Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Web Server**

FlexPred: a web-server for predicting residue positions involved in conformational switches in proteins

Igor B. Kuznetsov[1,*] and Michael McDuffie[1]

[1]Gen*NY*sis Center for Excellence in Cancer Genomics, Department of Epidemiology and Biostatistics, One Discovery Drive, University at Albany, Rensselaer, NY 12144, USA; Igor B. Kuznetsov* - E-mail: ikuznetsov@albany.edu; * Corresponding author

received October 30, 2008; accepted November 01, 2008; published November 05, 2008

Abstract:
Conformational switches observed in the protein backbone play a key role in a variety of fundamental biological activities. This paper describes a web-server that implements a pattern recognition algorithm trained on the examples from the Database of Macromolecular Movements to predict residue positions involved in conformational switches. Prediction can be performed at an adjustable false positive rate using a user-supplied protein sequence in FASTA format or a structure in a Protein Data Bank (PDB) file. If a protein sequence is submitted, then the web-server uses sequence-derived information only (such as evolutionary conservation of residue positions). If a PDB file is submitted, then the web-server uses sequence-derived information and residue solvent accessibility calculated from this file.

Availability: FlexPred is publicly available at http://flexpred.rit.albany.edu

Keywords: conformational variability; support vector machine; protein flexibility; structural transition; prediction

Background:
Proteins are flexible macromolecules. The protein backbone can switch from one specific folded conformation to another. Conformational switches have been shown to be involved in a variety of biological functions, such as catalysis, macromolecular recognition, signal transduction, locomotion, and a number of pathogenic disorders **[1-2]**. Molecular dynamics simulations of long time scale conformational transitions are very computationally expensive and therefore impractical for large-scale studies **[3]**. Several bioinformatics methods that attempt to predict conformational switches from sequence information alone have been developed. Most of these methods were not trained to predict conformational switches directly by using a dataset of experimental examples of such switches, but rather to identify them indirectly by predicting certain structural properties related to protein flexibility in general. Flexibility-related properties used to train these methods include the crystallographic B-factor **[4-5]**, the ambiguity in secondary structure assignment **[6-7]**, and the magnitude of large-scale fluctuations obtained from coarse grained protein dynamics modeling **[8]**. A dataset of experimental examples of flexible linkers connecting structurally rigid domains was used directly to develop a sequence-based method for predicting hinge points **[2]**.

Recently, we used a dataset of experimentally characterized conformational switches to develop predictors of flexible residue positions and studied the performance of these predictors. The results of this study showed that positions involved in conformational switches can be predicted with balanced sensitivity and specificity for all types of secondary structure and all types of protein movements **[9]**. Here, we use predictive models from our previous work to develop FlexPred, a web-server that uses a protein sequence alone or in combination with solvent accessibility to predict residue positions involved in conformational switches. A detailed description of the training and testing methods was provided in our original paper **[9]**. We therefore only briefly describe the methodology here. We used a non-redundant dataset obtained from the Database of Macromolecular Movements that contains examples of conformational switches derived by comparing experimental atomic-level structures of the same protein solved under different conditions **[10]**. This dataset was used to train a supervised pattern recognition method, Support Vector Machine (SVM), to distinguish between flexible and rigid residue positions. We implemented two types of encoding of the input sequence. One is the binary encoding which utilizes the input sequence alone and represents the 20 amino acid types as 20 mutually orthogonal binary vectors. The other is the PSSM encoding which accounts for evolutionary conservation of the input sequence and is based on the PSI-BLAST position-specific scoring matrix (PSSM). If the user submits a protein structure in a

ISSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(2): 134-136 (2008)

Protein Data Bank (PDB) file, then the normalized residue solvent accessibility calculated from this file is also used for prediction along with one of the two types of sequence encoding. Thus, we have four possible ways of encoding protein sequence with or without solvent accessibility. Accordingly, four SVM predictors, one for each of the four combinations, were implemented.

Input:

FlexPred is freely available at http://flexpred.rit.albany.edu. It has a simple intuitive user interface that consists of four input fields described below. Instructions for each field and general information about the methodology and the output format can be found by clicking a corresponding help hyperlink on the input page.

Field 1: ***Protein sequence or PDB file to be analyzed.*** For sequence-based prediction, the user can paste or upload an amino acid sequence in FASTA format. For the prediction based on a protein sequence and solvent accessibility of its residue positions, the user can either upload a PDB file or provide a four-character PDB id and let the server automatically download a corresponding file from ftp://ftp.wwpdb.org.

Field 2: ***Selection of encoding method.*** The user can select either binary or PSSM sequence encoding. The PSSM encoding performs better if protein sequence information alone is used for prediction, whereas the binary encoding performs better if both protein sequence and residue solvent accessibility are used. Therefore, the PSSM encoding is the default method for the sequence-based submissions, while the binary encoding is the default method for the PDB-based submissions.

Field 3: ***Selection of false positive rate.*** The false positive rate (FPr) gives the fraction of rigid positions incorrectly predicted as flexible, whereas the true positive rate (TPr) gives the fraction of flexible positions correctly predicted as flexible. For any prediction method, when FPr is decreased, TPr is also decreased, and *vice versa*. The user can choose FPr of 5%, 10%, 15%, or 20%. Since most statistical tests consider the 5% chance of false positive prediction to be an acceptable level, the FPr of 5% is selected by default.

Field 4: ***Selection of retrieval method***. The user can choose to receive results by E-mail (default) or manually retrieve them using a temporary URL provided upon submission. The results of prediction are kept on the web-server for one day from the moment of submission, and deleted afterwards.

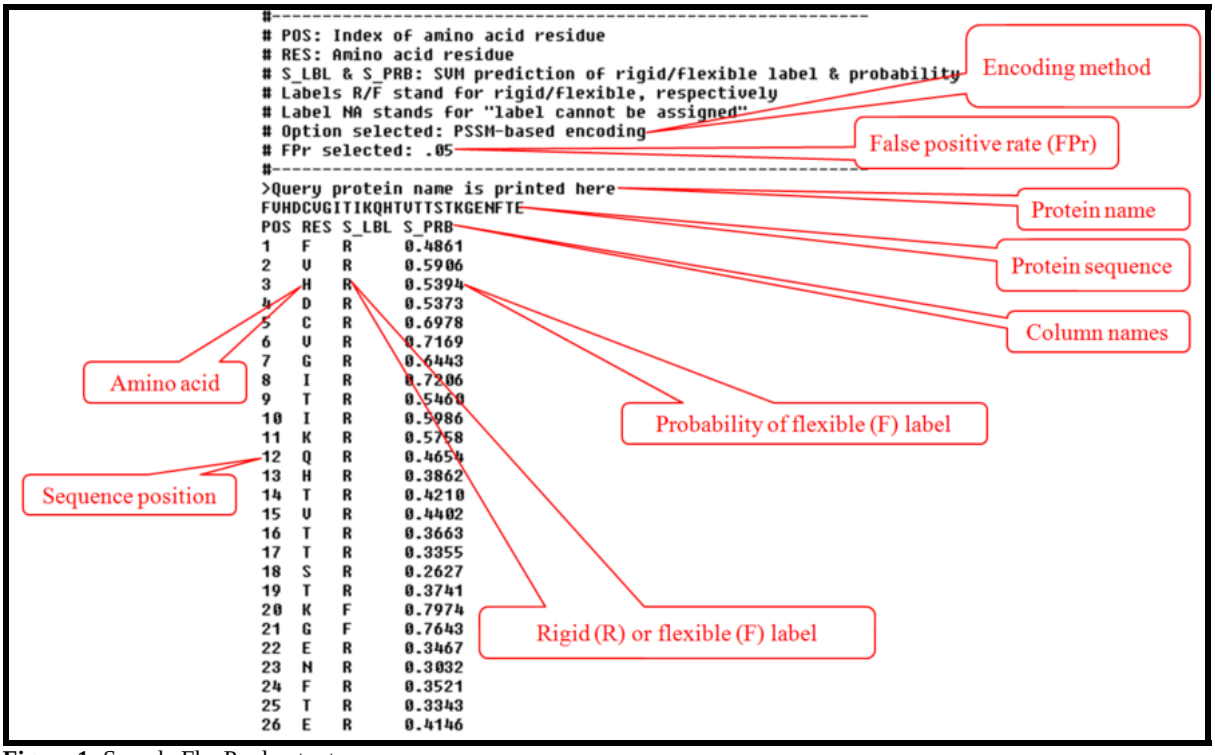

Figure 1: Sample FlexPred output.

ISSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(2): 134-136 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Web Server**

Output:

The output from FlexPred consists of a header that describes the output format itself, the selected encoding type, the selected false positive rate, the submitted sequence in FASTA format, and the predicted labels for each residue position (Figure 1). The labels are 'R' (rigid) and 'F' (flexible). The column 'S_PRB' shows the probability of label 'F' for each residue position. The probability of label 'R' is (1-P_F), where P_F is the probability of label 'F'. The probabilities are in range [0.0, 1.0]. Higher probability corresponds to a greater prediction confidence.

Future development:

To the best of the authors' knowledge, FlexPred is the only on-line method for predicting conformational switches in proteins directly trained on a large dataset of experimentally characterized examples. We will continue updating FlexPred by adding new predictive models and new experimental examples as they become available.

Acknowledgement:

This work was supported by grant number R03LM009034 from the National Institutes of Health.

References:

[01] M. Gross, *Curr. Protein Pept. Sci.,* 1: 339 (2000) [PMID: 12369904]

[02] S. C. Flores *et al.*, *BMC Bioinformatics,* 8: 167 (2007) [PMID: 17519025]

[03] L. Yang *et al.*, *Biophys. J.*, 93: 920 (2007) [PMID: 17483178]

[04] Z. Yuan *et al.*, *Proteins*, 58: 905 (2005) [PMID: 15645415]

[05] Schlessinger *et al.*, *Bioinformatics*, 22: 891 (2006) [PMID: 16455751]

[06] M. Young *et al.*, *Protein Sci.*, 8: 1752 (1999) [PMID: 10493576]

[07] M. Bodén and T.L. Bailey, *Bioinformatics*, 22: 1809 (2006) [PMID: 16720586]

[08] J. Gu *et al.*, *PLoS Comput. Biol.*, 2: e90 (2006) [PMID: 16839194]

[09] I. B. Kuznetsov, *Proteins*, 72: 74 (2008) [PMID: 18186479]

[10] S. C. Flores *et al.*, *Nucleic Acids Res.*, 34: D296 (2006) [PMID: 16381870]

Edited by P. Kangueane

Citation: **Kuznetsov & McDuffie,** Bioinformation 3(3): 134-136 (2008)

ISSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(2): 134-136 (2008)