protein sequence for prediction, the error might be caused by one of the following reasons: \(1) Input sequence\(s) is not in the FASTA format. \(2) Input sequence\(s) is less than 30 AAs and such sequence is considered to be too short to constitute a complete protein domain. \(3) The input sequence\(s) contains invalid characters. The valid single-letter characters for a protein sequence are ACDEFGHIKLMNPQRSTVWY. \(4) More than 10 sequences were entered.

Anticipated Results

The quality of predictions of PSSC-core[9], SCEC[17], and PFRES[23] is evaluated and discussed in the corresponding publications. Independent tests of the cross-evaluation procedure of _iFC[2]_ server show

that: \(1) The MAE \(mean absolute error) of helix and strand content predicted by _iFC^2_ server equal 0.085 and 0.049, respectively. The PCC \(Pearson correlation coefficient) values equal 0.94 for the helix content prediction and 0.89 for the strand content prediction. \(2) _iFC^2_ server assigns “correct” labels for 79.3% of predictions made by SCEC[17]. Among these “correct” predictions, the accuracy of SCEC equals 98.2%, while the accuracy of SCEC for the predictions deemed as “incorrect” by _iFC^2_ server equals 14.6%. \(3) _iFC^2_ server labels 81.8% of the PFRES[23] predictions as “correct” and the accuracy of these predictions equals 71.8%. At the same time, the accuracy of predictions performed with PFRES for the sequences predicted by _iFC^2_ server as “incorrect” equals 38.5%.

References

1. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. _J Mol Biol._ 292, 195-202 \(1999). 2. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. _Bioinformatics._ 16, 404-5 \(2000). 3. Rost B, Yachdav G, Liu J. The PredictProtein server. _Nucleic Acids Res._ 32, W321-6 \(2004). 4. Cole C, Barber JD, Barton GJ. The Jpred 3 secondary structure prediction server. _Nucleic Acids Res._ \(2008). 5. Kurgan L. On the relation between the predicted secondary structure and the protein size. _Protein J._ 27, 234-9 \(2008). 6. Cai YD, Liu XJ, Chou KC. Prediction of protein secondary structure content by artificial neural network. _J Comput Chem._ 24, 727-31 \(2003). 7. Ruan J, Wang K, Yang J, Kurgan L, Cios KJ. Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. _Artif. Intel. Med._ 35, 19-35 \(2005). 8. Lee S, Lee BC, Kim D. Prediction of protein secondary structure content using amino acid composition and evolutionary information. _Proteins._ 62, 1107-14 \(2006). 9. Homaeian L, Kurgan LA, Ruan J, Cios KJ, Chen K. Prediction of protein secondary structure content for the twilight zone sequences. _Proteins._ 69, 486-98 \(2007). 10. Chou KC, Cai YD. Predicting protein structural class by functional domain composition. _Biochem Biophys Res Commun._ 321, 1007-9 \(2004). 11. Chou KC. Progress in protein structural class prediction and its impact to bioinformatics and proteomics. _Curr Protein Pept Sci._ 6, 423-36 \(2005). 12. Xiao X, Shao SH, Huang ZD, Chou KC. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. _J Comput Chem._ 27, 478-82 \(2006). 13. Kedarisetti KD, Kurgan L, Dick S. Classifier ensembles for protein structural class prediction with varying homology. _Biochem Biophys Res Commun._ 348, 981-8 \(2006). 14. Kurgan L, Chen K. Prediction of protein structural class for the twilight zone sequences. _Biochem Biophys Res Commun._ 357, 453-60 \(2007). 15. Kurgan L, Cios K, Chen K. SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. _BMC Bioinformatics._ 9, 226 \(2008). 16. Xiao X, Lin WZ, Chou KC. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. _J Comput Chem._ 29, 2018-24 \(2008). 17. Chen K, Kurgan LA, Ruan J. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. _J Comput Chem._ 29, 1596-604 \(2008). 18. Ding CH, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. _Bioinformatics._ 17, 349-58 \(2001). 19. Shen HB, Chou KC. Ensemble classifier for protein fold pattern recognition. _Bioinformatics._ 22, 1717-22 \(2006). 20. Jeong J, Berman P, Przytycka T. Fold classification based on

secondary structure--how much is gained by including loop topology? _BMC Struct Biol._ 6, 3 \(2006). 21. Taguchi Y, Gromiha M. Application of amino acid occurrence for discriminating different folding types of globular proteins. _BMC Bioinformatics._ 8, 404 \(2007). 22. Melvin I, Ie E, Kuang R, Weston J, Stafford WN, Leslie C. SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition. _BMC Bioinformatics._ 8, Suppl 4:S2 \(2007). 23. Chen K, Kurgan L. PFRES: protein fold classification by using evolutionary information and predicted secondary structure. _Bioinformatics._ 23, 2843-50 \(2007). 24. Shamim MT, Anwaruddin M, Nagarajaram HA. Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. _Bioinformatics._ 23, 3320-7 \(2007). 25. Damoulas T, Girolami MA. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. _Bioinformatics._ 24, 1264-70 \(2008). 26. Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. _Bioinformatics._ 18, 819–824 \(2002). 27. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. _Proteins._ 50, 629-35 \(2003). 28. Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. _Proteins._ 54, 557–562 \(2004). 29. Garg A, Kaur H, Raghava GP. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. _Proteins._ 61, 318-24 \(2005). 30. Nguyen MN, Rajapakse JC. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. _Proteins._ 63, 542-50 \(2006). 31. Dor O, Zhou Y Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. _Proteins._ 68, 76-81 \(2007). 32. Chen K, Kurgan M, Kurgan L. Sequence Based Prediction of Relative Solvent Accessibility Using Two-stage Support Vector Regression with Confidence Values. _J. Biom. Science and Eng._ 1, 1-9 \(2008). 33. Pollastri G, Baldi P, Fariselli P, Casadio R. Improved prediction of the number of residue contacts in proteins by recurrent neural networks. _Bioinformatics._ Suppl 1, S234-42 \(2001). 34. Kinjo AR, Horimoto K, Nishikawa K. Predicting absolute contact numbers of native protein structure from amino acid sequence. _Proteins._ 58, 158-65 \(2005). 35. Kinjo AR, Nishikawa K. Predicting secondary structures, contact numbers, and residue-wise contact orders of native protein structures from amino acid sequences using critical random networks. _Biophysics._ 1, 67-74. \(2005). 36. Yuan Z. Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. _BMC Bioinformatics._ 6, 248 \(2005). 37. Song JN, Burrage K. Predicting residue-wise contact orders in proteins by support vector regression. _BMC Bioinformatics._ 7: 425 \(2006). 38. Yuan Z, Wang Z-X. Quantifying the relationship of protein burying depth and sequence. _Proteins._ 70, 509–516 \(2008). 39. Kinjo AR, Nishikawa K. Recoverable one-dimensional encoding of protein three-dimensional structures. _Bioinformatics._ 21, 2167-70 \(2005). 40. Bahar I, Atilgan AR, Jernigan RL, Erman B. Understanding the recognition of protein structural classes by amino acid composition. _Proteins_ 29, 172-185 \(1997). 41. Redfern OC, Harrison A, Dallman T, Pearl FM, Orengo CA. CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. _PLoS Comput Biol._ 3, e232 \(2007). 42. Smith J, Diez G, Klemm AH, Schewkunow V, Goldmann WH. CapZ-lipid membrane interactions: A computer analysis. _Theo. Bio. Med. Model._ 3, 33-7 \(2006). 43. Concepcion GP, David MP, Padlan EA. Why don't humans get scrapie from eating sheep? A possible explanation based on secondary structure predictions. _Med Hypotheses_ 64,

919-24 \(2005). 44. Gromiha M. Motifs in outer membrane protein sequences: applications for discrimination. _Biophys Chem._ 117, 65-71 \(2005). 45. Gromiha M, Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy. _Bioinformatics_ 21, 961-8 \(2005). 46. Gromiha M, Selvaraj S. Protein secondary structure prediction in different structural classes. _Protein Eng._ 11, 249-251 \(1998). 47. Liu J, Gough J, Rost B. Distinguishing protein-coding from non-coding RNAs through support vector machines. _PLoS Genet_ 2, 529-36 \(2006). 48. Gong H, Isom DG, Srinivasan R, Rose GD. Local secondary structure content predicts folding rates for simple, two-state proteins. _J Mol Biol._ 327, 1149-54 \(2003). 49. Gromiha M, Selvaraj S. Bioinformatics approaches for understanding and predicting protein folding rates. _Cur. Bioinformatics_ 3, 1-9 \(2008). 50. Ivankov DN, Finkelstein AV. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. _Proc Natl Acad Sci USA_ 101, 8942-4 \(2004). 51. Gromiha M. A statistical model for predicting protein folding rates from amino acid sequence with structural class information. _J. Chem Inf Model._ 45, 494-501 \(2005). 52. Gromiha M, Selvaraj S, Thangakani AM. A Statistical method for predicting protein unfolding rates from amino acid sequence, _J Chem Inf Model_ 46, 1503-1508 \(2006). 53. Huang JT, Cheng JP. Prediction of folding transition-state position \(T) of small, two-state proteins from local secondary structure content. _Proteins_ 68, 218-22 \(2007). 54. Kuznetsov IB, Gou Z, Li R and Hwang S. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. _Proteins_ 64, 19-27 \(2006). 55. Dobson PD, Doig AJ. Predicting enzyme class from protein structure without alignments. _J Mol Biol._ 345, 187-99 \(2005). 56. Dobson PD, Doig AJ. Distinguishing enzyme structures from non-enzymes without alignments. _J Mol Biol._ 330, 771-83 \(2003).

Acknowledgements

This work was supported in part by iCORE, Alberta Ingenuity Fund, and NSERC \(Natural Sciences and Engineering Research Council of Canada).

Figures

Figure 1

The interface for accessing the _iFC^2^_ server. The web page is located at "http://biomine.ece.ualberta.ca/1D/1D.html":http://biomine.ece.ualberta.ca/. The full size version of this figure can be found "here":http://protocols.nature.com/image/show/1057.

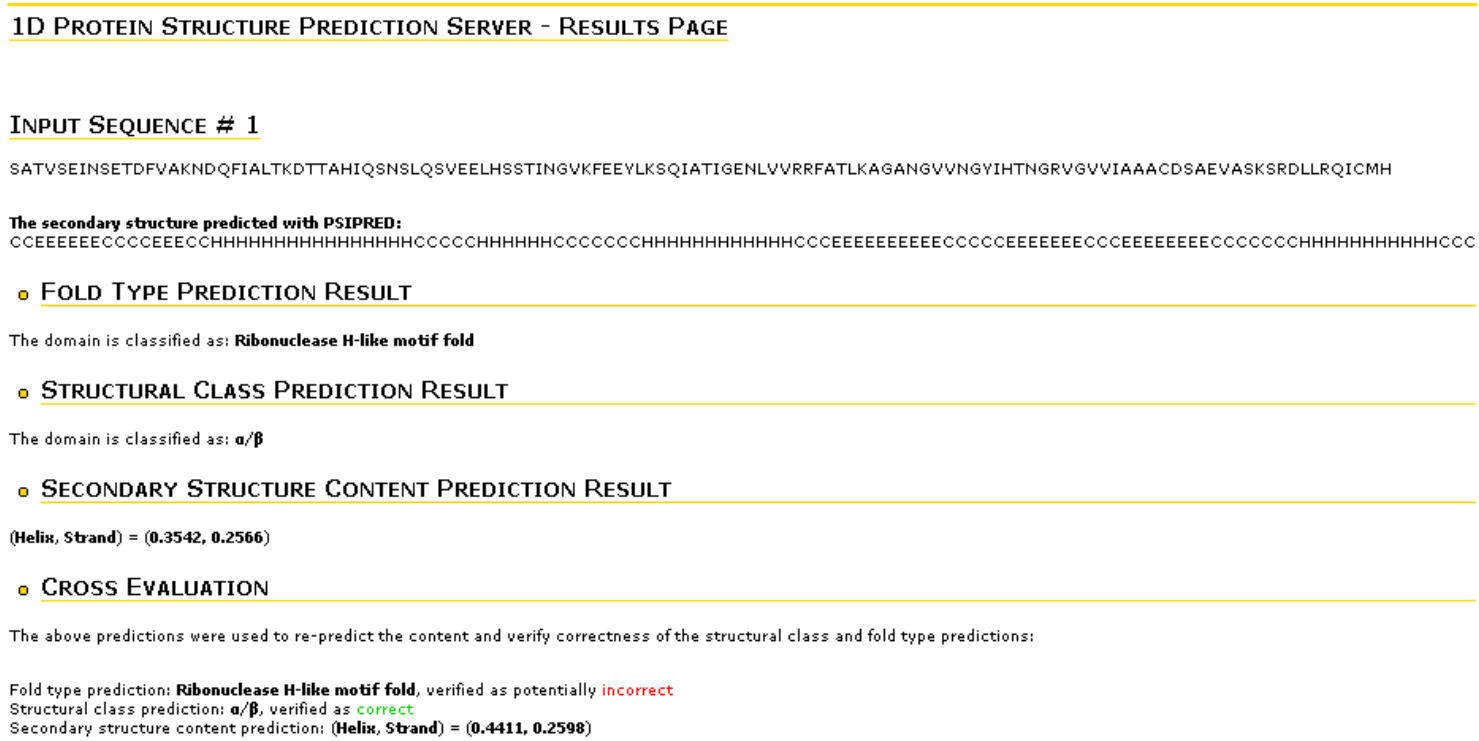

Figure 2

Example prediction results computed with _iFC^2^_ server. The full size version of this figure can be found "here":http://protocols.nature.com/image/show/1058.