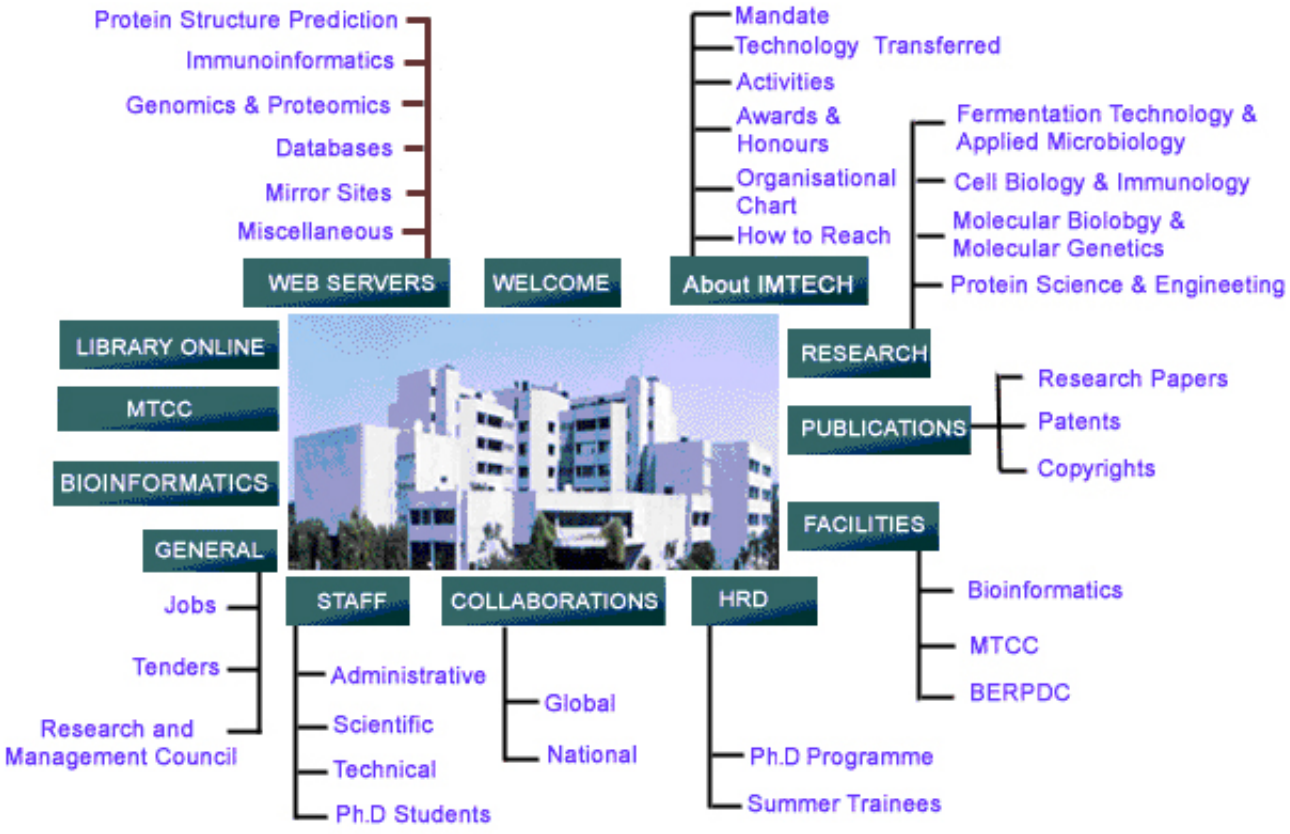
Webservers and Databases for Vaccine and Drug Design

GPS Raghav

Bioinformatics Center

Institute of Microbial Technology, Chandigarh

Email: raghava@imtech.res.in



Vaccine Informatics	
Description of Server	**Web Site**
1. **Propred:** Prediction of Promiscuous HLA-DR binders	*$R/propred/
2. **Propred1:** Prediction of Promiscuous MHC-I binders	*$R/propred1/
3. **nHLAPred:** Predicting promiscuous MHC-I binders using ANN	*$R/nhlapred/
4. **CTLPred:** Prediction of CTL epitopes using QM, SVM and ANN	*$R/ctlpred/
5. **HLA-DR4Pred:** Predicting HLA-DRB1*0401 binders using SVM	*$R/hladr4pred/
6. **MMBpred:** Promiscuous and high-affinity mutated MHC binders.	*$R/mmbpred/

7. **TAPPred:** Prediction of affinity of TAP binders using cascade SVM.	*$R/tappred/
8. **Pcleavage:** Prediction of proteasome cleavage sites in antigen	*$R/pcleavage/
9. **BcePred:** B-Cell epitopes using Physico-chemical properties	*$R/bcepred/
10. **ABCPred:** Prediction of B-cell epitopes using neural network.	*$R/abcpred/
11. **AlgPred:** Prediction of allergens and mapping of IgE epitopes.	*$R/algpred/
Drug Informatics	
Protein Structure Prediction	
12. **APSSP2:** Protein secondary structure prediction using KNN and ANN.	*$R/apssp2/
13. **PepBuild:** Building peptides/proteins with desired structure	*$B/pepbuilt/
14. **ProClass:** Classification of proteins based on secondary structure.	*$R/proclass/
15. **SARpred:** Solvent accessibility of amino acids in proteins.	*$R/sarpred/
16. **TBBpred:** Trans-membrane regions in beta-barrel proteins	*$R/tbbpred/
17. **BhairPred:** Prediction of beta-hairpins in a protein.	*$R/bhairpred/
18. **ArNHPred:** Aromatic-backbone NH interactions in proteins.	*$R/arnhpred/
19. **CHpredict:** Prediction of Cα-H...O and Cα-H...pi interactions.	*$R/chpredict/
20. **BTEVAL:** A server for evaluation of β-turn prediction methods	*$R/bteval/
21. **BetaTPred:** β-TURNS in a protein using statistical algorithms.	*$R/betatpred/
22. **BetatPred2:** Highly accurate method for prediction of β-turns.	*$R/betapred2/
23. **Betaturns:** Prediction of β-turn types in proteins.	*$R/betaturns/
24. **AlphaPred:** α-turns in proteins using PSI-BLAST profiles	*$R/alphapred/
25. **GammaPred:** γ-turns in proteins using multiple sequence alignment	*$R/gammapred/
26. **PepStr:** Prediction of tertiary structure small bioactive peptides.	*$R/pepstr/
Genome Annotation	
27. **SRF:** Identification of spectral repeats using Fourier transformation	*$R/srf/
28. **FTG:** Locating probable prokaryotic genes using FFT	*$R/ftg/
29. **EGPred:** Prediction of eukaryotic genes and their structure.	*$R/egpred/
30. **GWFASTA:** Genome wide FASTA search	*$R/gwfasta/
Subcellular localization	
31. **ESLpred**: Subcellular localization of eukaryotic proteins	*$R/eslpred/
32. **PSLpred:** Prediction of subcellular localization of bacterial proteins.	*$R/pslpred/
33. **HSLPred:** Subcellular localization of human proteins	*$R/hslpred/
34. **TBpred:** Predicting subcellular localization of mycobacterial proteins	*$R/tbpred/
35. **MitPred:** Prediction of mitochondrial proteins using SVM and HMM	*$R/mitpred/
Functional annotation of Proteomes	
36. **NRpred:** Classification of nuclear receptors based composition	*$R/nrpred/
37. **GPCRpred:** Families and subfamilies of G-protein coupled receptors.	*$R/gpcrpred/
38. **GPCRsclass:** Classification of amine type of GPCR.	*$R/gpcrsclass/
39. **BTXpred:** Prediction of bacterial toxins.	*$R/btxpred/
40. **NTXpred:** Neurotoxins based on their function and source.	*$R/ntxpred/
41. **VGIchan:** Prediction and classification of voltage-gated ion channels.	*$R/vgichan/
42. **VICMpred:** Functional annotation of Gram-negative bacteria	*$R/vicmpred/
43. **Oxypred:** Prediction and classification of oxygen binding proteins	*$R/oxypred/
44. **GSTPred:** Prediction of glutathione S-transferase proteins.	*$R/gstpred/
45. **Pprint:** Prediction of RNA-binding sites in a protein	*$R/pprint/
46. **MANGO:** Prediction GO class of a protein from its amino acid.	*$R/mango/
Databases and Miscellaneous web-servers/software	
47. **MHCBN:** A database of MHC/TAP binders and T-cell epitopes	*$R/mhcbn/
48. **Bcipep:** a database of B-cell epitopes.	*$R/bcipep/
49. **HaptenDB:** A database of haptens, carrier-proteins and antibodies.	*$R/haptendb/
50. **AntiBP:** Analysis and prediction of antibacterial peptides.	*$R/antibp/
51. **RBpred:** Forecasting of rice blast disease from weather conditions	*$R/rbpred/
52. **LGEpred:** Correlation and prediction of gene expression level	*$R/lgepred/

53. **OXBench:** Benchmarking of methods of multiple sequence alignment	*$R/oxbench/
54. **PSAweb:** Analysis of protein sequences and alignments	*$R/psa/
55. **AbAg:** Computing Titer and Concentration of Antibody/Antigen.	*$R/abag/
56. **DNA OPT:** Optimization of DNA gel electrophoresis and SDS-PAGE	*$RP/dnaopt/
57. **Hemo:** Measuring of hemolytic potency of drugs	*$RP/hemo/
58. **DNA SIZE:** Estimation of DNA fragment lengths from Gel	*$RP/dnasize/
59. **ELISA_eq:** Calculation of antibody and antigen	*$RP/elisaeq/
60. **IL4IFNG:** Measurement and computation of IL-4 and IFN-γ	*$RP/il4ifng/
61. **Ab_affi:** Affinity of antibody using non-competitive ELISA	*$RP/abaffi/
62. **GMAP:** Mapping of restriction sites into DNA sequences.	*$RP/gmap/

***$R: http://www.imtech.res.in/raghava/ $B: http://www.imtech.res.in/bvs/ $RP: http://www.imtech.res.in/raghava/progs/**

1. ProPred: prediction of HLA-DR binding sites.

Abstract: ProPred is a graphical web tool for predicting MHC class II binding regions in antigenic protein sequences. The server implement matrix based prediction algorithm, employing amino-acid/position coefficient table deduced from literature. The predicted binders can be visualized either as peaks in graphical interface or as colored residues in HTML interface. This server might be a useful tool in locating the promiscuous binding regions that can bind to several HLA-DR alleles.

Web-server:http://www.imtech.res.in/raghava/propred/
Useful for: Prediction of promiscuous HLA-DR binding sites in an antigen.
Reference: *Bioinformatics.* 2001;**17**:1236-1237.

2. ProPred1: prediction of promiscuous MHC Class-I binding sites.

Abstract: ProPred1 is an on-line web tool for the prediction of peptide binding to MHC class-I alleles. This is a matrix-based method that allows the prediction of MHC binding sites in an antigenic sequence for 47 MHC class-I alleles. The server represents MHC binding regions within an antigenic sequence in user-friendly formats. These formats assist user in the identification of promiscuous MHC binders in an antigen sequence that can bind to large number of alleles. ProPred1 also allows the prediction of the standard proteasome and immunoproteasome cleavage sites in an antigenic sequence. This server allows identification of MHC binders, who have the cleavage site at the C terminus. The simultaneous prediction of MHC binders and proteasome cleavage sites in an antigenic sequence leads to the identification of potential T-cell epitopes.

Web-server:http://www.imtech.res.in/raghava/propred1/
http://bioinformatics.uams.edu/mirror/propred1/
Useful for: Prediction of promiscuous MHC class-I binders
Reference: *Bioinformatics.* 2003; **19**:1009-1014.

3. nHLAPred: A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes.

Abstract: In the present study, a systematic attempt has been made to develop an accurate method for predicting MHC class I restricted T cell epitopes for a large number of MHC class I alleles. Initially, a quantitative matrix (QM)-based method was developed for 47 MHC class I alleles having at least 15 binders. A secondary artificial neural network (ANN)-based method was developed for 30 out of 47 MHC alleles having a minimum of 40 binders. Combination of these ANN-and QM-based prediction methods for 30 alleles improved the accuracy of prediction by 6% compared to each individual method. Average accuracy of hybrid method for 30 MHC alleles is 92.8%. This method also allows prediction of binders for 20 additional alleles using QM that has been reported in the literature, thus allowing prediction for 67 MHC class I alleles. The performance of the method was evaluated using jack-knife validation test. The

performance of the methods was also evaluated on blind or independent data. Comparison of our method with existing MHC binder prediction methods for alleles studied by both methods shows that our method is superior to other existing methods. This method also identifies proteasomal cleavage sites in antigen sequences by implementing the matrices described earlier. Thus, the method that we discover allows the identification of MHC class I binders (peptides binding with many MHC alleles) having proteasomal cleavage site at C-terminus. The user-friendly result display format (HTML-II) can assist in locating the promiscuous MHC binding regions from antigen sequence. The method is available on the web at www.imtech.res.in/raghava/nhlapred and its mirror site is available at http://bioinformatics.uams.edu/mirror/nhlapred/.

Web-server: http://www.imtech.res.in/raghava/nhlapred/
Useful for: Prediction of MHC class I T-cell epitops.
Reference:*J Biosci.* 2007;**32**; 31-42.

4. CTLPred: Prediction of CTL epitopes using QM, SVM and ANN techniques.

Abstract: Cytotoxic T lymphocyte (CTL) epitopes are potential candidates for subunit vaccine design for various diseases. Most of the existing T cell epitope prediction methods are indirect methods that predict MHC class I binders instead of CTL epitopes. In this study, a systematic attempt has been made to develop a direct method for predicting CTL epitopes from an antigenic sequence. This method is based on quantitative matrix (QM) and machine learning techniques such as Support Vector Machine (SVM) and Artificial Neural Network (ANN). This method has been trained and tested on non-redundant dataset of T cell epitopes and non-epitopes that includes 1137 experimentally proven MHC class I restricted T cell epitopes. The accuracy of QM-, ANN- and SVM-based methods was 70.0, 72.2 and 75.2%, respectively. The performance of these methods has been evaluated through Leave One Out Cross-Validation (LOOCV) at a cutoff score where sensitivity and specificity was nearly equal. Finally, both machine-learning methods were used for consensus and combined prediction of CTL epitopes. The performances of these methods were evaluated on blind dataset where machine learning-based methods perform better than QM-based method. We also demonstrated through subgroup analysis that our methods can discriminate between T-cell epitopes and MHC binders (non-epitopes). In brief this method allows prediction of CTL epitopes using QM, SVM, ANN approaches. The method also facilitates prediction of MHC restriction in predicted T cell epitopes.

Web-server: http://www.imtech.res.in/raghava/ctlpred/
Useful for: Prediction of Cytotoxic T cell epitopes.
Reference: *Vaccine.* 2004; **22**:3195-3204.

5. HLA-DR4Pred: SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence.

Abstract: Prediction of peptides binding with MHC class II allele HLA-DRB1(*)0401 can effectively reduce the number of experiments required for identifying helper T cell epitopes. This paper describes support vector machine (SVM) based method developed for identifying HLA-DRB1(*)0401 binding peptides in an antigenic sequence. SVM was trained and tested on large and clean data set consisting of 567 binders and equal number of non-binders. The accuracy of the method was 86% when evaluated through 5-fold cross-validation technique.

Web-server:http://www.imtech.res.in/raghava/hladr4pred/
http://bioinformatics.uams.edu/mirror/hladr4pred/ (Mirror Site).
Useful for: Prediction of bacterial toxins.
Reference: *Bioinformatics.* 2004;**20**:421-423.

6. MMBpred: Prediction of promiscuous and high-affinity mutated MHC binders.

Abstract: The identification of peptides in an antigenic sequence that can bind with high affinity to a wide range of MHC alleles is one of the challenges in subunit vaccine design. The mutation of natural peptides is an alternative to obtaining peptides that can bind to a wide range of MHC alleles with high affinity. A large number of experiments are typically necessary to identify mutations that define high-affinity binding peptides. Therefore there is a need to develop a computational method for detecting amino acid mutations in a peptide for making it high-affinity or promiscuous MHC binders. This report describes a high-throughput computer driven solution for the identification of promiscuous and high-affinity mutated binders of 47 MHC class I alleles by introducing mutations in an antigenic sequence. The method implements quantitative matrices for creating optimal mutations in an antigenic sequence. It has two major options: (i) prediction of promiscuous MHC binders and (ii) prediction of high-affinity binders. In case of prediction of promiscuous binders, the server allows a user to select (i) permissible mutations in a peptide; (ii) MHC alleles to whom it should bind; and (iii) positions at which mutation is allowed. In the case of prediction of high-affinity binders, the server allows users to specify the positions that should be conserved in the native protein. In both cases, the method computes the type of mutations and position of mutations in 9-mer peptides required to have the desired results.

Web-server:http://www.imtech.res.in/raghava/mmbpred/
Useful for: Prediction of high affinity and promiscuous MHC binders.
Reference: *Hybrid Hybridomics.* 2003; **22**:229-234.

7. TAPPred: Analysis and prediction of affinity of TAP binding peptides using cascade SVM.

Abstract: The generation of cytotoxic T lymphocyte (CTL) epitopes from an antigenic sequence involves number of intracellular processes, including production of peptide fragments by proteasome and transport of peptides to endoplasmic reticulum through transporter associated with antigen processing (TAP). In this study, 409 peptides that bind to human TAP transporter with varying affinity were analyzed to explore the selectivity and specificity of TAP transporter. The abundance of each amino acid from P1 to P9 positions in high-, intermediate-, and low-affinity TAP binders were examined. The rules for predicting TAP binding regions in an antigenic sequence were derived from the above analysis. The quantitative matrix was generated on the basis of contribution of each position and residue in binding affinity. The correlation of r = 0.65 was obtained between experimentally determined and predicted binding affinity by using a quantitative matrix. Further a support vector machine (SVM)-based method has been developed to model the TAP binding affinity of peptides. The correlation (r = 0.80) was obtained between the predicted and experimental measured values by using sequence-based SVM. The reliability of prediction was further improved by cascade SVM that uses features of amino acids along with sequence. An extremely good correlation (r = 0.88) was obtained between measured and predicted values, when the cascade SVM-based method was evaluated through jackknife testing. A Web service, TAPPred (http://www.imtech.res.in/raghava/tappred/ or), has been developed based on this approach.

Web-server:http://www.imtech.res.in/raghava/tappred/
http://bioinformatics.uams.edu/mirror/tappred/ (mirror site).
Useful for: Prediction of TAP binding peptides in a protein.
Reference: *Protein Sci.* 2004; **13**:596-607.

8. Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences.

Abstract: This manuscript describes a support vector machine based method for the prediction of constitutive as well as immunoproteasome cleavage sites in antigenic sequences. This method achieved Matthew's correlation coefficents of 0.54 and 0.43 on in vitro and major histocompatibility complex ligand data, respectively. This shows that the performance of our method is comparable to that of the NetChop method, which is currently considered to be the best method for proteasome cleavage site prediction. Based on the method, a web server, Pcleavage, has also been developed. This server accepts protein sequences in any standard format and present results in a user-friendly format.

Web-server: http://www.imtech.res.in/raghava/btxpred/
http://bioinformatics.uams.edu/mirror/pcleavage/.
Useful for: Prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences.
Reference: *Nucleic Acids Res.* 2005; **33(Web Server issue)**:W202-207.

9. BcePred: Prediction of Continuous B-Cell Epitopes in Antigenic Sequences Using Physico-chemical Properties

Abstract: A crucial step in designing of peptide vaccines involves the identification of B-cell epitopes. In past, numerous methods have been developed for predicting continuous B-cell epitopes, most of these methods are based on physico-chemical properties of amino acids. Presently, its difficult to say which residue property or method is better than the others because there is no independent evaluation or benchmarking of existing methods. In this study the performance of various residue properties commonly used in B-cell epitope prediction has been evaluated on a clean dataset. The dataset used in this study consists of 1029 non-redundant B cell epitopes obtained from Bcipep database and equally number of non-epitopes obtained randomly from SWISS-PROT database. The performance of each residue property used in existing methods has been computed at various thresholds on above dataset. The accuracy of prediction based on properties varies between 52.92% and 57.53%. We have also evaluated the combination of two or more properties as combination of parameters enhance the accuracy of prediction. Based on our analysis we have developed a method for predicting B cell epitopes, which combines four residue properties. The accuracy of this method is 58.70%, which is slightly better than any single residue property. A web server has been developed to predict B cell epitopes in an antigen sequence.

Web-server: http://www.imtech.res.in/raghava/bcepred/
Useful for: Prediction of continuous B-cell epitopes.
Reference: *ICARIS* 2004, **LNCS 3239**, 197-204, Springer,2004.

10. ABCPred: Prediction of continuous B-cell epitopes in an antigen using recurrent neural network.

Abstract: B-cell epitopes play a vital role in the development of peptide vaccines, in diagnosis of diseases, and also for allergy research. Experimental methods used for characterizing epitopes are time consuming and demand large resources. The availability of epitope prediction method(s) can rapidly aid experimenters in simplifying this problem. The standard feed-forward (FNN) and recurrent neural network (RNN) have been used in this study for predicting B-cell epitopes in an antigenic sequence. The networks have been trained and tested on a clean data set, which consists of 700 non-redundant B-cell epitopes obtained from Bcipep database and equal number of non-epitopes obtained randomly from Swiss-Prot database. The networks have been trained and tested at different input window length and hidden units. Maximum accuracy has been obtained using recurrent neural network (Jordan network) with a single hidden layer of 35 hidden units for window length of 16. The final network yields an overall prediction accuracy of 65.93% when tested by fivefold cross-validation. The corresponding sensitivity, specificity, and positive prediction values are 67.14, 64.71, and 65.61%, respectively. It has been observed that RNN (JE) was more successful than FNN in the prediction of B-cell epitopes. The length of the peptide is also important in the prediction of B-cell epitopes from antigenic sequences.

Web-server: http://www.imtech.res.in/raghava/gstpred/
Useful for: Prediction of continuous B-cell epitopes in a protein.
Reference: *Proteins.* 2006; **65**; 40-48.

11. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes.

Abstract: In this study a systematic attempt has been made to integrate various approaches in order to predict allergenic proteins with high accuracy. The dataset used for testing and training consists of 578

allergens and 700 non-allergens obtained from A. K. Bjorklund, D. Soeria-Atmadja, A. Zorzet, U. Hammerling and M. G. Gustafsson (2005) Bioinformatics, 21, 39-50. First, we developed methods based on support vector machine using amino acid and dipeptide composition and achieved an accuracy of 85.02 and 84.00%, respectively. Second, a motif-based method has been developed using MEME/MAST software that achieved sensitivity of 93.94 with 33.34% specificity. Third, a database of known IgE epitopes was searched and this predicted allergenic proteins with 17.47% sensitivity at specificity of 98.14%. Fourth, we predicted allergenic proteins by performing BLAST search against allergen representative peptides. Finally hybrid approaches have been developed, which combine two or more than two approaches. The performance of all these algorithms has been evaluated on an independent dataset of 323 allergens and on 101 725 non-allergens obtained from Swiss-Prot. A web server AlgPred has been developed for the predicting allergenic proteins and for mapping IgE epitopes on allergenic proteins.

Web-server: http://www.imtech.res.in/raghava/algpred/
Useful for: Prediction of allergenic proteins.
Reference: *Nucleic Acids Res.* 2006; **34**; W202-209.

12. APSSP2: Advanced Protein Secondary Structure Prediction Server

Abstract: This server allow to predict the secondary structure of protein's from their amino acid sequence. This is an advanced version of our PSSP server, which participated in CASP3 and in CASP4. PSSP was also part of CAFASP2. Raghava, G. P. S. (2000) Protein secondary structure prediction using nearest neighbor and neural network approach. CASP4: 75-76. This server is also participating in world-wide Live-Bench competition EVA, so you can get the perfomance of methods including APSSP2 from EVA Server. This server is also part of Meta II Prediction server. Please visit, ExPASy Tools for more protein structure prediction tools. APSSP2 particpated in CASP5 / CAFASP3 and predicted all CASP5 target proteins with high accoracy (See CAFASP3 results) . It got 2nd rank with Q3 = 82.5% (slighlty lower to SSPRO(Ist) Q3=82.7%), and 4th rank Q3=79.0% (Slightly lower than PSIPred, SAM-T02sec and PROFphd) in two categories of CAFASP3. The performance of of APSSP2 in EVA was impressive when all proteins were considered, where it perform better than all other methods (Q3 = 82.9%).

Web-server: http://www.imtech.res.in/raghava/apssp2/
Useful for: Prediction of protein secondary structure
Reference:*CASP5.* A-132.

13.PepBuild: a web server for building structure data of peptides/proteins

Abstract: PepBuild, a web server, will aid in designing and building a capped or uncapped peptide/protein with known secondary and tertiary structure. The user can build a peptide/protein by choosing the required amino acid residue with regular secondary structure. The torsional angles can be supplied by the user, if desired. The server also allows the user to add relevant protecting groups at the N- and/or C-terminal of the peptide. The amino acid side chains of the designed peptide are optimized using rotameric libraries. Finally, the server provides the option of displaying the result or downloading the complete file in PDB (Protein Data Bank) format. This PDB file can later be used as an input for various molecular simulation programs or for graphical display.

Web-server: http://www.imtech.res.in/bvs/pepbuild/
Useful for: designing and generating a capped or uncapped peptide/protein with known secondary and tertiary structure
Reference:*Nucleic Acids Research* 2004 **32(Web Server Issue)**:W559-W561

14. PROCLASS: Protein Structure Classification Server

Abstract: A Computer program called PROCLASS has been described which is developed for predicting the structural class of protein from its amino acid sequence. This program implement a statistical algorithm

described recently by Chou, 1995 (2). The unique feature of this statistical algorithm is that it incorporate the coupling of different amino acid components, which distinguish it from the previous algorithms. The program PROCLASS utilize the novel approach of prediction of protein structural classes in a (20-1) D amino acid composition space and the matrices described by Chou, 1995. This server allow to predict the class of protein from its amino acid sequence. It predict weather protein belong to class Alpha or Beta or Alpha+Beta or Alpha/Beta.

Web-server: http://www.imtech.res.in/raghava/proclass/
Useful for: Prediction of protein structure classification.
Reference:*J. Biosciences*, 1999; **24**;176.

15. SARpred: Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure.

Abstract: The present study is an attempt to develop a neural network-based method for predicting the real value of solvent accessibility from the sequence using evolutionary information in the form of multiple sequence alignment. In this method, two feed-forward networks with a single hidden layer have been trained with standard back-propagation as a learning algorithm. The Pearson's correlation coefficient increases from 0.53 to 0.63, and mean absolute error decreases from 18.2 to 16% when multiple-sequence alignment obtained from PSI-BLAST is used as input instead of a single sequence. The performance of the method further improves from a correlation coefficient of 0.63 to 0.67 when secondary structure information predicted by PSIPRED is incorporated in the prediction. The final network yields a mean absolute error value of 15.2% between the experimental and predicted values, when tested on two different nonhomologous and nonredundant datasets of varying sizes. The method consists of two steps: (1) in the first step, a sequence-to-structure network is trained with the multiple alignment profiles in the form of PSI-BLAST-generated position-specific scoring matrices, and (2) in the second step, the output obtained from the first network and PSIPRED-predicted secondary structure information is used as an input to the second structure-to-structure network. Based on the present study, a server SARpred (http://www.imtech.res.in/raghava/sarpred/) has been developed that predicts the real value of solvent accessibility of residues for a given protein sequence. We have also evaluated the performance of SARpred on 47 proteins used in CASP6 and achieved a correlation coefficient of 0.68 and a MAE of 15.9% between predicted and observed values.

Web-server: http://www.imtech.res.in/raghava/sarpred/
Useful for: prediction of solvent accessibility (real value) of amino acid residues of a protein.
Reference: *Proteins.* 2005; **61**; 318-324.

16. TBBpred: Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods.

Abstract: This article describes a method developed for predicting transmembrane beta-barrel regions in membrane proteins using machine learning techniques: artificial neural network (ANN) and support vector machine (SVM). The ANN used in this study is a feed-forward neural network with a standard back-propagation training algorithm. The accuracy of the ANN-based method improved significantly, from 70.4% to 80.5%, when evolutionary information was added to a single sequence as a multiple sequence alignment obtained from PSI-BLAST. We have also developed an SVM-based method using a primary sequence as input and achieved an accuracy of 77.4%. The SVM model was modified by adding 36 physicochemical parameters to the amino acid sequence information. Finally, ANN- and SVM-based methods were combined to utilize the full potential of both techniques. The accuracy and Matthews correlation coefficient (MCC) value of SVM, ANN, and combined method are 78.5%, 80.5%, and 81.8%, and 0.55, 0.63, and 0.64, respectively. These methods were trained and tested on a nonredundant data set of 16 proteins, and performance was evaluated using "leave one out cross-validation" (LOOCV).

Web-server:http://www.imtech.res.in/raghava/tbbpred/

Useful for: Prediction of trans-membrane regions of beta-barrel proteins.
Reference:*Proteins.* 2004; **56** :11-18.

17. BhairPred: prediction of beta-hairpins in a protein from multiple alignment information using ANN and SVM techniques.

Abstract: This paper describes a method for predicting a supersecondary structural motif, beta-hairpins, in a protein sequence. The method was trained and tested on a set of 5102 hairpins and 5131 non-hairpins, obtained from a non-redundant dataset of 2880 proteins using the DSSP and PROMOTIF programs. Two machine-learning techniques, an artificial neural network (ANN) and a support vector machine (SVM), were used to predict beta-hairpins. An accuracy of 65.5% was achieved using ANN when an amino acid sequence was used as the input. The accuracy improved from 65.5 to 69.1% when evolutionary information (PSI-BLAST profile), observed secondary structure and surface accessibility were used as the inputs. The accuracy of the method further improved from 69.1 to 79.2% when the SVM was used for classification instead of the ANN. The performances of the methods developed were assessed in a test case, where predicted secondary structure and surface accessibility were used instead of the observed structure. The highest accuracy achieved by the SVM based method in the test case was 77.9%. A maximum accuracy of 71.1% with Matthew's correlation coefficient of 0.41 in the test case was obtained on a dataset previously used by X. Cruz, E. G. Hutchinson, A. Shephard and J. M. Thornton (2002) Proc. Natl Acad. Sci. USA, 99, 11157-11162. The performance of the method was also evaluated on proteins used in the '6th community-wide experiment on the critical assessment of techniques for protein structure prediction (CASP6)'.

Web-server:http://www.imtech.res.in/raghava/bhairpred/
Useful for: Prediction of beta-hairpins in a protein.
Reference: *Nucleic Acids Res.* 2005; **33(Web Server issue)**:W154-159.

18. Ar_NHPred: Role of evolutionary information in prediction of aromatic-backbone NH interactions in proteins.

Abstract: In this study, an attempt has been made to develop a neural network-based method for predicting segments in proteins containing aromatic-backbone NH (Ar-NH) interactions using multiple sequence alignment. We have analyzed 3121 segments seven residues long containing Ar-NH interactions, extracted from 2298 non-redundant protein structures where no two proteins have more than 25% sequence identity. Two consecutive feed-forward neural networks with a single hidden layer have been trained with standard back-propagation as learning algorithm. The performance of the method improves from 0.12 to 0.15 in terms of Matthews correlation coefficient (MCC) value when evolutionary information (multiple alignment obtained from PSI-BLAST) is used as input instead of a single sequence. The performance of the method further improves from MCC 0.15 to 0.20 when secondary structure information predicted by PSIPRED is incorporated in the prediction. The final network yields an overall prediction accuracy of 70.1% and an MCC of 0.20 when tested by five-fold cross-validation. Overall the performance is 15.2% higher than the random prediction. The method consists of two neural networks: (i) a sequence-to-structure network which predicts the aromatic residues involved in Ar-NH interaction from multiple alignment of protein sequences and (ii) a structure-to structure network where the input consists of the output obtained from the first network and predicted secondary structure. Further, the actual position of the donor residue within the 'potential' predicted fragment has been predicted using a separate sequence-to-structure neural network.

Web-server:http://www.imtech.res.in/raghava/ar_nhpred/
http://bioinformatics.uams.edu/mirror/ar_nhpred/ (mirror site).
Useful for: Prediction of aromatic-backbone NH interactions in proteins.
Reference:*FEBS Lett.* 2004;**564**:47-57.

19. CHpredict:Prediction of C alpha-H...O and C alpha-H...pi interactions in proteins using recurrent neural network.

Abstract: In this study, an attempt has been made to develop a method for predicting weak hydrogen bonding interactions, namely, C alpha-H...O and C alpha-H...pi interactions in proteins using artificial neural network. Both standard feed-forward neural network (FNN) and recurrent neural networks (RNN) have been trained and tested using five-fold cross-validation on a non-homologous dataset of 2298 protein chains where no pair of sequences has more than 25% sequence identity. It has been found that the prediction accuracy varies with the separation distance between donor and acceptor residues. The maximum sensitivity achieved with RNN for C alpha-H...O is 51.2% when donor and acceptor residues are four residues apart (i.e. at delta D-A = 4) and for C alpha-H...pi is 82.1% at delta D-A = 3. The performance of RNN is increased by 1-3% for both types of interactions when PSIPRED predicted protein secondary structure is used. Overall, RNN performs better than feed-forward networks at all separation distances between donor-acceptor pair for both types of interactions.

Web-server:http://www.imtech.res.in/raghava/chpredict/
Useful for: Prediction of donor and acceptor residues in C alpha-H...O and C alpha-H...pi interactions in proteins.
Reference: *In Silico Biol.* 2006; **6**; 111-125.

20. BTEVAL: a server for evaluation of beta-turn prediction methods.

Abstract: This paper describes a web server BTEVAL, developed for assessing the performance of newly developed beta-turn prediction method and it's ranking with respect to other existing beta-turn prediction methods. Evaluation of a method can be carried out on a single protein or a number of proteins. It consists of clean data set of 426 non-homologous proteins with seven subsets of these proteins. Users can evaluate their method on any subset or a complete set of data. The method is assessed at amino acid level and performance is evaluated in terms of Qtotal, Qpredicted, Qobserved and MCC measures. The server also compares the performance of the method with other existing beta-turn prediction methods such as Chou-Fasman algorithm, Thornton's algorithm, GORBTURN, 1-4 and 2-3 Correlation model, Sequence coupled model and BTPRED. The server is accessible from http://imtech.res.in/raghava/bteval/

Web-server:http://www.imtech.res.in/raghava/bteval/
Useful for: Evaluation of beta-turn prediction methods.
Reference: *J Bioinform Comput Biol.* 2003; **1**:495-504.

21. BetaTPred: prediction of beta-TURNS in a protein using statistical algorithms.

Abstract: beta-turns play an important role from a structural and functional point of view. beta-turns are the most common type of non-repetitive structures in proteins and comprise on average, 25% of the residues. In the past numerous methods have been developed to predict beta-turns in a protein. Most of these prediction methods are based on statistical approaches. In order to utilize the full potential of these methods, there is a need to develop a web server. This paper describes a web server called BetaTPred, developed for predicting beta-TURNS in a protein from its amino acid sequence. BetaTPred allows the user to predict turns in a protein using existing statistical algorithms. It also allows to predict different types of beta-TURNS e.g. type I, I', II, II', VI, VIII and non-specific. This server assists the users in predicting the consensus beta-TURNS in a protein.

Web-server:http://www.imtech.res.in/raghava/betatpred/
Useful for: Prediction of beta turns in proteins.
Reference: *Bioinformatics.* 2002;**18**:498-499.

22. BetatPred2 : Prediction of beta-turns in proteins from multiple alignment using neural network.

Abstract: A neural network-based method has been developed for the prediction of beta-turns in proteins by using multiple sequence alignment. Two feed-forward back-propagation networks with a single hidden layer are used where the first-sequence structure network is trained with the multiple sequence alignment in

the form of PSI-BLAST-generated position-specific scoring matrices. The initial predictions from the first network and PSIPRED-predicted secondary structure are used as input to the second structure-structure network to refine the predictions obtained from the first net. A significant improvement in prediction accuracy has been achieved by using evolutionary information contained in the multiple sequence alignment. The final network yields an overall prediction accuracy of 75.5% when tested by sevenfold cross-validation on a set of 426 nonhomologous protein chains. The corresponding Q(pred), Q(obs), and Matthews correlation coefficient values are 49.8%, 72.3%, and 0.43, respectively, and are the best among all the previously published beta-turn prediction methods.

Web-server:http://www.imtech.res.in/raghava/betapred2/
Useful for: Prediction of beta-turns in proteins.
Reference:*Protein Sci.* 2003;**12**:627-634.

23. Betaturns: A neural network method for prediction of beta-turn types in proteins using evolutionary information.

Abstract: The prediction of beta-turns is an important element of protein secondary structure prediction. Recently, a highly accurate neural network based method Betatpred2 has been developed for predicting beta-turns in proteins using position-specific scoring matrices (PSSM) generated by PSI-BLAST and secondary structure information predicted by PSIPRED. However, the major limitation of Betatpred2 is that it predicts only beta-turn and non-beta-turn residues and does not provide any information of different beta-turn types. Thus, there is a need to predict beta-turn types using an approach based on multiple sequence alignment, which will be useful in overall tertiary structure prediction. In the present work, a method has been developed for the prediction of beta-turn types I, II, IV and VIII. For each turn type, two consecutive feed-forward back-propagation networks with a single hidden layer have been used where the first sequence-to-structure network has been trained on single sequences as well as on PSI-BLAST PSSM. The output from the first network along with PSIPRED predicted secondary structure has been used as input for the second-level structure-to-structure network. The networks have been trained and tested on a non-homologous dataset of 426 proteins chains by 7-fold cross-validation. It has been observed that the prediction performance for each turn type is improved significantly by using multiple sequence alignment. The performance has been further improved by using a second level structure-to-structure network and PSIPRED predicted secondary structure information. It has been observed that Type I and II beta-turns have better prediction performance than Type IV and VIII beta-turns. The final network yields an overall accuracy of 74.5, 93.5, 67.9 and 96.5% with MCC values of 0.29, 0.29, 0.23 and 0.02 for Type I, II, IV and VIII beta-turns, respectively, and is better than random prediction.

Web-server:http://www.imtech.res.in/raghava/betaturns/
http://bioinformatics.uams.edu/mirror/betaturns/
Useful for: Prediction of beta-turn types in proteins.
Reference: *Bioinformatics.* 2004; **20**:2751-2758.

24. AlphaPred: Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information.

Abstract: In this paper a systematic attempt has been made to develop a better method for predicting alpha-turns in proteins. Most of the commonly used approaches in the field of protein structure prediction have been tried in this study, which includes statistical approach "Sequence Coupled Model" and machine learning approaches; i) artificial neural network (ANN); ii) Weka (Waikato Environment for Knowledge Analysis) Classifiers and iii) Parallel Exemplar Based Learning (PEBLS). We have also used multiple sequence alignment obtained from PSIBLAST and secondary structure information predicted by PSIPRED. The training and testing of all methods has been performed on a data set of 193 non-homologous protein X-ray structures using five-fold cross-validation. It has been observed that ANN with multiple sequence alignment and predicted secondary structure information outperforms other methods. Based on our observations we have developed an ANN-based method for predicting alpha-turns in proteins. The main

components of the method are two feed-forward back-propagation networks with a single hidden layer. The first sequence-structure network is trained with the multiple sequence alignment in the form of PSI-BLAST-generated position specific scoring matrices. The initial predictions obtained from the first network and PSIPRED predicted secondary structure are used as input to the second structure-structure network to refine the predictions obtained from the first net. The final network yields an overall prediction accuracy of 78.0% and MCC of 0.16.

Web-server:http://www.imtech.res.in/raghava/alphapred/
Useful for: Prediction of alpha-turns in proteins.
Reference:*Proteins.* 2004; **55**:83-90.

25. GammaPred: A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment.

Abstract: In the present study, an attempt has been made to develop a method for predicting gamma-turns in proteins. First, we have implemented the commonly used statistical and machine-learning techniques in the field of protein structure prediction, for the prediction of gamma-turns. All the methods have been trained and tested on a set of 320 nonhomologous protein chains by a fivefold cross-validation technique. It has been observed that the performance of all methods is very poor, having a Matthew's Correlation Coefficient (MCC)

Web-server:http://www.imtech.res.in/raghava/gammapred/
Useful for: Prediction of gamma-turns in proteins.
Reference:*Protein Sci.* 2003;**12**:923-929.

26. PepStr: A de novo Method for Tertiary Structure Prediction of Small Bioactive Peptides.

Abstract: Among secondary structure elements, beta-turns are ubiquitous and major feature of bioactive peptides. We analyzed 77 biologically active peptides with length varying from 9 to 20 residues. Out of 77 peptides, 58 peptides were found to contain at least one beta-turn. Further, at the residue level, 34.9% of total peptide residues were found to be in beta-turns, higher than the number of helical (32.3%) and beta-sheet residues (6.9%). So, we utilized the predicted beta-turns information to develop an improved method for predicting the three-dimensional (3D) structure of small peptides. In principle, we built four different structural models for each peptide. The first 'model I' was built by assigning all the peptide residues an extended conformation (phi = Psi = 180 degrees). Second 'model II' was built using the information of regular secondary structures (helices, beta-strands and coil) predicted from PSIPRED. In third 'model III', secondary structure information including beta-turn types predicted from BetaTurns method was used. The fourth 'model IV' had main-chain phi, Psi angles of model III and side chain angles assigned using standard Dunbrack backbone dependent rotamer library. These models were further refined using AMBER package and the resultant C(alpha) rmsd values were calculated. It was found that adding the beta-turns to the regular secondary structures greatly reduces the rmsd values both before and after the energy minimization. Hence, the results indicate that regular and irregular secondary structures, particularly beta-turns information can provide valuable and vital information in the tertiary structure prediction of small bioactive peptides.

Web-server:http://www.imtech.res.in/raghava/pepstr/
Useful for:prediction of tertiary structure of small bioactive peptides.
Reference:*Protein Pept Lett.* 2007;**14(7)**:626-631.

27. Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation

Abstract: Repetitive DNA sequences, besides having a variety of regulatory functions, are one of the principal causes of genomic instability. Understanding their origin and evolution is of fundamental importance for genome studies. The identification of repeats and their units helps in deducing the intra-

genomic dynamics as an important feature of comparative genomics. A major difficulty in identification of repeats arises from the fact that the repeat units can be either exact or imperfect, in tandem or dispersed, and of unspecified length. The Spectral Repeat Finder program circumvents these problems by using a discrete Fourier transformation to identify significant periodicities present in a sequence. The specific regions of the sequence that contribute to a given periodicity are located through a sliding window analysis, and an exact search method is then used to find the repetitive units. Efficient and complete detection of repeats is provided together with interactive and detailed visualization of the spectral analysis of input sequence. We demonstrate the utility of our method with various examples that contain previously unannotated repeats. A Web server has been developed for convenient access to the automated program.

Web-server:http://www.imtech.res.in/raghava/srf
http://www2.imtech.res.in/raghava/srf
Useful for: Prediction of spectral repeats.
Reference:: *Bioinformatics.* 2004; **20**:1405-1412.

28. FTG: Locating probable genes using Fourier transform approach.

Abstract: FTG is a web server for analyzing nucleotide sequences to predict the genes using Fourier transform techniques. This server implements the existing Fourier transform algorithms for gene prediction and allows the rapid visualization of analysis by output in GIF format.

Web-server:http://www.imtech.res.in/raghava/ftg/
Useful for: Gene prediction using FFT.
Reference:*Bioinformatics.* 2002;**18**:196-197.

29. EGPred: prediction of eukaryotic genes using ab initio methods after combining with sequence similarity approaches.

Abstract: EGPred is a Web-based server that combines ab initio methods and similarity searches to predict genes, particularly exon regions, with high accuracy. The EGPred program proceeds in the following steps: (1) an initial BLASTX search of genomic sequence against the RefSeq database is used to identify protein hits with an E-value <1; (2) a second BLASTX search of genomic sequence against the hits from the previous run with relaxed parameters (E-values <10) helps to retrieve all probable coding exon regions; (3) a BLASTN search of genomic sequence against the intron database is then used to detect probable intron regions; (4) the probable intron and exon regions are compared to filter/remove wrong exons; (5) the NNSPLICE program is then used to reassign splicing signal site positions in the remaining probable coding exons; and (6) finally ab initio predictions are combined with exons derived from the fifth step based on the relative strength of start/stop and splice signal sites as obtained from ab initio and similarity search. The combination method increases the exon level performance of five different ab initio programs by 4%-10% when evaluated on the HMR195 data set. Similar improvement is observed when ab initio programs are evaluated on the Burset/Guigo data set. Finally, EGPred is demonstrated on an approximately 95-Mbp fragment of human chromosome 13. The list of predicted genes from this analysis are available in the supplementary material. The EGPred program is computationally intensive due to multiple BLAST runs during each analysis.

Web-server:http://www.imtech.res.in/raghava/egpred/
Useful for: Annotation of eukaryotic genome.
Reference:*Genome Res.* 2004;**14**:1756-1766.

30. GWFASTA: server for FASTA search in eukaryotic and microbial genomes.

Abstract: Similarity searches are a powerful method for solving important biological problems such as database scanning, evolutionary studies, gene prediction, and protein structure prediction. FASTA is a widely used sequence comparison tool for rapid database scanning. Here we describe the GWFASTA

server that was developed to assist the FASTA user in similarity searches against partially and/or completely sequenced genomes. GWFASTA consists of more than 60 microbial genomes, eight eukaryote genomes, and proteomes of annotatedgenomes. Infact, it provides the maximum number of databases for similarity searching from a single platform. GWFASTA allows the submission of more than one sequence as a single query for a FASTA search. It also provides integrated post-processing of FASTA output, including compositional analysis of proteins, multiple sequences alignment, and phylogenetic analysis. Furthermore, it summarizes the search results organism-wise for prokaryotes and chromosome-wise for eukaryotes. Thus, the integration of different tools for sequence analyses makes GWFASTA a powerful tool for biologists.

Web-server:http://www.imtech.res.in/raghava/gwfasta/
Useful for: Genome wide FASTA search.
Reference:*Biotechniques.* 2002;**33**:548-550, 552, 554-556.

31.ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST.

Abstract: Automated prediction of subcellular localization of proteins is an important step in the functional annotation of genomes. The existing subcellular localization prediction methods are based on either amino acid composition or N-terminal characteristics of the proteins. In this paper, support vector machine (SVM) has been used to predict the subcellular location of eukaryotic proteins from their different features such as amino acid composition, dipeptide composition and physico-chemical properties. The SVM module based on dipeptide composition performed better than the SVM modules based on amino acid composition or physico-chemical properties. In addition, PSI-BLAST was also used to search the query sequence against the dataset of proteins (experimentally annotated proteins) to predict its subcellular location. In order to improve the prediction accuracy, we developed a hybrid module using all features of a protein, which consisted of an input vector of 458 dimensions (400 dipeptide compositions, 33 properties, 20 amino acid compositions of the protein and 5 from PSI-BLAST output). Using this hybrid approach, the prediction accuracies of nuclear, cytoplasmic, mitochondrial and extracellular proteins reached 95.3, 85.2, 68.2 and 88.9%, respectively. The overall prediction accuracy of SVM modules based on amino acid composition, physico-chemical properties, dipeptide composition and the hybrid approach was 78.1, 77.8, 82.9 and 88.0%, respectively. The accuracy of all the modules was evaluated using a 5-fold cross-validation technique. Assigning a reliability index (reliability index > or =3), 73.5% of prediction can be made with an accuracy of 96.4%.

Web-server:http://www.imtech.res.in/raghava/eslpred/
Useful for: Prediction of subcellular localization of eukaryotic proteins.
Reference: *Nucleic Acids Res.* 2004; **32(Web Server issue)**:W414-419.

32.PSLpred: prediction of subcellular localization of bacterial proteins.

Abstract: We developed a web server PSLpred for predicting subcellular localization of gram-negative bacterial proteins with an overall accuracy of 91.2%. PSLpred is a hybrid approach-based method that integrates PSI-BLAST and three SVM modules based on compositions of residues, dipeptides and physico-chemical properties. The prediction accuracies of 90.7, 86.8, 90.3, 95.2 and 90.6% were attained for cytoplasmic, extracellular, inner-membrane, outer-membrane and periplasmic proteins, respectively. Furthermore, PSLpred was able to predict approximately 74% of sequences with an average prediction accuracy of 98% at RI = 5.

Web-server:http://www.imtech.res.in/raghava/pslpred/
Useful for: Prediction of subcellular localization of bacterial proteins.
Reference: *Bioinformatics.* 2005; **21**: 2522-2524.

33. HSLPred: Support vector machine-based method for subcellular localization of human proteins

using amino acid compositions, their order, and similarity search.

Abstract: Here we report a systematic approach for predicting subcellular localization (cytoplasm, mitochondrial, nuclear, and plasma membrane) of human proteins. First, support vector machine (SVM)-based modules for predicting subcellular localization using traditional amino acid and dipeptide (i + 1) composition achieved overall accuracy of 76.6 and 77.8%, respectively. PSI-BLAST, when carried out using a similarity-based search against a nonredundant data base of experimentally annotated proteins, yielded 73.3% accuracy. To gain further insight, a hybrid module (hybrid1) was developed based on amino acid composition, dipeptide composition, and similarity information and attained better accuracy of 84.9%. In addition, SVM modules based on a different higher order dipeptide i.e. i + 2, i + 3, and i + 4 were also constructed for the prediction of subcellular localization of human proteins, and overall accuracy of 79.7, 77.5, and 77.1% was accomplished, respectively. Furthermore, another SVM module hybrid2 was developed using traditional dipeptide (i + 1) and higher order dipeptide (i + 2, i + 3, and i + 4) compositions, which gave an overall accuracy of 81.3%. We also developed SVM module hybrid3 based on amino acid composition, traditional and higher order dipeptide compositions, and PSI-BLAST output and achieved an overall accuracy of 84.4%.

Web-server: http://www.imtech.res.in/raghava/hslpred/
Useful for: Prediction of subcellular localization of human proteins.
Reference: *J Biol Chem.* 2005; **280**:14427-14432.

34. TBpred: Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs.

Abstract: In past number of methods have been developed for predicting subcellular location of eukaryotic, prokaryotic (Gram-negative and Gram-positive bacteria) and human proteins but no method has been developed for mycobacterial proteins which may represent repertoire of potent immunogens of this dreaded pathogen. In this study, attempt has been made to develop method for predicting subcellular location of mycobacterial proteins.The models were trained and tested on 852 mycobacterial proteins and evaluated using five-fold cross-validation technique. First SVM (Support Vector Machine) model was developed using amino acid composition and overall accuracy of 82.51% was achieved with average accuracy (mean of class-wise accuracy) of 68.47%. In order to utilize evolutionary information, a SVM model was developed using PSSM (Position-Specific Scoring Matrix) profiles obtained from PSIBLAST (Position-Specific Iterated BLAST) and overall accuracy achieved was of 86.62% with average accuracy of 73.71%. In addition, HMM (Hidden Markov Model), MEME/MAST (Multiple Em for Motif Elicitation / Motif Alignment and Search Tool) and hybrid model that combined two or more models were also developed. We achieved maximum overall accuracy of 86.8% with average accuracy of 89.00% using combination of PSSM based SVM model and MEME/MAST. Performance of our method was compared with that of the existing methods developed for predicting sub-cellular locations of Gram-positive bacterial proteins. A highly accurate method has been developed for predicting subcellular location of mycobacterial proteins. This method also predicts very important class of proteins that is membrane-attached proteins. This method will be useful in annotating newly sequenced or hypothetical mycobacterial proteins.

Web-server:http://www.imtech.res.in/raghava/tbpred/
Useful for: Functional annotation of Mycobacterium tuberculosis proteins
Reference:*BMC Bioinformatics.* 2007; **8(1)**:337

35. MitPred: Prediction of mitochondrial proteins using support vector machine and hidden Markov model.

Abstract: Mitochondria are considered as one of the core organelles of eukaryotic cells hence prediction of mitochondrial proteins is one of the major challenges in the field of genome annotation. This study describes a method, MitPred, developed for predicting mitochondrial proteins with high accuracy. The data set used in this study was obtained from Guda, C., Fahy, E. & Subramaniam, S. (2004) Bioinformatics 20,

1785-1794. First support vector machine-based modules/methods were developed using amino acid and dipeptide composition of proteins and achieved accuracy of 78.37 and 79.38%, respectively. The accuracy of prediction further improved to 83.74% when split amino acid composition (25 N-terminal, 25 C-terminal, and remaining residues) of proteins was used. Then BLAST search and support vector machine-based method were combined to get 88.22% accuracy. Finally we developed a hybrid approach that combined hidden Markov model profiles of domains (exclusively found in mitochondrial proteins) and the support vector machine-based method. We were able to predict mitochondrial protein with 100% specificity at a 56.36% sensitivity rate and with 80.50% specificity at 98.95% sensitivity. The method estimated 9.01, 6.35, 4.84, 3.95, and 4.25% of proteins as mitochondrial in Saccharomyces cerevisiae, Drosophila melanogaster, Caenorhabditis elegans, mouse, and human proteomes, respectively.

Web-server:http://www.imtech.res.in/raghava/mitpred/
Useful for: Prediction of mitochondrial proteins.
Reference:*J Biol Chem.* 2006; **281**; 5357-5363.

36. NRpred: Classification of nuclear receptors based on amino acid composition and dipeptide composition.

Abstract: Nuclear receptors are key transcription factors that regulate crucial gene networks responsible for cell growth, differentiation, and homeostasis. Nuclear receptors form a superfamily of phylogenetically related proteins and control functions associated with major diseases (e.g. diabetes, osteoporosis, and cancer). In this study, a novel method has been developed for classifying the subfamilies of nuclear receptors. The classification was achieved on the basis of amino acid and dipeptide composition from a sequence of receptors using support vector machines. The training and testing was done on a non-redundant data set of 282 proteins obtained from the NucleaRDB data base (1). The performance of all classifiers was evaluated using a 5-fold cross validation test. In the 5-fold cross-validation, the data set was randomly partitioned into five equal sets and evaluated five times on each distinct set while keeping the remaining four sets for training. It was found that different subfamilies of nuclear receptors were quite closely correlated in terms of amino acid composition as well as dipeptide composition. The overall accuracy of amino acid composition-based and dipeptide composition-based classifiers were 82.6 and 97.5%, respectively. Therefore, our results prove that different subfamilies of nuclear receptors are predictable with considerable accuracy using amino acid or dipeptide composition.

Web-server:http://www.imtech.res.in/raghava/nrpred/
Useful for: Classification of nuclear receptors.
Reference: *J Biol Chem.* 2004; **279**:23262-23266.

37. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors.

Abstract: G-protein coupled receptors (GPCRs) belong to one of the largest superfamilies of membrane proteins and are important targets for drug design. In this study, a support vector machine (SVM)-based method, GPCRpred, has been developed for predicting families and subfamilies of GPCRs from the dipeptide composition of proteins. The dataset used in this study for training and testing was obtained from http://www.soe.ucsc.edu/research/compbio/gpcr/. The method classified GPCRs and non-GPCRs with an accuracy of 99.5% when evaluated using 5-fold cross-validation. The method is further able to predict five major classes or families of GPCRs with an overall Matthew's correlation coefficient (MCC) and accuracy of 0.81 and 97.5% respectively. In recognizing the subfamilies of the rhodopsin-like family, the method achieved an average MCC and accuracy of 0.97 and 97.3% respectively. The method achieved overall accuracy of 91.3% and 96.4% at family and subfamily level respectively when evaluated on an independent/blind dataset of 650 GPCRs. We have also suggested subfamilies for 42 sequences which were previously identified as unclassified ClassA GPCRs.

Web-server: http://www.imtech.res.in/raghava/gpcrpred/

Useful for: Prediction of GPCR families and subfamilies.
Reference:*Nucleic Acids Res.* 2004;**32(Web Server issue)** :W383-389.

38. GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors.

Abstract: The receptors of amine subfamily are specifically major drug targets for therapy of nervous disorders and psychiatric diseases. The recognition of novel amine type of receptors and their cognate ligands is of paramount interest for pharmaceutical companies. In the past, Chou and co-workers have shown that different types of amine receptors are correlated with their amino acid composition and are predictable on its basis with considerable accuracy [Elrod and Chou (2002) Protein Eng., 15, 713-715]. This motivated us to develop a better method for the recognition of novel amine receptors and for their further classification. The method was developed on the basis of amino acid composition and dipeptide composition of proteins using support vector machine. The method was trained and tested on 167 proteins of amine subfamily of G-protein-coupled receptors (GPCRs). The method discriminated amine subfamily of GPCRs from globular proteins with Matthew's correlation coefficient of 0.98 and 0.99 using amino acid composition and dipeptide composition, respectively. In classifying different types of amine receptors using amino acid composition and dipeptide composition, the method achieved an accuracy of 89.8 and 96.4%, respectively. The performance of the method was evaluated using 5-fold cross-validation. The dipeptide composition based method predicted 67.6% of protein sequences with an accuracy of 100% with a reliability index > or =5.

Web-server:http://www.imtech.res.in/raghava/gpcrclass/
Useful for: Prediction of amine-binding receptors from its amino acid sequence.
Reference:*Nucleic Acids Res.* 2005; **33(Web Server issue)**:W143-147.

39. BTXpred: Prediction of bacterial toxins.

Abstract: This paper describes a method developed for predicting bacterial toxins from their amino acid sequences. All the modules, developed in this study, were trained and tested on a non-redundant dataset of 150 bacterial toxins that included 77 exotoxins and 73 endotoxins. Firstly, support vector machines (SVM) based modules were developed for predicting the bacterial toxins using amino acids and dipeptides composition and achieved an accuracy of 96.07% and 92.50%, respectively. Secondly, SVM based modules were developed for discriminating entotoxins and exotoxins, using amino acids and dipeptides composition and achieved an accuracy of 95.71% and 92.86%, respectively. In addition, modules have been developed for classifying the exotoxins (e. g. activate adenylate cyclase, activate guanylate cyclase, neurotoxins) using hidden Markov models (HMM), PSI-BLAST and a combination of the two and achieved overall accuracy of 95.75%, 97.87% and 100%, respectively.

Web-server:http://www.imtech.res.in/raghava/btxpred/
Useful for: Prediction of bacterial toxins.
Reference:*In-Silico Biol.* 2007; **7**; 0028

40. NTXpred: Prediction of neurotoxins based on their function and source.

Abstract: We have developed a method NTXpred for predicting neurotoxins and classifying them based on their function and origin. The dataset used in this study consists of 582 non-redundant, experimentally annotated neurotoxins obtained from Swiss-Prot. A number of modules have been developed for predicting neurotoxins using residue composition based on feed-forwarded neural network (FNN), recurrent neural network (RNN), support vector machine (SVM) and achieved maximum accuracy of 84.19%, 92.75%, 97.72% respectively. In addition, SVM modules have been developed for classifying neurotoxins based on their source (e.g., eubacteria, cnidarians, molluscs, arthropods have been and chordate) using amino acid composition and dipeptide composition and achieved maximum overall accuracy of 78.94% and 88.07% respectively. The overall accuracy increased to 92.10%, when the evolutionary information obtained from PSI-BLAST was combined with SVM module of source classification. We have also developed SVM

modules for classifying neurotoxins based on functions using amino acid, dipeptide composition and achieved overall accuracy of 83.11%, 91.10% respectively. The overall accuracy of function classification improved to 95.11%, when PSI-BLAST output was combined with SVM module. All the modules developed in this study were evaluated using five-fold cross-validation technique.

Web-server:http://www.imtech.res.in/raghava/ntxpred/
http://bioinformatics.uams.edu/mirror/ntxpred.
Useful for: Prediction and classification of neurotoxins
Reference:*In-Silico Biol.* 2007; **7**; 0028

41.VGIchan: prediction and classification of voltage-gated ion channels.

Abstract: This study describes methods for predicting and classifying voltage-gated ion channels. Firstly, a standard support vector machine (SVM) method was developed for predicting ion channels by using amino acid composition and dipeptide composition, with an accuracy of 82.89% and 85.56%, respectively. The accuracy of this SVM method was improved from 85.56% to 89.11% when combined with PSI-BLAST similarity search. Then we developed an SVM method for classifying ion channels (potassium, sodium, calcium, and chloride) by using dipeptide composition and achieved an overall accuracy of 96.89%. We further achieved a classification accuracy of 97.78% by using a hybrid method that combines dipeptide-based SVM and hidden Markov model methods. A web server VGIchan has been developed for predicting and classifying voltage-gated ion channels using the above approaches.

Web-server:http://www.imtech.res.in/raghava/vgichan/
Useful for: Prediction and classification of voltage-gated ion channels.
Reference: *Genomics Proteomics Bioinformatics.* 2006; **4**; 253-258.

42. VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition.

Abstract: In this study, an attempt has been made to predict the major functions of gram-negative bacterial proteins from their amino acid sequences. The dataset used for training and testing consists of 670 non-redundant gram-negative bacterial proteins (255 of cellular process, 60 of information molecules, 285 of metabolism, and 70 of virulence factors). First we developed an SVM-based method using amino acid and dipeptide composition and achieved the overall accuracy of 52.39% and 47.01%, respectively. We introduced a new concept for the classification of proteins based on tetrapeptides, in which we identified the unique tetrapeptides significantly found in a class of proteins. These tetrapeptides were used as the input feature for predicting the function of a protein and achieved the overall accuracy of 68.66%. We also developed a hybrid method in which the tetrapeptide information was used with amino acid composition and achieved the overall accuracy of 70.75%. A five-fold cross validation was used to evaluate the performance of these methods.

Web-server:http://www.imtech.res.in/raghava/vicmpred/
Useful for: Function prediction of gram-negative bacterial proteins.
Reference: *Genomics Proteomics Bioinformatics.* 2006; **4**; 42-47.

43. Oxypred: Prediction and classification of oxygen binding proteins using SVM

Abstract: OxyPred is to predict the Oxygen Binding Proteins, Which is carrying Erythrocruorin, Hemoglobin, Myoglobin, Hemerithrin, Leghemoglobin and Hemocyanin. These Proteins working as a Oxygen transport and binding throughout the animal and plant kingdom. Other organisms including Bacteria, Protozoans and Fungi all have hemoglobin like proteins, their known and predicted roles include the reversible binding of gaseous ligands. Myoglobin: Found in the muscle tissue of many vertebrates including humans (gives muscle tissue a distinct red or dark gray color). Is very similar to hemoglobin in structure and sequence, but is not arranged in tetramers, it is a monomer and lacks cooperative binding and

is used to store oxygen rather than transport it. Hemocyanin: Second most common oxygen transporting protein found in nature. Found in the blood of many arthropods and molluscs. Uses copper prosthetic group instead of iron heme groups and is blue in color when oxygenated. Hemerythrin: Some marine invertebrates and a few species of annelid use this iron containing non-heme protein to carry oxygen in their blood. Appears pink/violet when oxygenated, clear when not. Erythrocruorin: Found in many annelids, including earthworms. Giant free-floating blood protein, contains many dozens even hundreds of Iron heme containing protein subunits bound together into a single protein complex with a molecular masses greater than 3.5 million daltons. Leghemoglobin: In leguminous plants, such as alfalfa or soybeans, the nitrogen fixing bacteria in the roots are protected from oxygen by this iron heme containing, oxygen binding protein.

Web-server:http://www.imtech.res.in/raghava/oxypred/
Useful for: Prediction of oxygen binding proteins.
Reference: *Genomics, Proteomics & Bioinformatics* (In Press).

44. GSTPred: Support vector machine based prediction of glutathione S-transferase proteins.

Abstract: Glutathione S-transferase (GST) proteins play vital role in living organism that includes detoxification of exogenous and endogenous chemicals, survivability during stress condition. This paper describes a method developed for predicting GST proteins. We have used a dataset of 107 GST and 107 non-GST proteins for training and the performance of the method was evaluated with five-fold cross-validation technique. First a SVM based method has been developed using amino acid and dipeptide composition and achieved the maximum accuracy of 91.59% and 95.79% respectively. In addition we developed a SVM based method using tripeptide composition and achieved maximum accuracy 97.66% which is better than accuracy achieved by HMM based searching (96.26%).

Web-server:http://www.imtech.res.in/raghava/gstpred/
Useful for: Prediction of GST proteins.
Reference: *Protein Pept Lett.* 2007;**14**; 575-580.

45. Pprint: Prediction of RNA binding sites in a protein using SVM and PSSM profile

Abstract: RNA-binding proteins play key roles in post-transcriptional control of gene expression, which, along with transcriptional regulation, is a major way to regulate patterns of gene expression during development. Thus, the identification and prediction of RNA binding sites is an important step in comprehensive understanding of how RBPs control organism development. Combining evolutionary information and support vector machine (SVM), we have developed an improved method for predicting RNA binding sites or RNA interacting residues in a protein sequence. The prediction models developed in this study have been trained and tested on 86 RNA binding protein chains and evaluated using five-fold cross validation technique. First, a SVM model was developed that achieved a maximum Mathew s Correlation Coefficient (MCC) of 0.31. The performance of this SVM model further improved the MCC from 0.31 to 0.45, when multiple sequence alignment in the form of PSSM profiles was used as input to the SVM, which is far better than the maximum MCC achieved by previous methods (0.41) on the same dataset. In addition, SVM models were also developed on an alternative dataset that contained 107 RNA-binding protein chains. Utilizing PSSM as input information to the SVM, the training/testing on this alternate dataset achieved a maximum MCC of 0.32. Conclusively, the prediction performance of SVM models developed in this study is better than the existing methods on the same datasets.

Web-server:http://www.imtech.res.in/raghava/pprint/
Useful for: Prediction of RNA-interacting residues.
Reference:*Proteins (Structure, Function and Bioinformatics).* In Press

46. MANGO: Prediction of Genome Ontology (GO) class of a protein from its amino acid and dipeptide composition using nearest neighbor approach. CASP7: 93".

Abstract: One of the major challenges in era of genomics is to predict the function of proteins. As number of proteins whose sequence is known is growing with exponential rate due to advancement in DNA sequence techniques. This has pose a major challenge to the bio informatician to develop strategy to predict the function of protein. Fortunately, function of a large number proteins have been deduced using experimental techniques, one may obtained the information about manually annotated proteins from SWISSPROT database. Recently initiatives were taken to provide the uniform definition of class of protein. Genome ontology is one of the major source of information from where one can obtained the information of class of protein. In GO database the annotation of proteins are at three level i) Biological functions; ii) Biological Process and iii) cell. However, a large number of method already developed in past to predict the class of proteins are limited to predict few classes of proteins. It has been shown in past that dipeptide composition have more information than simple composition because order of neighbor is also considered. Thus we implement our approach using dipeptide composition, where dipeptide composition of proteins were used to calculate Euclidian distance between query protein and GO class of proteins instead of residue composition. We also compute the overall difference (residue composition and dipeptide composition) in query and GO class of proteins. In summary we used composition, dipeptide composition and comination of both for predictiog GO class of target proteins. A server for predicting functional class of a protein. It predict function according to GO categories. The method is developed on protein in UNIPROT database whoes function have been assigned manually according to GO criteria.

Web-server:http://www.imtech.res.in/raghava/mango/.
Useful for: Prediction protein function
Reference:*CASP7*: **93**.

47. MHCBN: a comprehensive database of MHC binding and non-binding peptides.

Abstract: MHCBN is a comprehensive database of Major Histocompatibility Complex (MHC) binding and non-binding peptides compiled from published literature and existing databases. The latest version of the database has 19 777 entries including 17 129 MHC binders and 2648 MHC non-binders for more than 400 MHC molecules. The database has sequence and structure data of (a) source proteins of peptides and (b) MHC molecules. MHCBN has a number of web tools that include: (i) mapping of peptide on query sequence; (ii) search on any field; (iii) creation of data sets; and (iv) online data submission. The database also provides hypertext links to major databases like SWISS-PROT, PDB, IMGT/HLA-DB, GenBank and PUBMED.

Web-server: http://www.imtech.res.in/raghava/mhcbn/
http://srs.ebi.ac.uk/ (SRS version).
Useful for: Searching of MHC binding and non-binding peptides.
Reference:*Bioinformatics.* 2003;**19**:665-666.

48. Bcipep: a database of B-cell epitopes.

Abstract: Bcipep is a database of experimentally determined linear B-cell epitopes of varying immunogenicity collected from literature and other publicly available databases. The current version of Bcipep database contains 3031 entries that include 763 immunodominant, 1797 immunogenic and 471 null-immunogenic epitopes. It covers a wide range of pathogenic organisms like viruses, bacteria, protozoa, and fungi. The database provides a set of tools for the analysis and extraction of data that includes keyword search, peptide mapping and BLAST search. It also provides hyperlinks to various databases such as GenBank, PDB, SWISS-PROT and MHCBN. A comprehensive database of B-cell epitopes called Bcipep has been developed that covers information on epitopes from a wide range of pathogens. The Bcipep will be source of information for investigators involved in peptide-based vaccine design, disease diagnosis and research in allergy. It should also be a promising data source for the development and evaluation of methods for prediction of B-cell epitopes.

Web-server:http://www.imtech.res.in/raghava/bcipep/
Useful for: Searching of B-cell epitopes.
Reference: *BMC Genomics.* 2005; **6**:79.

49. HaptenDB: a comprehensive database of haptens, carrier proteins and anti-hapten antibodies.

Abstract: The key requirement for successful immunochemical assay is the availability of antibodies with high specificity and desired affinity. Small molecules, when used as haptens, are not immunogenic. However, on conjugating with carrier molecule they elicit antibody response. The production of anti-hapten antibodies of desired specificity largely depends on the hapten design (preserving greatly the chemical structure and spatial conformation of target compound), selection of the appropriate carrier protein and the conjugation method. This manuscript describes a curated database HaptenDB, where information is collected from published literature and web resources. The current version of the database has 2021 entries for 1087 haptens and 25 carrier proteins, where each entry provides comprehensive details about (1) nature of the hapten, (2) 2D and 3D structures of haptens, (3) carrier proteins, (4) coupling method, (5) method of anti-hapten antibody production, (6) assay method (used for characterization) and (7) specificities of antibodies. The current version of HaptenDB covers a wide array of haptens including pesticides, herbicides, insecticides, drugs, vitamins, steroids, hormones, toxins, dyes, explosives, etc. It provides internal and external links to various databases/resources to obtain further information about the nature of haptens, carriers and respective antibodies. For structure similarity comparison of haptens, the database also integrates tools like JME Editor and JMOL for sketching, displaying and manipulating hapten 2D/3D structures online. So the database would be of great help in identifying functional group(s) in smaller molecules using antibodies as well as for the development of immunodiagnostics/therapeutics by providing data and procedures available so far for the generation of specific or cross-reactive antibodies.

Web-server: http://www.imtech.res.in/raghava/haptendb/,
http://bioinformatics.uams.edu/raghava/haptendb/
Useful for: Searching of haptens, carrier proteins and anti-hapten antibodies.
Reference:*Bioinformatics.* 2006 Jan 15;**22(2)**:253-255.

50. AntiBP:Analysis and prediction of antibacterial peptides.

Abstract: Antibacterial peptides are important components of the innate immune system, used by the host to protect itself from different types of pathogenic bacteria. Over the last few decades, the search for new drugs and drug targets has prompted an interest in these antibacterial peptides. We analyzed 486 antibacterial peptides, obtained from antimicrobial peptide database APD, in order to understand the preference of amino acid residues at specific positions in these peptides. It was observed that certain types of residues are preferred over others in antibacterial peptides, particularly at the N and C terminus. These observations encouraged us to develop a method for predicting antibacterial peptides in proteins from their amino acid sequence. First, the N-terminal residues were used for predicting antibacterial peptides using Artificial Neural Network (ANN), Quantitative Matrices (QM) and Support Vector Machine (SVM), which resulted in an accuracy of 83.63%, 84.78% and 87.85%, respectively. Then, the C-terminal residues were used for developing prediction methods, which resulted in an accuracy of 77.34%, 82.03% and 85.16% using ANN, QM and SVM, respectively. Finally, ANN, QM and SVM models were developed using N and C terminal residues, which achieved an accuracy of 88.17%, 90.37% and 92.11%, respectively. All the models developed in this study were evaluated using five-fold cross validation technique. These models were also tested on an independent or blind dataset. Among antibacterial peptides, there is preference for certain residues at N and C termini, which helps to demarcate them from non-antibacterial peptides. Both the termini play a crucial role in imparting the antibacterial property to these peptides. Among the methods developed, SVM shows the best performance in predicting antibacterial peptides followed by QM and ANN, in that order. AntiBP (Antibacterial peptides) will help in discovering efficacious antibacterial peptides, which we hope will prove to be a boon to combat the dreadful antibiotic resistant bacteria.

Web-server:http://www.imtech.res.in/raghava/antibp/

Useful for: Prediction of antibacterial peptides
Reference:*BMC Bioinformatics.* 2007; **8**; 263.

51. RBpred: Machine learning techniques in disease forecasting: a case study on rice blast prediction.

Abstract: Diverse modeling approaches viz. neural networks and multiple regression have been followed to date for disease prediction in plant populations. However, due to their inability to predict value of unknown data points and longer training times, there is need for exploiting new prediction softwares for better understanding of plant-pathogen-environment relationships. Further, there is no online tool available which can help the plant researchers or farmers in timely application of control measures. This paper introduces a new prediction approach based on support vector machines for developing weather-based prediction models of plant diseases. Six significant weather variables were selected as predictor variables. Two series of models (cross-location and cross-year) were developed and validated using a five-fold cross validation procedure. For cross-year models, the conventional multiple regression (REG) approach achieved an average correlation coefficient (r) of 0.50, which increased to 0.60 and percent mean absolute error (%MAE) decreased from 65.42 to 52.24 when back-propagation neural network (BPNN) was used. With generalized regression neural network (GRNN), the r increased to 0.70 and %MAE also improved to 46.30, which further increased to r = 0.77 and %MAE = 36.66 when support vector machine (SVM) based method was used. Similarly, cross-location validation achieved r = 0.48, 0.56 and 0.66 using REG, BPNN and GRNN respectively, with their corresponding %MAE as 77.54, 66.11 and 58.26. The SVM-based method outperformed all the three approaches by further increasing r to 0.74 with improvement in %MAE to 44.12. Overall, this SVM-based prediction approach will open new vistas in the area of forecasting plant diseases of various crops. Our case study demonstrated that SVM is better than existing machine learning techniques and conventional REG approaches in forecasting plant diseases. In this direction, we have also developed a SVM-based web server for rice blast prediction, a first of its kind worldwide, which can help the plant science community and farmers in their decision making process.

Web-server:http://www.imtech.res.in/raghava/rbpred/
Useful for: Prediction of rice blast disease.
Reference:*BMC Bioinformatics.* 2006; **7**; 485.

52. LGEpred: Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein.

Abstract: A large number of papers have been published on analysis of microarray data with particular emphasis on normalization of data, detection of differentially expressed genes, clustering of genes and regulatory network. On other hand there are only few studies on relation between expression level and composition of nucleotide/protein sequence, using expression data. There is a need to understand why particular genes/proteins express more in particular conditions. In this study, we analyze 3468 genes of Saccharomyces cerevisiae obtained from Holstege et al., (1998) to understand the relationship between expression level and amino acid composition. RESULTS: We compute the correlation between expression of a gene and amino acid composition of its protein. It was observed that some residues (like Ala, Gly, Arg and Val) have significant positive correlation (r > 0.20) and some other residues (Like Asp, Leu, Asn and Ser) have negative correlation (r < -0.15) with the expression of genes. A significant negative correlation (r = -0.18) was also found between length and gene expression. These observations indicate the relationship between percent composition and gene expression level. Thus, attempts have been made to develop a Support Vector Machine (SVM) based method for predicting the expression level of genes from its protein sequence. In this method the SVM is trained with proteins whose gene expression data is known in a given condition. Then trained SVM is used to predict the gene expression of other proteins of the same organism in the same condition. A correlation coefficient r = 0.70 was obtained between predicted and experimentally determined expression of genes, which improves from r = 0.70 to 0.72 when dipeptide composition was used instead of residue composition. The method was evaluated using 5-fold cross validation test. We also demonstrate that amino acid composition information along with gene expression data can be used for improving the function classification of proteins. CONCLUSION: There is a

correlation between gene expression and amino acid composition that can be used to predict the expression level of genes up to a certain extent.

Web-server: http://www.imtech.res.in/raghava/lgepred/.
Useful for: calculation of the correlation between amino acid composition and gene expression and prediction of expression level.
Reference:*BMC Bioinformatics.* 2005; **6**:59.

53. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy.

Abstract: The alignment of two or more protein sequences provides a powerful guide in the prediction of the protein structure and in identifying key functional residues, however, the utility of any prediction is completely dependent on the accuracy of the alignment. In this paper we describe a suite of reference alignments derived from the comparison of protein three-dimensional structures together with evaluation measures and software that allow automatically generated alignments to be benchmarked. We test the OXBench benchmark suite on alignments generated by the AMPS multiple alignment method, then apply the suite to compare eight different multiple alignment algorithms. The benchmark shows the current state-of-the art for alignment accuracy and provides a baseline against which new alignment algorithms may be judged. The simple hierarchical multiple alignment algorithm, AMPS, performed as well as or better than more modern methods such as CLUSTALW once the PAM250 pair-score matrix was replaced by a BLOSUM series matrix. AMPS gave an accuracy in Structurally Conserved Regions (SCRs) of 89.9% over a set of 672 alignments. The T-COFFEE method on a data set of families with <8 sequences gave 91.4% accuracy, significantly better than CLUSTALW (88.9%) and all other methods considered here. The OXBench suite of reference alignments, evaluation software and results database provide a convenient method to assess progress in sequence alignment techniques. Evaluation measures that were dependent on comparison to a reference alignment were found to give good discrimination between methods. The STAMP Sc Score which is independent of a reference alignment also gave good discrimination. Application of OXBench in this paper shows that with the exception of T-COFFEE, the majority of the improvement in alignment accuracy seen since 1985 stems from improved pair-score matrices rather than algorithmic refinements. The maximum theoretical alignment accuracy obtained by pooling results over all methods was 94.5% with 52.5% accuracy for alignments in the 0-10 percentage identity range. This suggests that further improvements in accuracy will be possible in the future.

Web-server:http://www.compbio.dundee.ac.uk/Software/Oxbench/oxbench.html
Useful for: Evaluation of protein multiple sequence alignment methods.
Reference:*BMC Bioinformatics.* 2003; **4**:47.

54. AbAg: A Web Based Method for Computing Endpoint Titer and Concentration of Antibody/Antigen.

Abstract: In this report we have described a web-server for calculating the endpoint titers and concentrations of antibody/antigen (Ab/Ag) from the optical density (OD) taken from ELISA data. The server utilize a graphical method (Raghava et al., 1992) for determining the concentration of either the antibody or the antigen. In order to calculate the endpoint titer, first we fit the OD verses concentration of control data using a least square curve-fitting method. Then we fit the OD verses concentration of standard sample using graphical method. Finally, we determine the intersection or nearest point of two curves which we have called the endpoint titer. In order to calculate concentrations of the antibody/antigen of unknown samples, we have to first fit OD verses concentrations of the known samples using graphical method and to determine the linear interpolation and hyperbolic formulas. Then we calculate the concentrations of the unknown samples from their OD using these formula's.

Web-server:http://www.imtech.res.in/raghava/abag/
Useful for: Computing Endpoint Titer and Concentration of Antibody/Antigen.
Reference:*Biotech Software and Internet Reports*, 2001, **2(5)**.

55. DNAOPT: a computer program to aid optimization of DNA gel electrophoresis and SDS-PAGE.

Abstract: Several methods and computer programs have been developed for estimating the size of DNA fragments from gel electrophoresis. However, methods are lacking that may facilitate in optimization of gel conditions. In this article, a computer program called DNAOPT is described, which was developed to assist researchers in tuning the gel conditions of gel electrophoresis. The DNAOPT program fits the reciprocal of the migration distance vs. the size of the DNA fragments using the hyperbolic regression method and computes the hyperbolic parameters such as signal, flatness and capacity (optimization parameters). The program further manipulates these parameters obtained by running gel electrophoresis under various conditions (i) to determine the relationship between the gel conditions (temperature, buffer concentration, electric field strength, etc.) and optimization parameters; (ii) to demonstrate gel electrophoresis curves and optimization parameters graphically; and (iii) to represent the optimizing parameters at different gel conditions in tabular form. The above-mentioned program options aid the users in selecting optimum gel conditions by running the gel repeatedly under various conditions in which the agarose concentration, electric field strength, temperature, buffer concentration and so on are varied in a systematic way for each set of gel conditions. Similarly, this program can also be used to optimize gel conditions of sodium dodecyl sulfate polyacrylamide gel electrophoresis.

Web-server:http://www.imtech.res.in/raghava/progs/dnaopt/README.html
Useful for: Optimization of conditions of gel electrophoresis.
Reference:*Biotechniques.* 1995 Feb;**18(2)**:274-278, 280.

56. Hemo: A simple microassay for computing the hemolytic potency of drugs

Abstract: A simple microassay and computer program are described for determining the erythrocyte hemolytic potency of drugs in vitro. This microassay is sensitive for both micro as well as macro ranges of hemoglobin concentration. An ELISA reader has been adapted to read erythrocyte lysis (hemolysis), which reduces the number and culture of replicates. A computer program was developed that calculates parameters such as C50 (concentration of drug causing 50% hemolysis), C100 (concentration of drug causing 100% hemolysis) and beta (slope of the curve) and graphically expresses the hemolytic patterns of various drugs simultaneously. The program can obtain optical densities directly from a 96-well plate ELISA reader by interfacing the microplate reader to the computer or by using a keyboard. This method is useful for screening a large number of hemolytic drugs and requires lower amounts of test compounds. It may also be applicable to quantitative functional assays, such as complement-mediated hemolysis and enumeration of antibody-secreting cells. The program can be obtained from the authors on request.

Web-server:http://www.imtech.res.in/raghava/progs/hemo/
Useful for: Prediction of hemolytic potency of drugs
Reference:*Biotechniques*. 1994 Dec;**17(6)**:1148-1153.

57. DNASIZE: Improved estimation of DNA fragment length from gel electrophoresis data using a graphical method.

Abstract: A computer program has been developed for computing DNA fragment size from its electrophoretic mobility using a graphical method. The program uses DNA marker data and selects the semilogarithmic linear range (sl-range), i.e., the linear portion of the semilogarithmic curve (mobility vs. log of DNA fragment length). Over this range a linear interpolation is derived for calculating the size of a DNA fragment whose mobility falls in the sl-range. The program also derives a hyperbolic interpolation formula that covers the entire range for determining the size of a DNA fragment whose mobility is beyond the semilogarithmic linear range. The method described in this paper is sensitive, accurate and reliable. This program can also be used to compute protein or polypeptide size from sodium dodecyl sulfate polyacrylamide gel electrophoresis data.

Web-server: http://www.imtech.res.in/raghava/progs/dnasize/
Useful for: Estimation of DNA fragment length from gel electrophoresis.
Reference:*Biotechniques.* 1994;**17**:100-104.

58. PSAweb: A Graphical Web Server for the Analysis of Protein Sequences and Multiple Sequence Alignment

Abstract: A dynamic web server has been developed to analysis the amino acid sequence of proteins and their multiple sequence alignment. This is a comprehensive on-line Internet tool that allows rapid visualization, via GIF output, of aligned sequences and their analysis. It allows to perform following analysis on primary structure of protein i) to compute the physical properties of amino acid (e.g., hydrophobicity, charge) in specific way of mean formation in a given window; ii) to present amino acid properties along primary structure in graphical and in table form; iii) to highlight a residue or group of residues in sequence that exhibit a specific function. This web server also allows to analysis the multiple alignment of protein sequences. It allows following analysis on multiple alignment i) to compute the overall physical property of each position in alignment; ii) to present the properties of each position in an alignment, graphically; iii) to compare the various properties of an alignment by presenting these properties in same graph and iv) to highlight the conserved residues in alignment. It generates the graphical output in GIF format and text output in HTML so that result can be view the result by using any web browser over the Internet.

Web-server: http://www.imtech.res.in/raghava/psa/
Useful for: Analysis of protein sequence and multiple sequence alignment.
Reference:*Biotech Software and Internet Report*. 2001; **2**: 255-258.

59. ELISA_eq: Calculation of antibody and antigen concentrations from ELISA data using a graphical method.

Abstract: A graphical method for determining the concentration of either the antibody or the antigen from ELISA data is presented in the form of a GWBASIC program. In the program, ELISAEQ, optical densities (OD) obtained from a 96-well ELISA plate can be input either directly by interfacing a microplate reader to the computer or manually. The program uses standard sample data, and selects the semilogarithmic linear range. Over this range, a least-squares method is used to determine the concentrations of interest. In addition, a hyperbolic interpolation formula is derived over the entire range for estimating the antibody or antigen concentration of the unknown samples whose OD is beyond the linear range.

Web-server:http://www.imtech.res.in/raghava/progs/elisaeq/README.html
Useful for:calculation of Ab/Ag concentration from ELISA data.
Reference: *J Immunol Methods.* 1992 Aug 30;**153(1-2)**:263-264.

60. IL4IFNG: Measurement and computation of murine interleukin-4 and interferon-gamma by exploiting the unique abilities of these lymphokines to induce the secretion of IgG1 and IgG2a.

Abstract: A specific and new method for measuring Interleukin-4 and Interferon-gamma, based on the estimation of IgG1 and IgG2a isotypes secretion from B cells is described. An antagonizing effect of Interferon-gamma in the production of IgG1 induced by Interleukin-4 was neutralized by using antibody to Interferon-gamma. Similarly, the interference of Interleukin-4 in the Interferon-gamma mediated enhancement of IgG2a production was blocked by anti-Interleukin-4 antibody. The high concentrations of Interleukin-4 and Interferon gamma inhibited the secretion of IgG1 and IgG2a respectively. Therefore, in the assay described, the samples containing the cytokines were so diluted that their activity fell into the non-inhibitory zone. A computer program has also been developed for determining the concentrations of lymphokines.

Web-server:http://www.imtech.res.in/raghava/progs/il4ifng/README.html

Useful for:calculation of affinity of an antibody using non-competitive ELISA
Reference: *J Immunoassay.* 1993 Mar-Jun;**14(1-2)**:83-97.

61. Ab_affi: Method for determining the affinity of monoclonal antibody using non-competitive ELISA: a computer program.

Abstract: A simple and reliable method based upon law of mass action for calculating affinity of a monoclonal antibody using non-competitive ELISA, is described. In this method, the binding of an antibody (Ab) with an antigen (Ag) is measured by ELISA using serial dilutions of both antigen (coated on the plate) as well as antibody. When the OD measured after the antigen antibody interaction was plotted against the concentration of Ab, added to the wells, a hyperbolic curve was obtained. The OD, at any point of the curve, was considered as a direct reflection of the amount of antibody bound to the antigen. The OD-100 denotes the occupancy of maximum no. of epitopes available on the antigen molecules, accessible to the antigen. The concentration of antibody (Ab, Ab') at corresponding levels of antigen concentration (Ag, Ag'), presents the value obtained at OD-50. The [Ag] and [Ag'] are not the true antigen concentrations but are the measurement of antigen density on the plate. The affinity constant K(aff) was calculated by using the formula K(aff) = (n - 1)/2(n[Ab'] - [Ab]), derived from law of mass action, where n = [Ag]/[Ag']. A computer program to calculate the affinity of antibody to the antigen using method described in this manuscript has been developed and discussed.

Web-server:http://www.imtech.res.in/raghava/progs/il4ifng/README.html
Useful for:calculation of affinity of an antibody using non-competitive ELISA
Reference: *J Immunoassay.* 1994 May;**15(2)**:115-128.

62. GMAP: a multi-purpose computer program to aid synthetic gene design, cassette mutagenesis and the introduction of potential restriction sites into DNA sequences.

Abstract: A computer program called GMAP has been developed for i) mapping the potential restriction endonuclease (R.E.) sites that can be introduced in a nonambiguous DNA sequence; ii) predicting the mutations required to introduce unique R.E. sites in the nonambiguous DNA sequences; and iii) searching all R.E. sites in ambiguous DNA sequence obtained by reverse translation of a given amino acid sequence. This allows the design of synthetic genes as well as the modular redesign after introducing limited base pair mismatches in wild-type genes in order to adapt them for "cassette" mutagenesis. The GMAP program uses an algorithm based on set theory that reduces the degree of complexity from an exponential to linear function of sequence length. Therefore, the speed of searching for potential R.E. sites in reverse-translated gene sequences and the prediction of new R.E. sites in natural genes by mutations are rapid.

Web-server:http://www.imtech.res.in/raghava/progs/gmap/README.html
Useful for: Mapping of potential of RE sites.
Reference: *Biotechniques.* 1994 Jun;**16(6)**:1116-1123.

NOTES

BIC GROUP

Scientists/Staffs:

Dr. GPS Raghava; Dr. Balvinder Singh; Dr. Manoj Kumar; Mr. Harvinder Jassal

PhD Students:

Aarti Garg; Firoz Ahmad; Hifzur-rahman; Manish Datt; Mamoon Rashid; Manish Kumar; Neetu Saxena; Nitish Kumar; Ruchi Sachdeva; Ruchi Verma; Snehlata; Vijaya Brahma

Project Assistants:

Deepak Garg; Deepti Sethi; Reena Saini; Sandeep Sharma; Surnider Kumar; Yogita Sharma

URLs:

IMTECH: http://www.imtech.res.in/
BIC: http://www.imtech.res.in/bic/
Dr. GPS Raghava: http://www.imtech.res.in/raghava/
Dr. B. Singh: http://www.imtech.res.in/bvs/
Reprints: http://www.imtech.res.in/raghava/reprints/
Slides: http://www.imtech.res.in/raghava/slides/