TWO-LAYER SVM CLASSIFIER FOR REMOTE PROTEIN HOMOLOGY DETECTION AND FOLD RECOGNITION

MOHD HILMI MUDA

UNIVERSITI TEKNOLOGI MALAYSIA

TWO-LAYER SVM CLASSIFIER FOR REMOTE PROTEIN HOMOLOGY DETECTION AND FOLD RECOGNITION

MOHD HILMI MUDA

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

November 2009

To my beloved mother, my late father, and my siblings

ACKNOWLEDGEMENTS

I would like to express my true and sincere thanks and gratitude to my supervisor, Assoc. Prof. Dr. Puteh Saad for her patience, guidance, encouragement, invaluable comments, and advice that made this research possible and completed early.

I would like to thank all members of the Laboratory of Computational Intelligence and Biology (LCIB) for their continuous support in many aspects of this research especially Dr. Muhamad Razib Othman. Not forget my English lecturer, Mrs Hanizah Ishak, thanks a lot for proofread my first draft.

Finally, my deepest thanks go to my parent. Their influence made me realize the importance of education from a very early age. I cannot give enough thanks to my parent for the great love and support that they has been giving me throughout my life.

ABSTRACT

Advances in molecular biology in the past years have yielded an unprecedented amount of new protein sequences. The resulting sequences describe a protein in terms of the amino acids that constitute them without structural or functional protein information. Therefore, remote protein homology detection and fold recognition algorithms have become increasingly important to detect the structural homology in proteins where there are small or no similarity at all in the sequences compared. However, it is a challenging task to detect and classify this similarity with more biological meaning in the context of Structural Classification of Proteins (SCOP) database. This study presents a new computational framework based on two-layer SVM classifier that uses protein sequences as a primary source. The first layer is used to detect up to superfamily level in the SCOP hierarchy using one-versus-all SVM binary classifiers and the Bio-kernel function. The second layer uses SVM with fold recognition codes and the profile-string kernel to leverage the unlabeled data and to detect up to fold level in the SCOP hierarchy. The proposed framework is tested using SCOP 1.53, 1.67 and 1.73 datasets and the results are evaluated using mean Receiver Operating Characteristics (ROC) and mean Median Rate of False Positives (MRFP). In terms of mean ROC, the experiment shows 4.19% improvement in SCOP 1.53 dataset, 4.75% in SCOP 1.67 dataset and 4.03% in SCOP 1.73 dataset compared to the existing SVM-based classifiers and kernel functions. This result shows that the proposed framework is capable to perform well using different versions of datasets and has outperformed existing methods, which implies the reliability of the framework.

ABSTRAK

Kemajuan dalam bidang biologi molekul kebelakangan ini telah menghasilkan banyak jujukan protein yang baru. Jujukan yang dihasilkan terdiri daripada asid amino dan tidak mengandungi maklumat struktur dan fungsi bagi protein tersebut. Justeru, pengesanan homologi protein yang jauh dan pengecaman lipatan protein menjadi keperluan yang penting bagi mengesan struktur homologi sesuatu protein di mana wujud sedikit persamaan atau tiada persamaan di dalam jujukan protein tersebut. Walau bagaimanapun, adalah satu tugas yang mencabar untuk mengesan dan mengelaskan protein dengan menggabungkan maklumat biologi iaitu evolusi protein berdasarkan kepada hierarki Struktur Pengelasan Protein (SCOP). Kajian ini mencadangkan satu rangka kerja yang baru berasaskan kepada pengelasan dua-lapisan Mesin Sokongan Vektor (SVM) yang menggunakan jujukan protein sebagai sumber utama. Lapisan pertama berfungsi untuk mengesan hingga ke peringkat superfamili berasaskan hierarki SCOP menggunakan pengelasan binari dan fungsi *Kernel-Bio*. Lapisan kedua pula menggunakan SVM bersama kod pengecaman lipatan dan profil rentetan kernel bagi mengurangkan data yang tidak berlabel dan mengesan sehingga peringkat lipatan protein dalam hierarki SCOP. Rangka kerja ini diuji menggunakan set data SCOP 1.53, 1.67 dan 1.73 serta hasilnya dinilai menggunakan purata Penerima Operator Karakter (ROC) dan purata Positif Palsu Berkadar Median (MRFP). Keputusan yang diperolehi menunjukkan peningkatan 4.19% pada SCOP 1.53, 4.75% pada SCOP 1.67 dan 4.03% pada SCOP 1.73 berbanding kaedah-kaedah SVM dan fungsi kernel yang lain. Ini menunjukkan rangka kerja yang dicadangkan menghasilkan prestasi yang lebih baik daripada kaedah lain menggunakan versi data yang berbeza-beza, yang mana ini membuktikan kebolehpercayaan rangka kerja yang dicadangkan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3D	-	Three-Dimensional
AIDS	-	Acquired Immune Deficiency Syndrome
CPU	-	Central Processing Unit
DNA	-	Deoxyribonucleic Acid
FDA	-	Food and Drug Administration
FN	-	False Negative
FP	-	False Positive
HMM	-	Hidden Markov Model
HIV	-	Human Immunodeficiency Virus
kDa	-	Kilo Dalton
kNN	-	**k-Nearest Neighbor**
MRFP	-	Median Rate of False Positives
NB	-	**Naive Bayesian**
NIC	-	Network Interface Card
NMR	-	Nuclear Magnetic Resonance
OSH	-	Optimal Separating Hyperplane
NN	-	Neural Network
PC	-	Personal Computer
PDB	-	Protein Data Bank
RAM	-	Random Access Memory
RBF	-	Radial Basis Function
ROC	-	Receiver Operating Character
RQA	-	Recurrence Quantitative Analysis
SCOP	-	Structural Classification of Proteins

SuSE	-	Software und System Entwicklung
SVM	-	Support Vector Machines
SW	-	Smith-Waterman
TN	-	True Negative
TP	-	True Positive
TPR	-	True Positive Rate
FPR	-	False Positive Rate
WCM	-	Word Correlation Matrices

CHAPTER 1

INTRODUCTION

1.1 Background

Advances in molecular biology in past years like large-scale sequencing and the human genome project, have yielded an unprecedented amount of new protein sequences. The resulting sequences describe a protein in terms of the amino acids that constitute it and no structural or functional protein information is available at this stage. To a degree, this information can be inferred by finding a relationship or known as homology between new sequences and proteins for which structural properties are already known. Protein classification is the prediction of a protein’s structural class from its primary sequence of amino acids. This prediction problem is fundamental in computational biology for a number of reasons. First, a protein’s structure is closely linked to its biological function, so knowledge of the structural category can allow the improvement of the prediction of protein function. Moreover, experimental methods for determining the full three dimensional (3D) structure of a protein such as traditional laboratory methods of protein homology detection depend on lengthy and expensive procedures like X-Ray Crystallography and Nuclear Magnetic Resonance (NMR). Second, prediction of a protein sequence’s structural class enables the selection of a template structure from the protein database, which can then be used with various comparative modeling techniques to predict a full 3D structure for the protein sequence. Predicted structures are important for more detailed biochemical analysis and in particular for drug design (Heather and McCammon, 2000).

Since using these procedures is unpractical for the amount of data available, researchers are increasingly relying on computational techniques to automate the process. Accurately detecting homologs at low levels of sequence similarity or known as remote homology detection still remains a challenging problem to biologists. Remote protein homology detection refers to detection of structural homology in proteins where there are small or no similarity in the sequence. The remote protein homology detection is a classic problem and it aims to identify for a given protein or protein family from a large database of sequences, all distantly protein sequences are related. The principal idea behind homology is based on evolution; proteins that belong to the same family have evolutionary pressure to retain common regions associated with their biochemical function and maintenance of 3D fold. Fold recognition on the other hand is a key step in the protein structure discovery process, especially when traditional protein sequence comparison methods fail to yield convincing structural homologies. Although many methods have been developed for protein fold recognition, their accuracies remain low. This can be attributed to insufficient exploitation of fold discriminatory features.

To detect protein structural classes from protein primary sequence information, homology-based methods have been developed, which can be divided to three types: discriminative classifiers (Jaakkola *et al.*, 2000), generative models for protein families (Krogh *et al.*, 1994) and pairwise sequence comparisons (Altschul *et al.*, 1990). Discriminative classifiers show superior performance when compared to other methods (Altschul *et al.,* 1990). On the other hand, classical approach in fold recognition can be divided to four approaches: sequence-sequence alignment methods (Thompson *et al.*, 1994), sequence profile alignment method (Eddy, 1998), profile-profile alignment method (Sadreyev and Grishin, 2003) and sequence structure method (Xu *et al.*, 2003).

The following section will describe details about the challenges that arise in remote protein homology detection and fold recognition, followed by a brief review of the current methods used in remote protein homology detection and fold recognition. The problem statement, objectives, significance and scope of the study will also be presented. The aim of this study is to improve the existing method of remote protein homology detection and extend it to fold recognition.

1.2 Methods for Detecting Remote Protein Homology and Fold Recognition

Basically, remote protein homology detection can be divided into three categories (the details are described in Chapter 2):

(i) Generative model is used to extract the feature vectors, involves building a model for a single protein family and then evaluating each candidate sequence to see how well it fits the model. If the fit of the sequence is above some threshold value, then the protein is classified as belonging to the family. Examples of related works are Latent Semantic Analysis (LSA: Dong *et al.*, 2006) and Hidden Markov Model (HMM: Bernardes *et al.*, 2007).

(ii) Pairwise sequence comparison model is used to arrange the primary sequences of protein in order to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Examples of related works are FORCE (Wittkop *et al.*, 2007) and PDBalert (Agarwal *et al.*, 2008).

(iii) Discriminative classifier model is able to discriminate all the protein sequences into positive (label) and negative (unlabelled) members. It is easier to extend and deal with multiple object classes and to update with new training data. Example algorithms are Neural Network (NN:

Hochreiter *et al.*, 2007) and Support Vector Machine (SVM: Rangwala and Karypis, 2005).

The current methods for fold recognition on the other hand can be divided to four main approaches as below:

(i) Sequence-sequence alignment methods are effective at detecting homologs with significant sequence identity (>40%). Examples of related tools are CLUSTALW (Thompson *et al.,* 1994) and PALIGN (Ohlson *et al.*, 2004).

(ii) Sequence profile are more sensitive at detecting distant homologs with lower sequence identity (>20%). Examples of related tools are HMMER-hhmsearch (Eddy, 1998) and IMPALA (Schäffer *et al.*, 1999).

(iii) Profile-profile are more sensitive at detecting distant homologs and comparable structures, and often achieve even better performance than sequence-structure alignment methods that leverage template structural information (Rychlewski *et al.*, 2000). Examples of related tools are COMPASS (Sadreyev and Grishin, 2003) and HHSearch (So¨ding, 2005).

(iv) Sequence-structure alignment methods (or threading) align query sequences with template structures and compute compatibility scores according to structural environment fitness and contact potentials. These methods are particularly useful for detecting proteins with similar folds but no recognizable evolutionary relationship. Examples of the related work are Fralanyzer (Harpreet and Daniel, 2007) and Prospect II (Kim *et al.*, 2003).

1.3 Statement of the Problem

The remote protein homology detection and fold recognition problem to be solved in this study can be described as follows:

"Given protein sequences, the main problem is to classify the protein sequences into various levels of SCOP hierarchy and at the same time incorporates biological information in the classification using kernel function. The classification should provide higher mean Receiver Operating Character (ROC) and lower mean Median Rate of False Positives (MRFP), which indicate lower misclassification."

In order to solve the remote protein homology detection and fold recognition problems, two factors need to be considered. First, the holistic detection multi-layer classification should be able to detect not only family and superfamily, but also fold in Structural Classification of Proteins (SCOP: Andreeva *et al.*, 2008) hierarchy. On the other hand, different kernel functions will give different result in remote protein homology detection and fold recognition. Therefore, kernel functions that consider all optimal local alignments score with gaps between all of their possible subsequences is the second factor to be studied. Thus, it will provide more accurate results of remote protein homology detection and fold recognition.

1.4 Challenges of Detecting Remote Protein Homology and Fold Recognition

The underlying protein classification problem is in fact a huge multi-layer problem, with over 1,000 protein folds and even more structural subcategories organized into a hierarchy. Even though highly accurate SVM-based binary classifiers can go a long way in addressing some of the biologist's requirements, it is still unknown how to best combine the predictions of a set of SVM-based binary classifiers to solve the multi-layer classification problem and assign a protein sequence to a particular family, superfamily or fold precisely. Moreover, it is not clear whether method that combine binary classifiers are inherently better suited for solving the remote homology detection and fold recognition problems over method that directly builds an SVM-based multi-layer classification model. Some proteins have a very similar structure but do not share significant sequence similarity. Meanwhile, some unrelated protein sequences do not share any structural similarity yet their protein sequences have a high similarity. Based on that, our first challenge that arises is on how to make an accurate holistic detection of remote protein homology and fold recognition in the context of SCOP protein classification, as the SCOP provides a comprehensive and detailed description of the evolutionary and structural relationships of the proteins of known structure. Within the SCOP classification, the problem of remote homology detection corresponds to the detection of superfamily of a particular protein sequence under the constraint that the protein is not similar to any of it descendant families. Whereas, the problem of fold recognition corresponds to that of predicting the fold under the constraint that the protein is not similar to any of it descendant superfamilies.

A core component of an SVM is the kernel function. The kernel function can be thought as a measure of similarity between sequences (Saigo *et al.*, 2004; Rangwala and Karypis, 2005). Different result performance will be achieved as different kernels correspond to different notions of similarity. Alignment score between sequences provides a relevant measure of similarity between protein sequences which incorporates biological information about the protein evolution. Thus, our next challenge is to incorporate biological information by implementing the kernel function which takes into account all local alignments.

1.5 Objectives of the Study

The goal of this study is to develop a framework to detect remote protein homology and recognize fold. In order to achieve the goal, several objectives need to be accomplished:

(i) To design a framework in order to detect remote protein homology and fold recognition using SVM based classifier.

(ii) To develop algorithm named SVM-2L, which based on multi-layer classification using SVM to predict and classify accurately the protein to various levels of protein group based on SCOP hierarchy.

(iii) To develop BioSVM-2L algorithm, by improving SVM-2L with local alignment kernel function for the SVM in order to incorporate the biological information in the classification process.

(iv) To test the stability of the algorithms using three different versions of SCOP datasets (1.53, 1.67 and 1.73).

1.6 Significance and Scope of the Study

Protein homology refers to homology between different proteins, that is the proteins are derived from a common “ancestor”. The proteins may be in different species, with the ancestral protein being the form of the protein that existed in the ancestral species (orthology). Or the proteins may be in the same species, but have evolved from a single protein whose gene was duplicated in the genome (parology). The complete repository of known protein structures, deposited in the Protein Data Bank (PDB: Berman *et al.*, 2002), contains just 27,000 structures, while there are about 1.5 million protein sequences in the Non-redundant Database (Pruitt *et al.*, 2007) of protein sequences. Therefore, the classification of each protein to the right class is an essential task and there is a need to use the accurate method. Therefore, this study attempts to predict and classify the protein accurately to various levels according to SCOP database. This will contribute to the discovery of new type of proteins that can be useful in biological field. Meanwhile, fold recognition methods are widely used and effective because it is believed that there are a strictly limited number of different protein folds in nature, mostly as a result of evolution but also due to constraints imposed by the basic physics and chemistry of polypeptide chains. Therefore, a good chance (currently 70-80%) that a protein which has a similar fold to the target protein has already been studied by X-ray crystallography or NMR spectroscopy and can be found in the PDB. Currently there are just over 1,100 different protein folds known. A protein’s structure is closely linked to its biological function, so knowledge of the structural category can allow improved prediction of protein’s function. On the other hand, accurate detection of remote protein homology and fold recognition can be used to design a new drug as medicine (Carlson and McCammon, 2000), can gain more knowledge to find the cure for deadly diseases like pancreatic cancer (Honda *et al.*, 2005a) and development of anti-Human immunodeficiency virus (HIV) drugs (Kliger *et al.*, 2000).

In this study, we select the protein dataset from the SCOP database which is a manually inspected database of protein as the data to this study. We limit our scope of datasets to three different versions: 1.53, 1.67 and 1.73. The scopes of capabilities are as follows: (i) the first layer implements the alternative structural formulation of the SVM optimization problem for conventional binary classification (Thorsten, 2006a); and (ii) the second layer applies the fold recognition code to learn the optimal weight of the classifier to fit into the training dataset. By combining SVM-based binary classifiers with fold recognition problem, we are able to create two-layers classifier which is capable to detect the protein up to fold level. Besides that, we used the local alignment kernel in the first layer which shows the best detection performance on widely-used homology detection setups (Lingner and Meinicke, 2008) to measure the similarities between the protein sequences. This can be done by taking into account all the optimal local alignment scores with gaps between all the possible sequences. The performances of the proposed two-layer classifier are measured by mean ROC and mean MRFP.

1.7 Organization of the Thesis

A general content description of the subsequent chapters in this thesis is given as follows:

(i) Chapter 1 describes the challenges, current methods, statement of the problems, objectives, scope and significance of the study.

(ii) In Chapter 2, we present the basic concept in remote protein homology detection and fold recognition and followed by concise description regarding classification algorithm used. Exhaustive review of the previous related work is also presented.

(iii) Chapter 3 begins with a brief review of the proposed computational framework. This will be followed by detailed description for all instruments involved, such as hardware and software requirements, testing and analysis and performance measurement.

(iv) Chapter 4 describes the SVM-2L algorithm, which is the multi-layer classification using SVM to detect remote protein homology and recognize protein fold.

(v) Chapter 5 describes the BioSVM-2L, the algorithm that combined the fold recognition codes and the Bio-kernel function which incorporated the biological information in the classification of remote homology and fold recognition.

(vi) In Chapter 6, the conclusion of the study and the achieved results to date is described. The contributions and future works of the study are also described.

REFERENCES

Agarwal, V., Remmert, M., Biegert, A. and Soding, J. (2008). PDBalert: Automatic, Recurrent Remote Homology Tracking and Protein Structure Prediction, *BMC Structural Biology*, 8(1): 51-57.

Al-Lazikani, B., Sheinerman, F. B. and Honig, B. (2001). Combining Multiple Structure and Sequence Alignments to Improve Sequence Detection and Alignment: Application to the SH2 Domains of Janus Kinases, *PNAS*, 98(26): 14796-14801.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic Local Alignment Search Tool, *Journal of Molecular Biology*, 215(3): 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs, *Nucleic Acids Research*, 25(17): 3389-3402.

Andreas, O., Axel, P., Michael, F. and Peter, A. (2006). Generic Object Recognition With Boosting, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3): 416-431.

Andreeva, A., Howorth, D., Chandonia, J., Brenner, S., Hubbard, T., Chothia, C. and Murzin, A. (2008). Data Growth and its Impact on the SCOP Database: New Developments, *Nucleic Acids Research*, 36(1): 419-425.

Andrew, T. S. and Charles, E. (2007). Making Generative Classifiers Robust to Selection Bias. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 12-15 August. San Jose, California, USA: ACM, 657-666.

Bailey, T. and Gribskov, M. (1997). Score Distributions for Simultaneous Matching to Multiple Motifs, *Journal of Molecular Biology*, 4(1): 45-59.

Baldi, P., Chauvin, Y., Hunkapiller, T. and McClure, M. A. (1994). Hidden Markov Models of Biological Primary Sequence Information, *PNAS*, 91(3): 1059-1063.

Ben-Hur, A. and Brutlag, D. (2003). Remote Homology Detection: A Motif Based Approach, *Bioinformatics*, 19(1): 26-33.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. and Zardecki, C. (2002). The Protein Data Bank, *Acta Crystallogr D Biol Crystallogr*, 58 (1): 899-907.

Bernardes, J., Dávila, A., Costa, V. and Zaverucha, G. (2007). Improving Model Construction of Profile HMMs for Remote Homology Detection through Structural Alignment, *BMC Bioinformatics*, 8(1): 435.

Björn, W., Huisheng, F., Tomas, O., Johannes, F.-S. and Arne, E. (2004). Using Evolutionary Information for the Query and Target Improves Fold Recognition, *Proteins: Structure, Function, and Bioinformatics*, 54(2): 342-350.

Boisvert, S., Marchand, M., Laviolette, F. and Corbeil, J. (2008). HIV-1 Coreceptor Usage Prediction without Multiple Alignments: An Application of String Kernels, *Retrovirology*, 5(11): 1-14.

Bowie, J. U., Luthy, R. and Eisenberg, D. (1991). A Method to Identify Protein Sequences that Fold into a Known Three-dimensional Structure, *Science*, 253(5016): 164-170.

Bramer, M. (2007). Using Decision Trees for Classification. In Bramer, M. (Ed.) *Principles of Data Mining*. (pp. 41-50). London: Springer-Verlag.

Busuttil, S., Abela, J. and Pace, G. J. (2004). Support Vector Machines with Profile-Based Kernels for Remote Protein Homology Detection, *Genome Informatics*, 15(2): 191-200.

Carlson, H. A. and McCammon, J. A. (2000). Accommodating Protein Flexibility in Computational Drug Design, *Molecular Pharmacology*, 57(2): 213-218.

Chandonia, J.-M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M. and Brenner, S. E. (2004). The ASTRAL Compendium in 2004, *Nucleic Acids Research*, 32(Suppl 1): D189-D192.

Chinnasamy, A., Sung, W. K. and Mittal, A. (2005). Protein Homology Structure and Fold Prediction using Tree-Augmented Naïve Bayesian Classifier, *Journal of Bioinformatics and Computational Biology*, 3(4): 803-819.

Corinna, C. and Vladimir, V. (1995). Support-Vector Networks, *Machine Learning*, 20(3): 273-297.

Damoulas, T. and Girolami, M. A. (2008). Probabilistic Multi-class Multi-kernel Learning: On Protein Fold Recognition and Remote Homology Detection, *Bioinformatics*, 24(10): 1264-1270.

David, R., Korenberg, M. J. and Hunter, I. W. (2000). 3D-1D Threading Methods for Protein Fold Recognition, *Pharmacogenomics*, 1(4): 445-455.

Dayhoff, M. O., Barker, W. C. and Hunt, L. T. (1983). Establishing Homologies in Protein Sequences, *Methods in Enzymology*, 91(1): 524-545.

Demichelis, F., Magni, P., Piergiorgi, P., Rubin, M. and Bellazzi, R. (2006). A Hierarchical Naive Bayes Model for Handling Sample Heterogeneity in Classification Problems: An Application to Tissue Microarrays, *BMC Bioinformatics*, 7(1): 514.

Ding, C. H. Q. and Dubchak, I., (2001). Multi-class Protein Fold Recognition Using Support Vector Machines and Neural Networks, *Bioinformatics*, 17(4): 349-358.

Diplaris, S., Tsoumakas, G., Mitkas, P. and Vlahavas, I. (2005). Protein Classification with Multiple Algorithms. In Bozanis, P. (Ed.) *Advances in Informatics* (pp. 448-456). London: Springer Berlin.

Domingues, F. S., Lackner, P., Andreeva, A. and Sippl, M. J. (2000). Structure-based Evaluation of Sequence Comparison and Fold Recognition Alignment Accuracy, *Journal of Molecular Biology*, 297(4): 1003-1013.

Dong, Q. W., Lin, L., Wang, X. L. and Li, M. H. (2005). A Pattern-Based SVM for Remote Homology Detection. *Proceedings of the 4th International Conference on Machine Learning and Cybernetics.* 18-21 August. Guangzhou, China: IEEE, 3363-3368.

Dong, Q. W., Wang, X. L. and Lin, L. (2006). Application of Latent Semantic Analysis to Protein Remote Homology Detection, *Bioinformatics*, 22(3): 285-290.

Eddy, S., (1998). Profile Hidden Markov Models, *Bioinformatics*, 14(9): 755-763.

Eddy, S. R., (2004). Where Did the BLOSUM62 Alignment Score Matrix Come

From?, *Nature Biotechnology*, 22(8): 1035-19036.

Edgar, R. C. and Sjolander, K. (2003). SATCHMO: Sequence Alignment and Tree Construction Using Hidden Markov Models, *Bioinformatics*, 19(11): 1404-1411.

Elofsson, A., Fischer, D., Rice, D. W., Le Grand, S. M. and Eisenberg, D. (1996). A Study of Combined Structure/Sequence Profiles, *Folding and Design*, 1(6): 451-461.

Enright, A. J., Van, D. S. and Ouzounis, C. A. (2002). An Efficient Algorithm for Large-Scale Detection of Protein Families, *Nucleic Acids Research*, 30(7): 1575-1584.

Enrique, A., Peter, K., Thomas, E. and Raymond, C. S. (2000). Automation of X-ray Crystallography, *Nature Structural Biology,* 7(Suppl 1): 7973-977.

Eugene, I., Jason, W., William Stafford, N. and Christina, L. (2005). Multi-class Protein Fold Recognition Using Adaptive Codes. *Proceedings of the 22nd International Conference on Machine Learning*. 11-15 August. Bonn, Germany: ACM, 329-336.

Feng, K.-Y., Cai, Y.-D. and Chou, K.-C. (2005). Boosting Classifier for Predicting Protein Domain Structural Class, *Biochemical and Biophysical Research Communications*, 334(1): 213-217.

Fitch, W. M. (1970). Distinguishing Homologous from Analogous Proteins, *Systematic Zoology*, 19(2): 99-113.

Ganyun, L. V., Haozhong, C., Haibao, Z. and Lixin, D. (2005). Fault Diagnosis of Power Transformer Based on Multi-layer SVM Classifier, *Electric Power Systems Research*, 74(1): 1-7.

Gao, D. and Yang, Y. (2005) Fuzzily Modular Multilayer Perceptron Classifiers for Large-Scale Learning Problems. *Proceedings of the 14th International Conference on Fuzzy Systems*. 22-25 May. Nevada, USA: IEEE, 625-630.

Ginalski, K., Pas, J., Wyrwicz, L. S., Grotthuss, M. V., Bujnicki, J. M. and Rychlewski, L. (2003). ORFeus: Detection of Distant Homology Using Sequence Profiles and Predicted Secondary Structure, *Nucleic Acids Research*, 31(13): 3804-3807.

Girolami, M. (2002). Mercer Kernel-Based Clustering in Feature Space, *IEEE Transactions on Neural Networks*, 13(3): 780-784.

Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001). Assignment of

Homology to Genome Sequences Using a Library of Hidden Markov Models that Represent All Proteins of Known Structure, *Journal of Molecular Biology*, 313(4): 903-919.

Griffiths-Jones, S. and Bateman, A. (2002). The Use of Structure Information to Increase Alignment Accuracy Does Not Aid Homologue Detection with Profile HMMs, *Bioinformatics*, 18(9): 1243-1249.

Grigorios, T., Lefteris, A. and Ioannis, V. (2005). Selective Fusion of Heterogeneous Classifiers, *Intelligent Data Analysis*, 9(6): 511-525.

Guoli, W., Roland, L. and Dunbrack, J. (2004). Scoring Profile-to-Profile Sequence Alignments, *Protein Science*, 13(6): 1612-1626.

Han, S., Lee, B. C., Yu, S. T., Jeong, C. S., Lee, S. and Kim, D. (2005). Fold Recognition by Combining Profile-Profile Alignment and Support Vector Machine, *Bioinformatics*, 21(11): 2667-2673.

Haoliang, Q., Sheng, L., Jianfeng, G., Zhongyuan, H. and Xinsong, X. (2008). Ordinal Regression for Information Retrieval, *Journal of Electronics*, 25 (1) 120-125.

Harpreet, K. S. and Daniel, F. (2007). FRalanyzer: A Tool for Functional Analysis of Fold-recognition Sequence–Structure Alignments, *Nucleic Acids Research*, 35(Web Server Issue): W499-W502.

Haussler, D. (1999). *Convolution Kernels on Discrete Structures*. Technical Report. Department of Computer Science, University of California.

Heather, A. C. and McCammon, J. A. (2000). Accommodating Protein Flexibility in Computational Drug Design, *Molecular Pharmacology*, 57(2): 213-218.

Henikoff, S. and Henikoff, J. G. (1992). Amino Acid Substitution Matrices from Protein Blocks, *PNAS*, 89(22): 10915-10919.

Hochreiter, S., Heusel, M. and Obermayer, K. (2007). Fast Model-Based Protein Homology Detection Without Alignment, *Bioinformatics*, 23(14): 1728-1736.

Honda, K., Hayashida, Y., Umaki, T., Okusaka, T., Kosuge, T., Kikuchi, S., Endo, M., Tsuchida, A., Aoki, T., Itoi, T., Moriyasu, F., Hirohashi, S. and Yamada, T. (2005). Possible Detection of Pancreatic Cancer by Plasma Protein Profiling, *Cancer Research*, 65(22): 10613-10622.

Honda, M., Kawai, H., Shirota, Y., Yamashita, T., Takamura, T. and Kaneko, S. (2005). cDNA Microarray Analysis of Autoimmune Hepatitis, Primary Biliary Cirrhosis and Consecutive Disease Manifestation, *Journal of*

Autoimmunity, 25(2): 133-140.

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J. and Nakai, K. (2007). WoLF PSORT: Protein Localization Predictor, *Nucleic Acids Research*, 35(Suppl 2): W585-W587.

Hou, Y., Hsu, W., Lee, M. and Bystroff, C. (2003). Efficient Remote Homology Detection Using Local Structure, *Bioinformatics*, 19(17): 2294-2301.

Hughey, R. and Krogh, A. (1996). Hidden Markov Models for Sequence Analysis: Extension and Analysis of the Basic Method, *Computer Applications in the Biosciences*, 12(2): 95-107.

Iain, M., Eugene, I., Rui, K., Jason, W., William, S. N. and Christina, L. (2007). SVM-Fold: A Tool for Discriminative Multi-class Protein Fold and Superfamily Recognition, *BMC Bioinformatics*, 8(Suppl 4): S2.

Ioannis, T., Thorsten, J., Thomas, H. and Yasemin, A. (2005). Large Margin Methods for Structured and Interdependent Output Variables, *The Journal of Machine Learning Research*, 6(1): 1453-1484.

Jaakkola, T., Diekhans, M. and Haussler, D. (2000). A Discriminative Framework for Detecting Remote Protein Homologies, *Journal of Computational Biology*, 7(1-2): 95-114.

Jaakkola, T., Diekhans, M. and Haussler, D. (1999). Using the Fisher Kernel Method to Detect Remote Protein Homologies. *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*. 6-10 August. Michigan, USA: AAAI Press, 149-158.

Jeffrey, S. and Daisuke, K. (2001). Defrosting the Frozen Approximation: PROSPECTOR - A New Approach to Threading, *Proteins: Structure, Function, and Genetics*, 42(3): 319-331.

Jinbo, X., Ming, L., Dongsup, K. and Ying, X. (2003). RAPTOR: Optimal Protein Threading by Linear Programming, *Journal of Bioinformatics and Computational Biology,* 1(1): 95-117.

Jones, D. T. (1999). GenTHREADER: An Efficient and Reliable Protein Fold Recognition Method for Genomic Sequences, *Journal of Molecular Biology*, 287(4): 797-815.

Jothi, R., Zotenko, E., Tasneem, A. and Przytycka, T. M. (2006). COCO-CL: Hierarchical Clustering of Homology Relations Based on Evolutionary Correlations, *Bioinformatics*, 22(7): 779-788.

Juan, D., Graña, O., Pazos, F., Fariselli, P., Casadio, R. and Valencia, A. (2003). A Neural Network Approach to Evaluate Fold Recognition Results, *Proteins: Structure, Function, and Genetics*, 50(4): 600-608.

Jung, I., Lee, J., Lee, S.-Y. and Kim, D. (2008). Application of Nonnegative Matrix Factorization to Improve Profile-Profile Alignment Features for Fold Recognition and Remote Homolog Detection, *BMC Bioinformatics*, 9(1): 298-309.

Karlin, S. and Altschul, S. F., (1990). Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes, *PNAS*, 87(6): 2264-268.

Karplus, K., Barrett, C. and Hughey, R. (1998). Hidden Markov Models for Detecting Remote Protein Homologies, *Bioinformatics*, 14(10): 846-856.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T. and Kanehisa, M. (2008). AAindex: Amino Acids Index Database, Progress Report 2008, *Nucleic Acids Research*, 36(Database Issue): D202-D205.

Kelley, L. A., MacCallum, R. M. and Sternberg, M. J. E. (2000). Enhanced Genome Annotation Using Structural Profiles in the Program 3D-PSSM, *Journal of Molecular Biology*, 299(2): 499-520.

Kim, D., Xu, D., Guo, J.-T., Ellrott, K. and Xu, Y. (2003). PROSPECT II: Protein Structure Prediction Program for Genome-Scale Applications, *Protein Engineering*, 16(9): 641-650.

Kliger, Y., Gallo, S. A., Peisajovich, S. G., Munoz-Barroso, I., Avkin, S., Blumenthal, R. and Shai, Y. (2000). Mode of Action of an Antiviral Peptide from HIV-1: Inhibition At a Post Lipid-Mixing Stage, *Journal of Biological Chemistry*, 276(2): 1391-1398.

Koretke, K. K., Russell, R. B., Copley, R. R. and Lupas, A. N. (1999). Fold Recognition Using Sequence and Secondary Structure Information, *Proteins*, (Suppl 3): 141-148.

Kristian, L. (1999). Limitations of the Paired t-test for Evaluation of Method Comparison Data, *Clinical Chemistry*, 45(1): 314-315.

Kristin, K. K., Robert, B. R. and Andrei, N. L. (2001). Fold Recognition from Sequence Comparisons, *Proteins: Structure, Function, and Genetics*, 45(S5): 68-75.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. and Haussler, D. (1994). Hidden

Markov Models in Computational Biology: Applications to Protein Modeling, *Journal of Molecular Biology*, 235(5): 1501-153.

Leslie, C., Eskin, E. and Noble, W., S. (2002). The Spectrum Kernel: A String Kernel for SVM Protein Classification. *Proceedings of the 7th Pacific Symposium on Biocomputing*. 3-7 January. Hawaii, USA: World Scientific Press, 565-575.

Leslie, C. S., Eskin, E., Cohen, A., Weston, J. and Noble, W. S. (2004). Mismatch String Kernels for Discriminative Protein Classification, *Bioinformatics*, 20(4): 467-476.

Liao, L. and Noble, W. (2002). Combining Pairwise Sequence Similarity and Support Vector Machines for Remote Protein Homology Detection. *Proceedings of the 6th Annual International Conference on Research in Computational Molecular Biology*. 18-21 April. Washington, DC, USA: ACM, 225-232.

Liao, L. and Noble, W. (2003). Combining Pairwise Sequence Similarity and Support Vector Machines for Detecting Remote Protein Evolutionary and Structural Relationships, *Journal of Computational Biology*, 10(6): 857-868.

Lingner, T. and Meinicke, P. (2008). Word Correlation Matrices for Protein Sequence Analysis and Remote Homology Detection, *BMC Bioinformatics*, 9(1): 259.

Liu, B., Lin, L., Wang, X., Dong, Q. and Wang, X. (2008a). A Discriminative Method for Protein Remote Homology Detection Based on N-Nary Profiles. *Proceedings of the 2nd Bioinformatics Research and Development*. 7-9 July. Vienna, Austria: Springer, 74-86.

Liu, B., Wang, X., Lin, L., Dong, Q. and Wang, X. (2008b). A Discriminative Method for Protein Remote Homology Detection and Fold Recognition Combining Top-n-grams and Latent Semantic Analysis, *BMC Bioinformatics*, 9(510): 1-16.

Madera, M. and Gough, J. (2002). A Comparison of Profile Hidden Markov Model Procedures for Remote Homology Detection, *Nucleic Acids Research*, 30(19): 4321-4328.

Mangasarian, O. and Musicant, D. (2001). Lagrangian Support Vector Machines *Journal of Machine Learning Research*, 1(1): 161-177.

Marc, A., Madhusudhan, M. S. and Andrej, S. (2004). Alignment of Protein

Sequences by Their Profiles, *Protein Science*, 13(4): 1071-1087.

McCallum, A. and Kamal, N. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*. 26-27 July. Madison, Wisconsin, USA: AAAI Press, 41-48.

McDonnell, A. V., Menke, M., Palmer, N., King, J., Cowen, L. and Berger, B. (2006). Fold Recognition and Accurate Sequence-Structure Alignment of Sequences Directing Beta-Sheet Proteins, *Proteins* 63(4): 976-985.

Melvin, I., Ie, E., Kuang, R., Weston, J., Noble, W. and Leslie, C. (2007). SVM-Fold: A Tool for Discriminative Multi-class Protein Fold and Superfamily Recognition, *BMC Bioinformatics*, 8(Suppl 4): S2.

Min, R., Bonner, A., Li, J. and Zhang, Z. (2009). Learned Random-Walk Kernels and Empirical-Map Kernels for Protein Sequence Classification, *Journal of Computational Biology*, 16(3): 457-474.

Mingjun, Z., Fabien, L., Mark, G. and Anatole, L. (2008). Classifying EEG for Brain Computer Interfaces Using Gaussian Processes, *Pattern Recognition Letters*, 29(3): 354-359.

Mittelman, D., Sadreyev, R. and Grishin, N. (2003). Probabilistic Scoring Measures for Profile-Profile Comparison Yield More Accurate Short Seed Alignments, *Bioinformatics*, 19(12): 1531-1539.

Mohseni, Z. S., Brézellec, P. and Risler, J. L. (2004). Cluster-C, an Algorithm For the Large-Scale Clustering of Protein Sequences Based on the Extraction of Maximal Cliques, *Computational Biology and Chemistry*, 28(3): 211-218.

Needleman, S. B. and Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *Journal of Molecular Biology*, 48(3): 443-453.

Noah, S. A. and Ismail, F. (2008). Automatic Classifications of Malay Proverbs Using Naive Bayesian Algorithm, *Information Technology Journal*, 7(7): 1016-1022.

Notredame, C., Higgins, D. G. and Heringa, J. (2000). T-coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment, *Journal of Molecular Biology*, 302(1): 205-217.

Octavian, T., Tamara, G., Jaroslaw, P. and Ron, E. (2004). Enriching the Sequence Substitution Matrix by Structural Information, *Proteins: Structure, Function,* *and Bioinformatics*, 54(1): 41-48.

Oğul, H. and Mumcuoğlu, E. U. (2007). A Discriminative Method for Remote Homology Detection Based on N-Peptide Compositions with Reduced Amino Acid Alphabets, *Biosystems*, 87(1): 75-81.

Oğul, H. and Mumcuoğlu, Ü. E. (2005). Discriminative Remote Homology Detection Using Maximal Unique Sequence Matches. In Famili, A.F. (Ed.) *Advances in Intelligent Data Analysis* (pp. 283-292). London: Springer-Berlin.

Ohlson, T., Wallner, B. and Elofsson, A. (2004). Profile-Profile Methods Provide Improved Fold Recognition: A Study of Different Profile-Profile Alignment Methods, *Proteins*, 57(1): 188-197.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G. and Notredame, C. (2004). 3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments, *Journal of Molecular Biology*, 340(2): 385-395.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1998). *The Page Rank Citation Ranking: Bringing Order to the Web*. Technical Report. Department of Computer Science. Stanford University.

Panchenko, A. R., Marchler-Bauer, A. and Bryant, S. H. (2000). Combination of Threading Potentials and Sequence Profiles Improves Fold Recognition, *Journal of Molecular Biology*, 296(5): 1319-1331.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. (1998). Sequence Comparisons Using Multiple Sequences Detect Three Times as Many Remote Homologues as Pairwise Methods, *Journal of Molecular Biology*, 284(4): 1201-1210.

Pearson, W. R. and Lipman, D. J. (1988). Improved Tools for Biological Sequence Comparison, *PNAS*, 85(8): 2444-2448.

Pettitt, C. S., McGuffin, L. J. and Jones, D. T. (2005). Improving Sequence-Based Fold Recognition by Using 3D Model Quality Assessment, *Bioinformatics*, 21(17): 3509-3515.

Pipenbacher, P., Schliep, A., Schneckener, S., Schonhuth, A., Schomburg, D. and Schrader, R. (2002). ProClust: Improved Clustering of Protein Sequences With an Extended Graph-Based Approach, *Bioinformatics*, 18(Suppl 2): S182-S191.

Polatkan, A. C., Ogul, H. and Sever, H. (2008). A Data Fusion Approach in Protein

Homology Detection. *Proceedings of the 2008 International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies*. 29 June-5 July. Washington, USA: IEEE Computer Society, 17-21.

Posner, K., Oquendo, M. A., Gould, M., Stanley, B. and Davies, M. (2007). Columbia Classification Algorithm of Suicide Assessment (C-CASA): Classification of Suicidal Events in the FDA's Pediatric Suicidal Risk Analysis of Antidepressants, *American Journal of Psychiatry*, 164(7): 1035-1043.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2007). NCBI Reference Sequence (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins, *Nucleic Acids Research*, 35(Database Issue): D61-D65.

Qian, B. and Goldstein, R. A. (2004). Performance of an Iterated T-HMM for Homology Detection, *Bioinformatics*, 20(14): 2175-2180.

Ralf, Z. and Ralf, T. (1997). Fast Protein Fold Recognition and Accurate Sequence-Structure Alignment. In Ralf, Z. (Ed.) *Lecture Notes in Computer Science* (pp. 137-146). Berlin: Springer-Verlag.

Rangwala, H. and Karypis, G. (2005). Profile-based Direct Kernels for Remote Homology Detection and Fold Recognition, *Bioinformatics*, 21(23): 4239-4247.

Rangwala, H. and Karypis, G. (2006). Building Multiclass Classifiers for Remote Homology Detection and Fold Recognition, *BMC Bioinformatics*, 7(1): 455.

Raval, A., Ghahramani, Z. and Wild, D. L. (2002). A Bayesian Network Model for Protein Fold and Remote Homologue Recognition. *Bioinformatics*, 18(6): 788-801.

Rifkin, R. and Klautau, A. (2004). In Defense of One-Versus-All Classification, *Journal of Machine Learning Research*, 5(1): 101-104.

Rocchio, J. J. (1966). *Document Retrieval Systems-optimization and Evaluation*. Doctor Philosophy, Harvard University, Cambridge.

Rossow, W. B. and Schiffer, R. A. (1999). Advances in Understanding Clouds from ISCCP, *Bulletin of the American Meteorological Society*, 80(11): 2261-2287.

Rui, K., Eugene, I., Ke, W., Kai, W., Mahira, S., Yoav, F. and Christina, L. (2004). Profile-Based String Kernels for Remote Homology Detection and Motif Extraction. *Proceedings of the 2004 IEEE Computational Systems*

Bioinformatics Conference. 16-19 August. Stanford, CA: IEEE, 152-160.

Runarsson, T. P. (2006). Ordinal Regression in Evolutionary Computation. In Runarsson, T.P. (Ed.) *Parallel Problem Solving from Nature* (pp. 1048-1057). Berlin: Springer-Verlag.

Rychlewski, L., Jaroszewski, L., Li, W. and Godzik, A. (2000). Comparison of Sequence Profiles: Strategies for Structural Predictions Using Sequence Information, *Protein Science*, 9(2): 232-241.

Sadreyev, R. and Grishin, N. (2003). COMPASS: A Tool for Comparison of Multiple Protein Alignments with Assessment of Statistical Significance, *Journal of Molecular Biology*, 326(1): 317-336.

Saigo, H., Vert, J.-P., Ueda, N. and Akutsu, T. (2004). Protein Homology Detection Using String Alignment Kernels, *Bioinformatics*, 20(11): 1682-1689.

Schölkopf, B., Smola, A. J., Williamson, R. C. and Bartlett, P. L. (2000). New Support Vector Algorithms, *Neural Computation*, 12(1): 1207-1245.

Schäffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L. and Altschul, S. F. (1999). IMPALA: Matching a Protein Sequence against a Collection of PSI-BLAST-Constructed Position-Specific Score Matrices, *Bioinformatics*, 15(12): 1000-1011.

Schoelkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. (1^{st} ed.) Cambridge, MA: The MIT Press.

Shah, A. R., Oehmen, C. S. and Webb-Robertson, B. J. (2008). SVM-HUSTLE-An Iterative Semi-Supervised Machine Learning Approach for Pairwise Protein Remote Homology Detection, *Bioinformatics*, 24(6): 783-790.

Shamim, M. T. A., Anwaruddin, M. and Nagarajaram, H. A. (2007). Support Vector Machine-Based Classification of Protein Folds using the Structural Properties of Amino Acid Residues and Amino Acid Residue Pairs, *Bioinformatics*, 23(24): 3320-3327.

Shi, J., Blundell, T. L. and Mizuguchi, K. (2001). FUGUE: Sequence-Structure Homology Recognition using Environment-Specific Substitution Tables and Structure-Dependent Gap Penalties, *Journal of Molecular Biology*, 310(1): 243-257.

Smith, T. and Waterman, M. S. (1981). Identification of Common Molecular Subsequences, *Journal of Molecular Biology*, 147(1): 195-197.

So¨ding, J. (2005). Protein Homology Detection by HMM-HMM Comparison, *Bioinformatics*, 21(9): 951-960.

Tan, A. C., David, G. and Deville, Y. (2003). Multi-Class Protein Fold Classification Using a New Ensemble Machine Learning Approach, *Genome Informatics*, 14(1): 206-217.

Tang, C. L., Xie, L., Koh, I. Y. Y., Posy, S., Alexov, E. and Honig, B. (2003). On the Role of Structural Information in Remote Homology Detection and Sequence Alignment: New Methods Using Hybrid Sequence Profiles, *Journal of Molecular Biology*, 334(5): 1043-1062.

Teichmann, S. A. (2002). The Constraints Protein-Protein Interactions Place on Sequence Divergence, *Journal of Molecular Biology*, 324(3): 399-407.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTALW: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-specific Gap Penalties and Weight Matrix Choice, *Nucleic Acids Research*, 22(22): 4673-4680.

Thorsten, J. (2005). A Support Vector Method for Multivariate Performance Measures. *Proceedings of the 22nd International Conference on Machine Learning*. August 11-13. Bonn, Germany: ACM, 377-384.

Thorsten, J. (2006). Training Linear SVMs in Linear Time. *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*. 20-23 August. Philadelphia, USA: ACM, 217-226.

Thorsten, J., Galor, T. and Elber, R. (2006). Learning to Align Sequences: A Maximum Margin Approach. In Leimkuhler, B (Ed.) *Lecture Notes in Computational Science and Engineering* (pp. 57-69). London: Springer-Verlag.

Tomas, O., Björn, W. and Arne, E. (2004). Profile-profile Methods Provide Improved Fold-Recognition: A Study of Different Profile-Profile Alignment Methods, *Proteins: Structure, Function, and Bioinformatics*, 57(1): 188-197.

Torralba, A., Murphy, K. and Freeman, W. (2006). Shared Features for Multiclass Object Detection. In J. Ponce, (Ed.) *Toward Category-Level Object Recognition*. (pp. 345-361). Berlin: Springer.

Tovinkere, V. R., Penaloza, M., Logar, A., Lee, J., Weger, R. C., Berendes, T. A. and Welch, R. M. (1992). An Intercomparison of Artificial Intelligence Approaches for Polar Scene Identification, *Journal of Geophysical Research*, 98(D3): 5001-5016.

Tsochantaridis, I., Joachims, T., Hofmann, T. and Altun, Y. (2005). Large Margin Methods for Structured and Interdependent Output Variables, *Journal of Machine Learning Research*, 6(1): 1453-1484.

Ukasz, J., Rychlewski., L., Baohong, Z. and Adam, G. (1998). Fold Prediction by a Hierarchy of Sequence, Threading, and Modeling Methods, *Protein Science*, 7(6): 1431-1440.

Vapnik, V. N. (1998). *Statistical Learning Theory*. (1st ed.) New York: Wiley.

Vendruscolo, M. and Dobson, C. M., (2005). A Glimpse at the Organization of the Protein Universe, *PNAS*, 102(16): 5641-5642.

Vingron, M. and Waterman, M. S. (1994). Sequence Alignment and Penalty Choice: Review of Concepts, Case Studies and Implications, *Journal of Molecular Biology*, 235(1): 1-12.

Waegeman, W., Baets, B. D. and Boullart, L. (2008). A Graph-Theoretic Approach for Reducing One-Versus-One Multi-Class Classification to Ranking. *Proceedings of 6th International Workshop on Mining and Learning with Graphs*. 4-5 July. Helsinki, Finland: Springer-Verlag, 1-3.

Wang, Z., Wang, Y., Xuan, J., Dong, Y., Bakay, M., Feng, Y., Clarke, R. and Hoffman, E. P. (2006). Optimized Multilayer Perceptrons for Molecular Classification and Diagnosis Using Genomic Data, *Bioinformatics*, 22(6): 755-761.

Watkins, C. (2000). Dynamic Alignment Kernels. In Smola, A.J. (Eds.) *Advances in Large Margin Classifiers*. (pp. 39-50). Cambridge: The MIT Press.

Webb-Robertson, B.-J. M., Oehmen, C. S. and Shah, A. R. (2008). A Feature Vector Integration Approach for a Generalized Support Vector Machine Pairwise Homology Algorithm, *Computational Biology and Chemistry*, 32(6): 458-461.

Welch, R. M., Sengupta, S. K., Goroch, A. K., Rabindra, P., Rangaraj, N. and Navar, M. S. (1992). Polar Cloud and Surface Classification Using AVHRR Imagery: An Intercomparison of Methods, *Journal of Applied Meteorology*, 31(5): 405-420.

Wicker, N., Perrin, G. R., Thierry, J. C. and Poch, O. (2001). Secator: A Program for Inferring Protein Subfamilies from Phylogenetic Trees, *Molecular Biology and Evolution*, 18(8): 1435-1441.

Wieser, D. and Niranjan, M. (2009). Remote Homology Detection Using a Kernel

Method that Combines Sequence and Secondary-Structure Similarity Scores, *In Silico Biology*, 9(3): 89-103.

Wittkop, T., Baumbach, J., Lobo, F. P. and Rahmann, S. (2007). Large Scale Clustering of Protein Sequences with FORCE-A Layout Based Heuristic for Weighted Cluster Editing, *BMC Bioinformatics*, 8(1): 396-340.

Xu, J., Li, M., Lin, G., Kim, D. and Xu, Y. (2003). Protein Threading by Linear Programming. *Proceedings of the 8th Pacific Symposium on Biocomputing*. 3-7 January. Hawaii, USA: World Scientific, 264-275.

Yibing, S., Guoli, W. and Huan-Xiang, Z. (2001). Fold Recognition and Accurate Query-Template Alignment by a Combination of PSI-BLAST and Threading, *Proteins: Structure, Function, and Genetics*, 42(1): 23-37.

Yona, G. and Levitt, M. (2002). Within the Twilight Zone: A Sensitive Profile-Profile Comparison Tool Based on Information Theory, *Journal of Molecular Biology*, 315(5): 1257-1275.

Yuna, H., Wynne, H., Lee, M. L. and Bystroff, C. (2004). Remote Homolog Detection Using Local Sequence-Structure Correlations, *PROTEINS: Structure, Function, and Bioinformatics,* 57(1): 518-530.

Zheng, S., Tang, H., Han, Z. and Zhang, H. (2006). Solving Large-scale Multiclass Learning Problems via an Efficient Support Vector Classifier, *Journal of Systems Engineering and Electronics*, 17(4): 910-915.

Zhou, H. and Zhou, Y. (2004). Single-Body Residue-Level Knowledge-Based Energy Score Combined with Sequence-Profile and Secondary Structure Information for Fold Recognition, *Proteins: Structure, Function, and Bioinformatics*, 55(4): 1005-1013.

Zhou, H. and Zhou, Y. (2005). Fold Recognition by Combining Sequence Profiles Derived from Evolution and from Depth-Dependent Structural Alignment of Fragments, *Proteins: Structure, Function, and Bioinformatics*, 58(2): 321-328.