Prediction of Biochemical Properties of Protein Active Site Residues with ANN Classifier

Brijesh Singh Yadav, Sweta Gupta, K. P. Mishra
Division of Bioinformatics Research, United Research Center, Allahabad, India

Abstract

In this study, we present a method for the prediction of physiochemical properties of catalytic sites residues using a suitable Artificial Neural Networking (ANN) Feed Forward Backpropagation algorithm coupled with a set of structural proteins with the properties of their amino acid residues. The method has been applied to a set of 100 structural proteins from the Protein Data Bank (PDB) having a ligand at their active site. Using *Ligplot* program for searching of active site residues and *Surface racer* for identifying the non active site moieties, the identified amino acid residues were classified in 15 different categories based on their physiochemical properties. After classification of active and non active site amino acids, their properties were converted into machine language. Furthermore, we created Neural Network Using *Matlab* software and generated algorithm for training and testing of data. Thereafter, analysis of results showed that 95% of active site's physiochemical properties were correctly predicted. It is hoped that this work would help in determining the surface topographic properties for ligand binding sites residues in protein. The computational outcome would be helpful in ligand designing, molecular docking, *de novo* drug designing and structural identification and functional sites Comparisons.

Keyword: Active Site Residues, Protein Databank, Ligand, Physiochemical Properties, Artificial Neural Networking

Introduction

Molecular design is important in various fields such as organic chemistry, physical chemistry, chemical engineering, chemical physics, bioengineering and molecular biology. No single strategy or method has come for ward that provides an optimum solution to the many different challenges involved in designing materials with new properties. Computational methods are needed for functional prediction of proteins. Advances in experimental and computational methods have quietly ushered in a new era in protein function annotation. This 'age of multiplicity' is marked by the notion that only the use of multiple tools, multiple evidence and considering the multiple aspects of

function can give us the broad picture that 21st century biology will need to link and alter micro- and macroscopic phenotypes. It might also help us to undo past mistakes by removing errors from our databases and prevent us from producing more. On the downside, multiplicity is often confusing, therefore systematically review methods and resources for automated protein function prediction, looking at individual (biochemical) and contextual (network) functions, respectively.[21]In particular, knowledge of the location of catalytic residues provides valuable insight into the mechanisms of enzyme catalyzed reactions. Many computational methods have been developed for predicting protein functions and functional residues involved in catalytic reactions, binding activities, and protein- protein interactions. Automated propagation of functional annotation from a protein with known function to homologous proteins is a well-established method for the assignment of protein function. However, reliable functional propagation generally requires a high degree of sequence similarity. For example, to transfer all four digits of an EC number at an error rate of below 10% needs at least 60% sequence identity [1], and only about 60% of the proteins can be annotated by a homology transfer of experimental functional information in 62 proteomes.[2]

The evolutionary trace (ET) method is used for prediction of active sites and functional interfaces in proteins with known structure. Based on the observation that functional residues are more conserved than other residues, the method finds the most conserved residues at different sequence identity cutoffs and, as a final step, relies on human visual examination of the residues on protein structures [3]. While the ET method was shown successful in many case studies [4-6], the need for manual inspection in this original implementation is not suitable for automated large-scale analysis. Modified and automated versions of the ET method have been developed and tested on two protein datasets [7]. In one study the catalytic residues were predicted correctly for 62 (77.5%) out of 80enzymes with the ACTSITE and SITE records from the PDB database in another study [8], ~60% (79%by manual analysis) of catalytic residues were predicted correctly for 29 enzymes with experimentally characterized active sites. Another group of methods, the *ab initio* methods [reviewed in [2, 9]], do not use sequence conservation for functional site prediction. These methods exploit general protein properties, such as residue buffer capacity [10], the electrostatic energy of charged residues [11], protein sub cellular localization [12], and conservation of local structural similarities. [13] These methods are potentially useful for the prediction of novel protein function seven if sequence conservation of the functional site in question is low. The last group of methods combines sequence conservation with different aspects of protein structure. [14-17]Three-dimensional cluster analysis predicted functional residues by examination of spatially-adjacent conserved residues [14], and achieved a high recovery (83%) with low error rate (2%) for the prediction of catalytic residues in 15 enzymes. A similar method enriched with two additional structural parameters predicted ~47% of catalytic residues at the 5% false positive rate among 39 enzymes from the CDD database with manually curate's catalytic sites. [15] A method for locating catalytic residues based on the sequence conservation, local special conservation, stability analysis, and geometrical

location of the residue predicted 56% of catalytic residues in 49 enzymes. [16]The method considered only highly conserved D, E, K, R,H, S, T, N, Y, and C residues. A trained neural network (NN) with spatial clustering predicted over 69% of catalytic residues with a high false positive rate among 189enzymes from the CATRES database containing manually curetted catalytic residues. [17] The method used sequence conservation, residue type, and four structural parameters as inputs for the NN. Direct comparison of methods is confounded by the use of different performance measure and different data sets of various size and quality. Nevertheless, the overall accuracy for the prediction of catalytic residues remains low (in the 70% range). [18] We describe a methodology that attempts to optimize two components, global shape and local physicochemical texture, for evaluating the similarity between a pair of surfaces. Surface shape similarity is assessed using a three-dimensional object recognition algorithm and physicochemical texture similarity is assessed through a spatial alignment of conserved residues between the surfaces. The comparisons are used in tandem to efficiently search the Global Protein Surface Survey (GPSS), a library of annotated surfaces derived from structures in the PDB, for studying evolutionary relationships and uncovering novel similarities between proteins.[19] We introduce a classifier to predict the small molecule–enzyme interaction, i.e., whether they can interact with each other. Small molecules are represented by their chemical functional groups, and enzymes are represented by their biochemical and physicochemical properties, resulting in a total of 160 features. These features are input into the AdaBoost classifier, which is known to have good generalization ability to predict interaction. As a result, the overall prediction accuracy, tested by tenfold cross-validation and independent sets, is 81.76% and 83.35%, respectively. [20] Protein-protein interactions play an important role in a number of biological activities. We developed two methods of predicting protein-protein interaction site residues. One method uses only sequence information and the other method uses both sequence and structural information. We used support vector machine (SVM) with a position specific scoring matrix (PSSM) as sequence information and accessible surface area (ASA) of polar and non-polar atoms as structural information. SVM is used in two stages. In the first stage, an interaction residue is predicted by taking PSSMs of sequentially neighboring residues or taking PSSMs and ASAs of spatially neighboring residues as features. The second stage acts as a filter to refine the prediction results. The recall and precision of the predictor using both sequence and structural information are 73.6% and 50.5%, respectively. We found that using PSSM instead of frequency of amino acid appearance was the main factor of improvement of our methods.[22]

This study was aimed to develop an improved fully-automated method for the prediction of physiochemical properties of catalytic residues of structural protein of PDB using a carefully selected and supervised Machine learning Backpropagation algorithm coupled with an optimal discriminative set of structural protein properties. This study helps in de novo prediction of properties of functional sites of proteins.

Material and Methods

(i) Compilation of benchmarking dataset (ii) Searching of active site and non active residues (iii) Classification of Residue features (iv) Binary Coding and generation of data set (v) Neural network training and testing.

Bench marking of dataset

The benchmarking dataset was compiled from the PDB the management of the PDB became the responsibility of the Research collaboratory for Structural Bioinformatics (RCSB). In general terms, the vision of the RCSB is to create a resource based on the fast modern technology that facilitates the use and analysis of structural data and thus creates an enabling resource for biological research. We collect the 100 Protein ligand complex hetero atoms in PDB database, for example: 1a4k.pdb, 1a4q.pdb, 1a5g.pdb, 1a42.pdb, 1a50.pdb, 1a94.pdb, 1aaq-1996.pdb, 1aaq-2002.pdb, abe-1983.Pdb, and 1abe-1983-2.pdb, 1abe-1996.pdb, 1abe-2002.pdb, 1abf etc.

Searching of active and non active residues

The LIGPLOT algorithm reads in the 3D structure of the ligand from the PDB file, together with the protein residues it interacts with, and `unrolls' each object about its rotatable bonds, flattening them out onto the 2D page. The program automatically generates schematic diagrams of protein-ligand interactions from the 3D coordinates in a PDB file, so we use the Ligplot for finding the active site amino acid. For searching of non active site Residues we use Surface *Racer***:** A computer programs, which perform fast calculations of the solvent accessible and molecular (solvent excluded) surface areas of macro molecules. All surface area and curvature calculations are analytical therefore yield exact values of these quantities. High calculation speed of this software is achieved primarily by avoiding computation- ally expensive mathematical procedures whenever possible and by efficient binding of surface data structures. We run the surface racer and find the non catalytic residues.

Encoding and generation of data set

For the initial analysis, each residue of the benchmarking dataset was represented as a vector with 15 residue property values and a label {put the value 1 for property which is present or 0 property which is not present} to indicate the hydrophobic or hydrophilic character, polar or non polar, contain aliphatic or aromatic side chain, positive charged, negative charged, uncharged, containing sulphur, optically inactive, making H bonding, cyclic, essential or nonessential etc important functional and structural parameters were derived for each residues(catalytic and non-catalytic) in all 100 proteins.

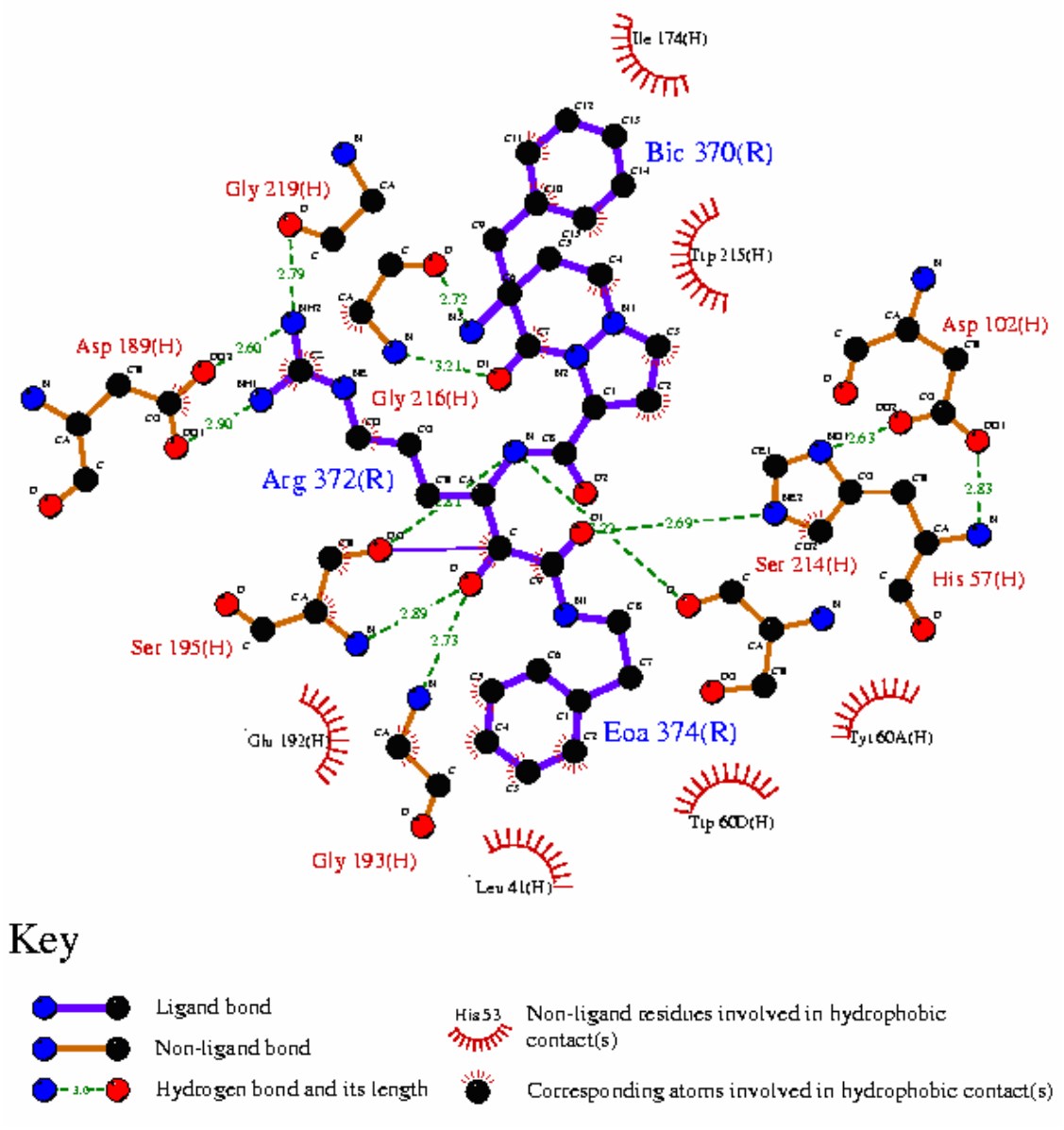

Fig1. Ligplot output of 1a46.pdb showing the catalytic site

Table1. The active and non active site amino acid of some structural proteins

Protein	Active site residue			Non catalytic Residue		
1a4k	Glu 81A	Asn35B	Gly230	SER	ARG	TYR
1a5g	Asp12	His57	Ser195	LYS	PRO	ARG
1a42	His199	Gln137	Thr199	ASP	ASN	GLN
1a46	Glu205	Gly216	Lys375	ARG	GLU	LYS

Table 2. Data represent the example of binary coding of residue properties

Protein	Residues	Property encoding (Polar, Non polar, Aromatic, Aliphatic, Acidic, Basic, Positive, Negative or Uncharged, Essential or not, Sulfur Contain, Optically inactive, Cyclic and H bonding)
1gld.pdb	Cys, Gln, His	100100100011000, 100010100010000, 0000001010010000

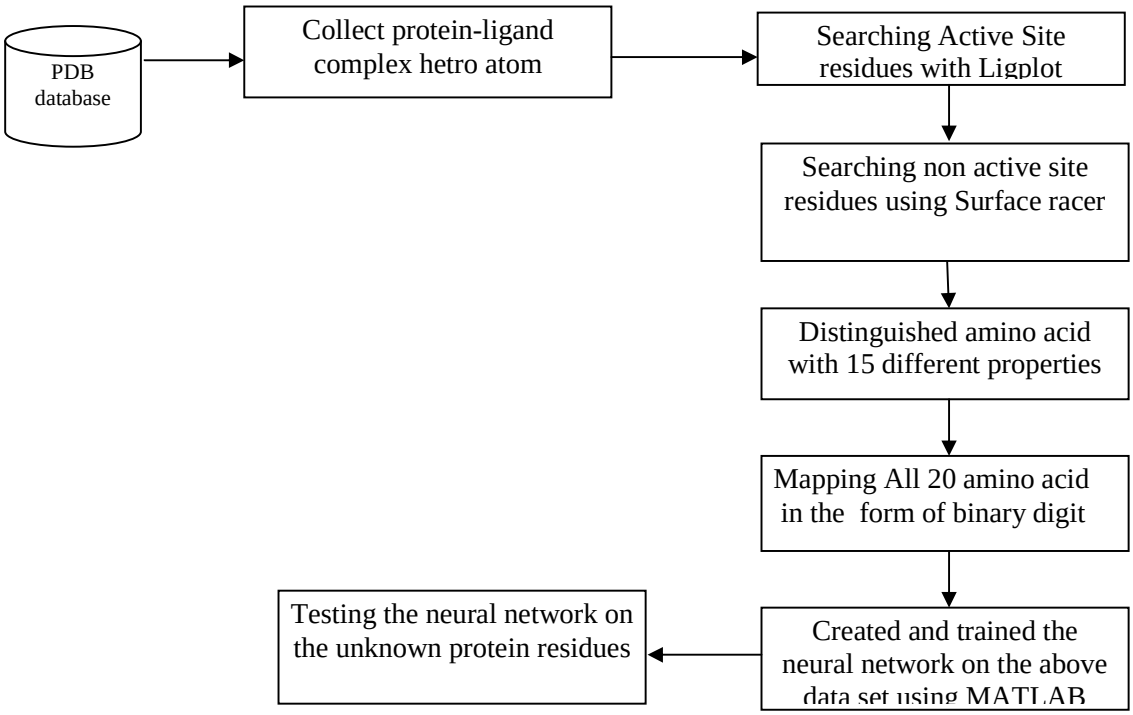

Fig 2. Diagrammatic representation of used methodology

Neural network training and testing of data

After generating the input data we divided data set in to group, one is training data for the train the network and another is testing data for the testing of network. Create the program using Matlab editor and fix input, output value and all variable in program for the network initiation. The three essential features of ANN are basic computing elements referred to as neurons, the network architecture describing the connections between the neurons and the training algorithm used to find values of the network parameters for performing a particular task. Each neuron performs a simple calculation, a scalar function of a scalar input. S Network architecture refers to the organization of the neurons and the types of connections. In the multilayer feed forward network, neurons are organized in a series of layers. Information flows only in one direction; units receive information only from units in higher layers of the network. The neural network that was used in this investigation consists of a 100 unit input layer and two-unit output layer, hidden layers were incorporated into the network. Furthermore, the network utilizes a feed-forward design, in which signals are transferred forward from the input units to the output unit. The output unit represents the prediction made by the neural network as to whether the central residue represents catalytic site or non catalytic site.

During each cycle, the inputs are presented to the network. The weights of the units are adjusted at the end of the cycle, and this procedure is repeated. Back-propagation, a type of learning algorithm, is used to optimize the adjustment of the weights. This form of

supervised training, in which the desired output is presented to the network along with the inputs, was used to train the neural network.
ANN can be 'trained' to perform a specific task by adjusting these weights. The weights are continually adjusted by comparing the output of the network with the target until the output of the network 'matches' the target in the sense that the error function measuring the difference between the target and the output is minimized. Many pairs of input and output are used tot rain the network and this mode of adjustment is called 'supervised' learning .A learning rule in which weights and biases are adjusted by error-derivative (delta) vectors back propagated through the network. Back propagation is commonly applied to feed forward multilayer networks. Output is compared with the target for each input and adjustments of the weights are made using a training algorithm, most often the back propagation algorithm. Incremental training is sometimes referred to 'adaptive' training. Back propagation involves two passes through the network, a forward pass and a backward pass. The forward pass generates the network's output activities and is generally the least computation intensive. The more time consuming backward pass involves propagating the error initially found in the output nodes back through the network to assign errors to each node that contributed to the initial error. Once all the errors are assigned, the weights are changed so as to minimize these errors.

Result and Discussion

The ANN models developed in this study are based on biochemical features of active site residue of structural protein. We found that the network reached an overall accuracy of 95% ± 2.86% based on amino acid derived features. In order to judge the neural network learning process, a suitable measure of performance is required. Total error (percentage of incorrect predictions) is not sufficient due to the highly unbalanced nature of the data set. All of the statistics are derived from the following quantities:

p = Number of correctly classified catalytic residues.
n = Number of correctly classified non-catalytic residues.
o = Number of non-catalytic residues incorrectly predicted to be catalytic (over-predictions).
u = Number of catalytic residues incorrectly predicted to be non-catalytic (under-predictions).
t = Total residues (p + n + o + u).
The total error (Q Total) is given by equation (1)
Q Total =p + n / t x 100 (2)

To complement this, two other measures of performance were used; Q Predicted measures the percentage of catalytic predictions that are correct and Q Observed measures the percentage of catalytic residues that are correctly predicted.

Training Result

(96/100)*100
ans = 96.00

Performance of Testing set

(38/40)*100
ans = 95.00

To determine the best machine learning algorithm, we train the neural network with training set of 100 PDB structural proteins with 50 catalytic residues and 50 non catalytic. The residue of the dataset was represented by a set of 15 residue properties (Biochemical Properties of amino acid)

Which encoded in binary digit (0 1) previously shown to be of functional relevance, as well as a label {1 0 / 0 1} to indicate catalytic/non-catalytic residue. This method correctly predicted 96 of the 100 residues, with an overall predictive accuracy of more than 90%. The results demonstrate that the developed ANN-based binary prediction of biochemical properties of catalytic site residue is adequate and can be considered an effective tool for *in silico* screening. The structure of a protein is an important determinant for the detailed molecular function of proteins, and would consequently also be useful for prediction of ligand molecules. Based on the analysis of biochemical features from protein structure, differences in the parameters between catalytic and non-catalytic have previously been shown to exist and used for prediction of catalytic/non-catalytic in archaeal. This agrees well with our result that sequence derived features can be used for predicting enzymes.
Presumably, accuracy of the approach operating by the structure derived features can be improved even further by expanding the parameters or by applying more powerful classification techniques such as Support Vector Machines or Bayesian Neural Networks.

Conclusion

An overall predictive accuracy of more than 90 %, missing only 4 % of the catalytic residues, The analysis of the optimal subset selected from the initial 15 residue properties indicated that the algorithm was capable to learn to distinguish catalytic from non-catalytic residues based on structural protein residues on protein surface properties like hydrophobic or hydrophilic character, polar or non polar, contain aliphatic or aromatic side chain, positive charged, negative charged, uncharged, containing sulphur, optically inactive, making H bonding, cyclic, essential or nonessential etc. This algorithm predicted fundamental features of catalytic residues, and could predict catalytic residues with accuracy > 90 % for proteins with known structure. This study shows that the choices of machine learning Neural network Backpropagation algorithm sets for the selected algorithm are critical for the prediction tasks. The results of the present work demonstrate that protein structure derived features with ANN Backpropagation classification method appear to be a very fast protein classification mechanism providing good results, comparable to some of the current efforts in the literature.

References

1. W Tian, J Skolnick: *J Mol Biol* **2003**, 333(4):863.
2. B Rost, J Liu, R Nair, KO Wrzeszczynski, Y Ofran: *CMLS Cell Mol Life Sci* **2003**, 60(12):2637.
3. O Lichtarge, HR Bourne, FE Cohen:.*J Mol Biol* **1996**, 257(2):342.
4. CA Innis, J Shi, TL Blundell: *Protein Engineering* **2000**, 13(12):839.

5. S Zhu, I Huys, K Dyason, F Verdonck, J Tytgat: *Proteins* **2004**, 54(2):361.
6. S Chakravarty, AM Hutson, MK Estes, BV Prasad: *J Virol* **2005**, 79(1):554.
7. P Aloy, E Querol, FX Aviles, MJE Sternberg: *J Mol Biol* **2001**, 311(2):395.
8. H Yao, DM Kristensen, I Mihalek, ME Sowa, C Shaw, M Kimmel, L Kavraki, O Lichtarge:. *J Mol Biol* **2003**, 326(1):255.
9. S Jones, JM Thornton: *Current Opinion in Chemical Biology* **2004**, 8(1):3.
10. MJ Ondrechen, JG Clifton, D Ringe:. *Proc Natl Acad Sci USA* **2001**, 98(22):12473.
11. AH Elcock:.*J Mol Biol* **2001**, 312(4):885.
12. PP Wangikar, AV Tendulkar, S Ramya, DN Mail, S Sarawagi. *J Mol Biol* **2003**, 326(3):955.
13. K Kinoshita , H Nakamura *Protein Sci* **2003**, 12(8):1589.
14. R Landgraf, I Xenarios, D Eisenberg: *J Mol Biol* **2001**, 307(5):1487.
15. AR Panchenko, F Kondrashov, S Bryant. *Protein Science* **2004**, 13(4):884.
16. M Ota, K Kinoshita, K Nishikawa. *J Mol Biol* **2003**, 327(5):1053.
17. A Gutteridge, GJ Bartlett, JM Thornton:. *J Mol Biol* **2003**, 330(4):719.
18. IH Witten, F Eibe: Data Mining: Practical machine learning tools and techniques. 2nd edition. Morgan Kaufmann, San Francisco; **2005.**
19. TA Binkowski, AJ Midwest. Protein Functional Surfaces: Global Shape Matching and Local Spatial Alignments of Ligand Binding Sites. Center for Structural Genomics and Structural Biology Center, Biosciences Division, Argonne National Laboratory, Argonne, Illinois, 60439, USA BMC Struct Biol. **2008**; 8: 45.
20. Bing Niu, Y Jin, L Lu, K Fen, L Gu, Z He, W Lu, Y Li, Y Cai : Prediction of interaction between small molecule and enzyme using AdaBoost Mol Divers, **2009.**
21. R Rentzsch, C A. Orengo :Protein function prediction – the power of multiplicity Institute of Structural and Molecular Biology, University College London, London, WC1E 6BT, UK Available online 27 February **2009.**
22. M Kakuta, S Nakamura, K Shimizu: Prediction of Protein-Protein Interaction Sites Using Only Sequence Information and Using Both Sequence and Structural Information. 1) Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo December 17, **2007.**