Molecular Informatics. 2015; 34(11-12): 736-741

Prediction of nucleotide binding peptides using star graph topological indices

Yong Liu[1,2,3], Cristian R. Munteanu[1], Enrique Fernández Blanco[1], Zhiliang Tan[3], Antonino Santos del Riego[1] andAlejandro Pazos[1]

[1] *Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruna, Campus de Elviña s/n, 15071, A Coruña, Spain, phone/fax: +34-981167000/+34-981167160*
[2] *Faculty of Veterinary Medicine and Animal Science, Autonomous University of the State of Mexico, Toluca, 50090, México*
[3] *Key Laboratory of Subtropical Agro-ecological Engineering, Institute of Subtropical Agriculture, the Chinese Academy of Sciences, Changsha, Hunan, 410125, P. R. China*

Abstract
The nucleotide binding proteins are involved in many important cellular processes, such as transmission of genetic information or energy transfer and storage. Therefore, the screening of new peptides for this biological function is an important research topic. The current study proposes a mixed methodology to obtain the first classification model that is able to predict new nucleotide binding peptides, using only the amino acid sequence. Thus, the methodology uses a Star graph molecular descriptor of the peptide sequences and the Machine Learning technique for the best classifier. The best model represents a Random Forest classifier based on two features of the embedded and non-embedded graphs. The performance of the model is excellent, considering similar models in the field, with an Area Under the Receiver Operating Characteristic Curve (AUROC) value of 0.938 and true positive rate (TPR) of 0.886 (test subset). The prediction of new nucleotide binding peptides with this model could be useful for drug target studies in drug development.

Keywords:
QSAR; nucleotide binding proteins; Star Graph; topological indice

1. Introduction

Nucleotides participate in different cellular processes, such as transmission of genetic information or energy transfer and storage. Therefore, they are in interaction with proteins in key cell processes[1]. The nucleotide binding proteins became important possible drug targets because they are part of signalling systems in mammalian cells, regulating systems in sensory perception, cell growth and hormonal regulation[2]. The nucleotide binding molecular function is defined by the gene ontology GO:0000166 as a selective and non-covalently interaction with a nucleotide, an (ortho) phosphate esterified nucleoside or an oligophosphate at any hydroxyl group on the ribose/deoxyribose. The child ontology terms are purine nucleotide binding, cyclic nucleotide binding, methyl-CpG binding, pyrimidine nucleotide binding, NADP binding, NAD binding, methyl-CpNpG binding, methyl-CpNpN binding, deoxyribonucleotide binding, ribonucleotide binding, flavin adenine dinucleotide binding, and ADP-D-ribose binding (http://www.ebi.ac.uk/QuickGO/).

The most known examples are the guanine nucleotide binding proteins (G proteins), the subject of a Nobel Prize. G proteins act as molecular switches inside cells and their malfunction leads to several diseases such as cancer, depression, diabetes, allergies, cardiovascular defects[3]. Due to the importance of G proteins, approximatively 30 % of the current drug targets are G Protein-coupled Receptors (GPCRs)[4].

New peptides with nucleotide binding function could be very useful as new treatment solutions. For this reason, the current paper is aimed at developing a new theoretical model that can predict new

peptides as nucleotide binders, in order to be used as *in silico* screening to reduce the number of molecules to be tested in pre-clinical experiments. This model uses the molecular information of proteins and represents Quantitative Structure Activity Relationships (QSARs)[5]. QSAR models presented in this study are based on topological indices (TIs) or molecular descriptors obtained by encoding the amino acid sequence information into Star-like graph descriptors[6]. The basic search for the best classification model that can predict a function for peptides uses the Linear Discriminant Analysis (LDA)[7] and the non-linear Artificial Neural Networks (ANNs)[8]. Several prediction models for protein biological properties based on graph/complex network molecular descriptors have been published by our group regarding transport proteins[9], lipid-binding proteins[10], cancer-related proteins[11], lectin proteins[12], cell-death proteins[13], enzyme regulatory protein[14], antioxidant proteins[15] or ATCUN DNA-cleavage proteins[16].

In this work, the authors propose the first classification QSAR model based on embedded/non-embedded Start-like graph descriptors to predict nucleotide binding proteins. This methodology, consisting of mixing the molecular information and Machine Learning techniques to obtain QSAR classifiers, has been proved successful within the above protein function models.

A series of recent publications[17] are based on Chou's 5-step rule[18], by establishing a very useful sequence-based statistical predictor for a biological system:

Construct or select a valid benchmark dataset to train and test the predictor;

Formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted;

Introduce or develop a powerful algorithm (or engine) to operate the prediction;

Properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor;

Establish a user-friendly web-server for the predictor that is accessible to the public.

It is described below how to deal with these steps one-by-one.

2. Materials and Methods

The workflow of the methodology used in the current work is presented in Figure 1. The input data source is composed of amino acid sequences (primary structure) from proteins with nucleotide (NB) and non-nucleotide binding (non-NB) properties in FASTA format. All sequences of amino acids are transformed into Star Graphs and the corresponding topological indices using S2SNet application[19]. These TIs are the input for data mining techniques from Weka software[20]. The dataset was standardized and resampled in order to balance the number of NB and non-NB cases. Eight Machine Learning techniques from Weka have been used to find the best QSAR classification model.

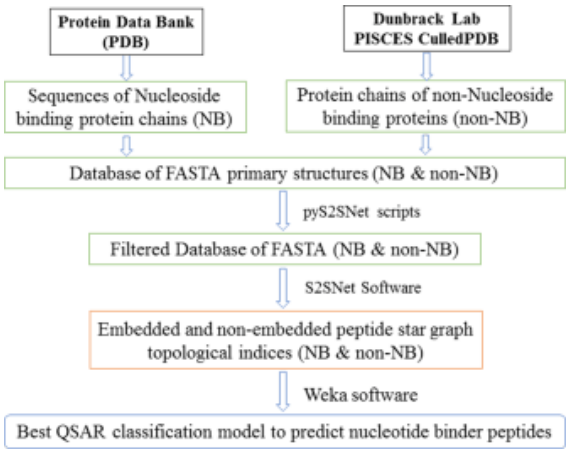

Figure 1. Flow chart of the study methodology: Machine Learning classifiers based on Star Graph descriptors of peptide amino acid sequences.

2.1. Protein Set

The datasets of this work are extracted from two protein databases. These sets of protein primary sequences are made up of 1911 proteins with nucleotide binding activity (NB) and 2333 proteins with no binding activity (non-NB). The protein FASTA sequences (positive group, NB) have been downloaded from the Protein Databank[21], the "nucleotide binding" list being obtained from the drop-down list of "carbohydrate derivative binding" from the "Molecular Function Browser" in the "Advanced Search Interface". The negative group (non-NB) was constructed from some protein sequences without nucleotide binding activity, using the PISCES CulledPDB[22] list of proteins with less than 20 % identity, resolution of 1.6 Å and R-factor of 0.25 (non-nucleotide binding proteins included, but any other possible biological function; http://dunbrack.fccc.edu/PISCES.php). Identity is the degree of correspondence between two sequences, and a value of 25 % or higher implies a similar function activity. Both sets have not been post-filtered for any source organism. The NB and non-NB lists of sequences have been transformed into S2SNet format and filtered for common sequences using pyS2SNet GitHub repository (https://github.com/muntisa/pyS2SNet). The full dataset has been resampled using Weka in order to equalize the number of positive and negative cases. The normalized dataset was the input for the Weka's Machine Learning methods.

2.2. Star Graph Topological Indices (TIs)

Each protein sequence was transformed into a Star Graph, where the vertices (nodes) are presented by each amino acid, connected in a specific sequence by peptide bonds. The Star Graph (SG) is a special type of tree with N vertices, where one has had N-1 degrees of freedom and the remaining N-1 vertices have had one single degree of freedom[23].

The current SGs have been constructed using 26 branches ("rays") of the star. Each branch contains the same amino acid type and the star centre is a non-amino acid vertex. The amino acid list (SG groups) was composed of the 20 standard amino acids (A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V), two amino acids that in some species are interpreted as stop codons (U, O) and four amino acid codes that are used when the determination of a residue is not clear (B, Z, J, X). Thus, the non-embedded SG connectivity encodes the following information on the protein primary structure: amino acid type, sequence and frequency. The embedded SGs have extra information, such as protein chain connectivity, due to the chemical peptide bonds.

The use of the graphical approaches to study biological problems can provide an intuitive picture or useful insights for contributing to the analysis of complicated relations in these systems[24], as demonstrated by many previous studies on a series of important biological topics: enzyme-catalyzed reactions[25], inhibition of HIV-1 reverse transcriptase[26], inhibition kinetics of processive nucleic acid polymerases and nucleases[27], protein folding kinetics[28], drug metabolism systems[29], and using wenxiang diagrams or graphs to study protein-protein interactions[30].

The encoding of this information of amino acid sequences into Star Graph TIs was possible using the Sequence to Star Networks (S2SNet) application[17]. For the current study, the embedded and non-embedded TIs have been calculated, without weights, with Markov normalization and power (n) of matrices/indices from 0 to 5. This power is used to simulate the interaction of the amino acids in the molecular graph at a distance of n (similar to Markov chains). The following TI types have been calculated: Shannon entropies (Sh_n), trace of the n connectivity matrices (Tr_n), Harary number (H), Wiener index (W), Gutman topological index ($S6$), Schultz topological index (non-trivial part) (S), Balaban distance connectivity index (J), Kier-Hall connectivity indices (${}^nX$, n=0,2–5), and Randic connectivity index (${}^1X$). All the formulas for these TIs are presented in Refs. [19]. The embedded TIs have an "*e*" prefix. A total of 42 TIs have been calculated. All TIs have been used to find the best classification model which predicts the nucleotide binding peptides, using Weka techniques.

2.3. Classification Methods

The following eight Weka classifiers have been used with the NB/non-NB dataset: LIBLINEAR – linear method[31], LibSVM - Support Vector Machines (SVM)[32], Multilayer Perceptrons (MLP) - a neural network technique[33], KStar[34], JRip[35], NaiveBayes – a Bayesian technique (Naïve Bayes)[36], RandomTree[37] and RandomF – Random Forest[38]. The parameters for each of the models were initialised with the default setting of Weka.

SVM introduces the key-concept of kernel, a function that provides data a higher dimensionality, which allows converting the original data from non-linearly separable to linearly separable. SVM yields very good results with high-dimensional data[39]. MLP is a non-linear Machine Learning technique and is represented by a feedforward artificial neural network model that maps a set of input onto a set of corresponding outputs. An MLP can contain multiple layers of nodes as a directed graph (each layer fully connected to the next one). Each node represents a neuron as a processing element (except for the inputs), with a non-linear activation function. The training of an MLP consists of a supervised learning technique such as backpropagation[40].

KStar algorithm differs from instance-based algorithms because it does not use a Euclidean measure approach, but it is based on an entropy distance function, extracted from the information theory[41]. JRip represents a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER). Random Tree constructs a tree that considers K randomly chosen attributes at each node. Random Forest combines many decision trees to make a prediction, giving as output the class that is the mode of the classes output by individual trees ("ensemble learning" technique by using multiple models). The main advantage of Random Forest over other techniques such as MLP or SVM is its robustness regarding solution overfitting, tending to converge always when the number of trees is too large. For all Machine Learning techniques, 10-fold cross-validation method has been used[42].

In order to choose the best two-class classifier, several well-known accuracy measures have been used: True Positive Rate (TPR), False Positive Rate (FPR), F-measure and Area Under the Receiver Operating Characteristic Curve (AUROC). The comparison of these performance measures could be found in Ref.[43]. F-measure, which considers both precision and recall equally important, is a trade-off between them[44]. The higher the precision, the less effort wasted in testing and inspection; and the higher the recall, the fewer defective modules go undetected. ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of TPR (also known as sensitivity) vs. FPR at various threshold settings. FPR is one minus the specificity or true negative rate. AUROC is considered by Jin et al.[45] to be a better measure for lassifier comparison.

3. Results and Discussion

The NB/non-NB dataset was composed of 4244 protein sequences, out of which 1911 have proved to have nucleotide binding activity (NB or positive group). The remaining 2333 proteins (non-NB or negative group) are sequences from the CulledPDB server with less than 20 % identity, without nucleotide binding activity. All protein sequences have been processed with S2SNet application[46] in order to calculate the corresponding Star Graph topological indexes. Thus, for each sequence, 42 attributes are calculated. These TIs correspond to the embedded/non-embedded protein Star Graph. The series of topological indices for each protein chain have been used to find the best nucleotide binding classification model with eight Weka linear and non-linear classifiers (10-fold cross validation).

The objective of this work is to select the technique with the highest classification score, providing a good precision and AUROC values, with the minimum number of features and using the simplest classification method. The aim of the results is to find accurate classification models for the prediction of nucleotide binding/non-nucleotide binding peptides based on Star Graph topological indexes.

In the first step, the dataset with all features has been tested with eight Weka classifiers: LIBLINEAR, LibSVM, MLP, KStar, JRip, NaiveBayes, RandomTree and RandomF (see Table 1). It can be observed that with all 42 features it is possible to obtain an NB/non-NB classification model characterized by TPR and AUROC values higher than 0.85, and FPR value lower than 0.14 (RandomTree, KStar and RandomF). KStar and RandomF have an even better AUROC value, higher than 0.94. Thus, the best model is a RandomF classifier with an AUROC value of 0.954, a TPR value of 0.896 and a FPR value of 0.104.

Table 1. NB/non-NB classification results using all 42 features.

Weka classifier	TPR	FPR	F-Measure	AUROC
LibLINEAR	0.753	0.247	0.753	0.753
LibSVM	0.752	0.248	0.752	0.752
MLP	0.753	0.247	0.753	0.821
KStar	0.868	0.132	0.868	0.946
JRip	0.759	0.241	0.759	0.793
NaiveBayes	0.687	0.313	0.667	0.799
RandomTree	0.871	0.129	0.871	0.871
RandomF	0.896	0.104	0.896	0.954

Due to the large number of features (42), in the next step a Weka feature selection technique was applied (Evaluator: *weka.attributeSelection.CfsSubsetEval -P 1 -E 1*; Search: *weka.attributeSelection.BestFirst -D 1 -N 5*) in order to obtain a reduced number of features. Therefore, only six features have been selected: H, ${}^{5}X$, eTr_3, eTr_4, eTr_5 and $e^{4}X$. The same eight Weka classifiers have been tested with only these six features and the results are shown in Table 2. RandomTree and RandomF classifiers maintained the results, but the KStar presents lower values for TPR (0.794) and AUROC (0.876). The best classifier with only six selected features is again RandomF with a TPR value of 0.896, a FPR value of 0.104 and an AUROC value of 0.952. Therefore, it was demonstrated that with less features (6 out of 42) RandomF classifier is able to obtain almost the same performance.

Table 2. NB/non-NB classification results using six selected features.

Weka classifier	TPR	FPR	F-Measure	AUROC
LibLINEAR	0.745	0.255	0.744	0.745
LibSVM	0.731	0.269	0.729	0.731
MLP	0.743	0.257	0.743	0.811
KStar	0.794	0.206	0.794	0.876
JRip	0.759	0.241	0.759	0.781
NaiveBayes	0.684	0.316	0.663	0.777
RandomTree	0.874	0.126	0.874	0.874
RandomF	0.896	0.104	0.896	0.952

In the last step, the correlation between the selected features has been tested with a Python script from pyS2SNet collection. A cutoff of 0.80 correlation has been used and only two features have passed this test: ${}^{5}X$ and eTr_5 (0.836). This new dataset with only two features have been tested with the six Weka classifiers and the results are presented in Table 3. It can be pointed out that with only two features RandomF classifier has a performance similar to the one based on six classifiers, even if the AUROC (0.938) and TPR (0.886) values are lower.

Table 3. NB/non-NB classification results using two features after the correlated, selected features were removed.

Weka classifier	TPR	FPR	F-Measure	AUROC
LibLINEAR	0.744	0.256	0.743	0.744
LibSVM	0.732	0.268	0.729	0.732
MLP	0.744	0.256	0.744	0.813
KStar	0.754	0.246	0.754	0.83
JRip	0.750	0.250	0.750	0.779
NaiveBayes	0.678	0.322	0.654	0.771
RandomTree	0.868	0.132	0.868	0.868
RandomF	0.886	0.114	0.886	0.938

Using these three datasets with all the features, with only selected features and removing the correlated selected features it was demonstrated the power of Random Forest technique for the NB/non-NB classification and the codification of the essential peptide sequence information into only two features such as X_5 (non-embedded Kier-Hall descriptor, power 5) and eTr_5 (embedded Trace of the connectivity matrix, power 5). The power 5 may indicate the importance of the interaction of the amino acids at a distance of five nodes in the molecular star graph for this type of classification.

As demonstrated in a series of recent publications (see, e.g., [2a,c,e,47]) on developing new prediction methods, user-friendly and publicly accessible web-servers will significantly enhance their impact[48], further work will provide tools for designing a web-server for the prediction method presented in this paper.

4. Conclusions

This study proposes the first model designed to identify proteins that have nucleotide binding activity, using Star Graph TIs obtained from protein amino acid sequences. The proposed model, based on only two attributes extracted from the embedded and non-embedded graph, shows good predictive capacity with the Random Forest technique, obtaining an AUROC value of 0.938 and a TPR value of 0.886. This study confirms the classification power of the mixture of Machine Learning classifiers with molecular Star graph descriptors. The current results may be used to predict new nucleotide binding peptides for future drug development.

Conflict of Interest

None declared.

Acknowledgements

Yong Liu acknowledges the Mexican Council for Science and Technology (CONACyT) for financial support for Ph.D. studies. This work was supported by the "Galician Network for Colorectal Cancer Research (REGICC)" (Ref. R2014/039), funded by Xunta de Galicia, by the "Collaborative Project on Medical Informatics (CIMED)" PI13/00280, funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013–2016, and the European Regional Development Funds (FEDER). The authors acknowledge the support offered by the Galician Network of Drugs R+D REGID (Xunta de Galicia R2014/025).

References

1. L. Parca, P. F. Gherardini, M. Truglio, I. Mangone, F. Ferre, M. Helmer-Citterich, G. Ausiello, PLoS ONE 2012, 7, e50240.
2. D. J. Roberts, M. Waelbroeck, Biochem. Pharmacol. 2004, 68, 799–806.
3. T. Schoneberg, A. Schulz, H. Biebermann, T. Hermsdorf, H. Rompler, K. Sangkuhl, Pharmacol. Ther. 2004, 104, 173–206.
4. D. E. Bosch, D. P. Siderovski, Exp. Mol. Med. 2013, 45, e15.
5.
5a Q. S. Du, R. B. Huang, K. C. Chou, Curr. Protein Pept. Sci. 2008, 9, 248–260.
5b Q. S. Du, R. B. Huang, Y. T. Wei, Z. W. Pang, L. Q. Du, K. C. Chou, J. Comput. Chem. 2009, 30, 295–304.
5c F. J. Prado-Prado, H. Gonzalez-Diaz, O. M. de la Vega, F. M. Ubeira, K. C. Chou, Bioorg. Med. Chem. 2008, 16, 5871–5880.
5d H. Wei, C. H. Wang, Q. S. Du, J. Meng, K. C. Chou, Med. Chem. 2009, 5, 305–317.
5e J. Devillers, A. T. Balaban, Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach, The Netherlands, 1999.
6. M. Randić, J. Zupan, D. Vikic-Topic, J. Mol. Graph. Model. 2007, 26, 290–305.
7.
7a H. Van Waterbeemd, in Chemometric methods in molecular design, Vol. 2 (Ed.: H. Van Waterbeemd), Wiley-VCH, New York, 1995, pp. 265–282
7b J. H. Friedman, J. Am. Stat. Assoc. 1989, 84, 165–175.
8. D. Rivero, E. Fernandez-Blanco, J. Dorado, A. Pazos, in Evolutionary Computation (CEC), 2011 IEEE Congress on, IEEE, 2011, pp. 587–592.
9. C. Fernandez-Lozano, M. Gestal, N. Pedreira-Souto, L. Postelnicu, J. Dorado, C. R. Munteanu, Curr. Top. Med. Chem. 2013, 13, 1681–1691.
10. H. Gonzalez-Diaz, C. R. Munteanu, L. Postelnicu, F. Prado-Prado, M. Gestal, A. Pazos, Mol. BioSyst. 2012, 8, 851–862.
11. C. R. Munteanu, A. L. Magalhaes, E. Uriarte, H. Gonzalez-Diaz, J. Theor. Biol. 2009, 257, 303–311.
12. C. R. Munteanu, N. Pedreira-Souto, J. Dorado, A. Pazos, L. G. Pérez-Montoto, F. M. Ubeira, H. González-Díaz, Molecular Informatics 2014, 33, 276–285.
13. C. Fernandez-Lozano, M. Gestal, H. Gonzalez-Diaz, J. Dorado, A. Pazos, C. R. Munteanu, J. Theor. Biol. 2014, 349, 12–21.
14. C. Fernandez-Lozano, E. Fernandez-Blanco, K. Dave, N. Pedreira, M. Gestal, J. Dorado, C. R. Munteanu, Mol. BioSyst. 2014, 10, 1063–1071.
15. E. Fernandez-Blanco, V. Aguiar-Pulido, C. R. Munteanu, J. Dorado, J. Theor. Biol. 2012, 317, 331–337.
16. C. R. Munteanu, J. M. Vazquez, J. Dorado, A. P. Sierra, A. Sanchez-Gonzalez, F. J. Prado-Prado, H. Gonzalez-Diaz, J. Proteome Res. 2009, 8, 5219–5228.
17.
17a S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, W. Chen, K. C. Chou, Bioinformatics 2014, 30, 1522–1529.
17b Y. Xu, X. Wen, L. S. Wen, L. Y. Wu, N. Y. Deng, K. C. Chou, PLoS ONE 2014, 9, e105018.
17c H. Lin, E. Z. Deng, H. Ding, W. Chen, K. C. Chou, Nucleic Acids Res. 2014, 42, 12961–12972;
17d B. Liu, L. Fang, F. Liu, X. Wang, J. Chen, K. C. Chou, PLoS ONE 2015, 10, e0121501.
17e J. Jia, Z. Liu, X. Xiao, B. Liu, K. C. Chou, J. Theor. Biol. 2015, 377, 47–56.
18. K. C. Chou, J. Theor. Biol. 2011, 273, 236–247.
19. C. R. Munteanu, A. L. Magalhaes, A. Duardo-Sanchez, A. Pazos, H. Gonzalez-Diaz, Curr. Bioinf. 2013, 8, 429–437.
20. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. A. Witten, SIGKDD EXPLOR. NEWSL. 2009, 11, 10–18.
21. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, Nucleic Acids Res. 2000, 28, 235–242.
22. G. Wang, J. R. L. Dunbrack, Bioinformatics 2003, 19, 1589–1591.
23. F. Harary, Graph Theory, Reading, MA, 1969.. 24S. X. Lin, J. Lapointe, J. Biomed. Sci. Eng. 2013, 6, 435–442.
25.
25a K. C. Chou, S. Forsen, J. Biochem. 1980, 187, 829–835.
25b G. P. Zhou, M. H. Deng, J. Biochem. 1984, 222, 169–176.
25c K. C. Chou, J. Biol. Chem. 1989, 264, 12074–12079.
26. I. W. Althaus, J. J. Chou, A. J. Gonzales, M. R. Diebel, K. C. Chou, F. J. Kezdy, D. L. Romero, P. A. Aristoff, W. G. Tarpley, F. Reusser, Biochemistry (Mosc.) 1993, 32, 6548–6554.
27. K. C. Chou, F. J. Kezdy, F. Reusser, Anal. Biochem. 1994, 221, 217–230.
28. K. C. Chou, Biophys. Chem. 1990, 35, 1–24.
29. K. C. Chou, Curr. Drug Metab. 2010, 11, 369–378.
30.
30a G. P. Zhou, J. Theor. Biol. 2011, 284, 142–148.
30b G. P. Zhou, R. B. Huang, Curr. Top. Med. Chem. 2013, 13, 1152–1163.
31. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, J. Mach. Learn. Res. 2008, 9, 1871–1874.
32. V.N. Vapnik, Nauka, English Translation Springer Verlang, 1982, 1979.
33.

33a F. Rosenblatt, Principles of neurodynamics; perceptrons and the theory of brain mechanisms, Spartan Books, Washington, 1962.
33b F. Rosenblatt, Psychol. Rev. 1958, 65, 386–408..
34. J. G. Cleary, L. E. Trigg, in Machine Learning International Workshop, Morgan Kaufmann Plublishers, Inc., 1995, pp. 108–114.
35. W. W. Cohen, in Twelfth International Conference on Machine Learning, 1995, pp. 115–123.
36. G. H. John, P. Langley, in 11th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufman, Montreal, Quebec, 1995, pp.338–345.
37. C. M. Bishop, Pattern recognition and machine learning, Springer, 2006.
38. L. Breiman, Machine Learning 2001, 45, 5–32.
39.
39a O. Chapelle, P. Haffner, V. N. Vapnik, IEEE Trans. Neural Netw. 1999, 10, 1055–1064.
39b L. S. Moulin, A. P. Alves Da Silva, M. A. El-Sharkawi, R. J. Marks, IEEE T. POWER SYST. 2004, 19, 818–825.
40. D.E. Rumelhart, G. E. Hinton, R. J. Williams, DTIC Document, 1985.
41.
41a C. E. Shannon, W. Weaver, R. E. Blahut, B. Hajek, The mathematical theory of communication, Vol. 117, University of Illinois press Urbana, 1949.
41b D. J. MacKay, Information theory, inference and learning algorithms, Cambridge university press, 2003.
42. G. J. McLachlan, K.-A. Do, C. Ambroise, Analyzing microarray gene expression data, Wiley, 2004.
43. C. Ferri, J. Hernandez-Orallo, R. Modroiu, Pattern Recogn. Lett. 2009, 30, 27–38.
44. I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems), Morgan Kaufmann, 2005.
45. H. Jin, IEEE Trans. Knowl. Data Eng. 2005, 17, 299–310.
46. C. R. Munteanu, A. L. Magalhães, E. Uriarte, H. González-Díaz, J. Theor. Biol. 2009, 257, 303–311.
47.
47a W. Chen, P. M. Feng, E. Z. Deng, H. Lin, K. C. Chou, Anal. Biochem. 2014, 462, 76–83.
47b Z. Liu, X. Xiao, W. R. Qiu, K. C. Chou, Anal. Biochem. 2015, 474, 69–77.
47c B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, K.-C. Chou, Nucleic Acids Res. 2015, doi:10.1093/nar/gkv458.
48. K. C. Chou, Med. Chem. 2015, 11, 218–234.