DEVELOPMENT OF A COMPUTATIONAL FRAMEWORK FOR PROTEIN HOMOLOGY DETECTION BY INCORPORATING REALIGNMENT ALGORITHM

MOHAMAD FIRDAUS ABDULLAH

UNIVERSITI TEKNOLOGI MALAYSIA

ABSTRACT

Remote protein homology detection is a problem of detecting evolutionary relationship between proteins at low sequence similarity level. Among several problems in remote protein homology detection include the questions of determining which combination of multiple alignment and classification techniques is the best as well as the misalignment of protein sequences during the alignment process. Therefore, this study deals with remote protein homology detection via assessing the impact of using structural information on protein multiple alignments over sequence information. This study further presents the best combinations of multiple alignment and classification programs to be chosen. This study also improves the quality of the multiple alignments via integration of a refinement algorithm. The framework of this study began with datasets preparation on datasets from SCOP version 1.73, followed by multiple alignments of the protein sequences using CLUSTALW, MAFFT, ProbCons and T-Coffee for sequence-based multiple alignments and 3DCoffee, MAMMOTH-mult, MUSTANG and PROMALS3D for structural-based multiple alignments. Next, a refinement algorithm was applied on the protein sequences to reduce misalignments. Lastly, the aligned protein sequences were classified using the pHMMs generative classifier such as HMMER and SAM and also SVMs discriminative classifier such as SVM-Fold and SVM-Struct. The performances of assessed programs were evaluated using Receiver Operating Characteristics (ROC), Precision and Recall tests. The result from this study shows that the combination of refined SVM-Struct and PROMALS3D performs the best against other programs, which suggests that this combination is the best for remote protein homology detection. This study also shows that the use of the refinement algorithm increases the performance of the multiple alignments programs by at least 4 percent.

ABSTRAK

Pengesanan homologi protein terpencil merupakan permasalahan dalam mengesan hubungan evolusi antara protein yang mempunyai kesamaan urutan yang rendah. Antara masalah yang terdapat dalam pengesanan homologi protein terpencil termasuklah menentukan kombinasi terbaik teknik penyelarasan dan pengklasifikasian selain kesalahan penyelarasan urutan di dalam proses penyelarasan protein. Oleh itu, kajian ini adalah berkaitan pengesanan homologi protein yang terpencil melalui penilaian terhadap kesan penggunaan maklumat struktur kepada penyelarasan berganda protein berbanding penggunaan maklumat urutan. Kajian ini seterusnya memaparkan pilihan kombinasi terbaik bagi teknik penyelarasan dan pengklasifikasian. Kajian ini turut mempertingkatkan kualiti penyelarasan berganda melalui algoritma penambahbaikan. Rangka kerja kajian ini bermula dengan penyediaan set data daripada SCOP versi 1.73, diikuti penyelarasan berganda menggunakan CLUSTALW, MAFFT, ProbCons dan T-Coffee yang berasaskan struktur primer dan 3DCoffee, MAMMOTH-mult, MUSTANG serta PROMALS3D yang berasaskan struktur sekunder. Seterusnya, algoritma penambahbaikan diaplikasikan untuk mengurangkan kesalahan semasa penyelarasan. Akhir sekali, urutan protein diklasifikasikan menggunakan HMMER dan SAM yang berasaskan Model Markov Tersembunyi Berprofil (pHMMs) dan SVM-Fold serta SVM-Struct yang berasaskan Mesin Vektor Sokongan (SVMs). Karakter Pengoperasian Penerima (ROC), ketepatan dan dapatan semula digunakan untuk menilai kemampuan rangka kerja yang dicadangkan ini. Hasil kajian menunjukkan bahawa kombinasi SVM-Struct dan PROMALS3D mengatasi kombinasi yang lain. Ini menunjukkan ia adalah kombinasi terbaik bagi pengesanan homologi protein terpencil. Kajian ini turut menunjukkan bahawa penggunaan algoritma penambahbaikan telah meningkatkan prestasi program penyelarasan berganda sebanyak sekurang-kurangnya 4 peratus.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3D	-	3-Dimensional
BLAST	-	Basic Local Alignment Tools
CASP	-	Critical Assessment of Techniques for Protein Structure Prediction
CRFs	-	Conditional Random Fields
DNA	-	Deoxyribonucleic Acid
GNU	-	General Public License
HMMs	-	Hidden Markov Models
LDA	-	Linear Discriminant Analysis
LSA	-	Latent Semantic Analysis
MSA	-	Multiple Sequence Alignment
MStA	-	Multiple structural alignments
NIC	-	Network Interface Card
NN	-	Neural Networks
PCR	-	Polymerase Chain Reaction
pHMMs	-	Profile Hidden Markov Models
RAM	-	Random Access Memory
RNA	-	Ribonucleic Acid
ROC	-	Receiver Operating Characteristics
SCOP	-	Structural Classification of Proteins
SVMs	-	Support Vector Machines

CHAPTER 1

INTRODUCTION

1.1 Background

Remote protein homology detection forms the basis for structure prediction, function prediction and evolution in protein. Being a core problem in computational biology, there are two different degrees of remote protein homology. The first one is sequence homology while the second one is structural homology. Protein sequence homology is where protein sequences are compared to each other as subtle similarity between the compared protein sequences defines homology. As for structural homology, whether or not there are homologies are detected by finding identical secondary structures and motifs in the compared proteins. The main objective in remote protein homology detection is to find homology of protein sequences when the actual sequence identity is low.

The use of multiple alignments has been proven to improve the detection of remote protein homology. There are two types of multiple alignments in bioinformatics which are multiple sequence alignments and multiple structural alignments. Multiple sequence alignments are often used to assess protein sequences shared evolutionary origins. Meanwhile, multiple structural alignments are essential in providing benchmarks dataset for improving sequence alignment algorithm as bases for bioinformatics research.

Meanwhile, another two fashionable methods in computational biology for detecting remote homologies are Hidden Markov Models (HMMs) and Support Vector Machines (SVMs). As probabilistic models, HMMs are initially used in speech recognition (Mendel, 1992). To date, HMMs are being applied in solving molecular biology problems such as gene finding (Brejova *et al.*, 2005; Majoros *et al.*, 2005), multiple sequence alignment (MSA: Mamitsuka, 2005; Knudsen and Miyamoto, 2003) and protein structure prediction (Lampros *et al.*, 2007; Camproux and Tufféry, 2005; Lin *et al.*, 2005). HMMs that are used to represents groups of homologues sequences are called profile Hidden Markov Models (pHMMs). The pHMMs are probabilistic models built from multiple sequence alignments. Madera and Gough (2002) has systematically compared the performance of HMMER (http://hmmer.janelia.org/) and SAM (http://compbio.soe.ucsc.edu/HMM-apps/) which is based on pHMMs over two protein families, globins and cupredoxins by using nrdb90 (Holm and Sander, 1998) database and an all-against-all experiment for the two systems using SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) database. In their works, several alignment strategies have been used, including manual alignment of the two protein families, SAM-T99 (http://compbio.soe.ucsc.edu/HMM-apps/T99-query.html) seeded from a single protein, WU-BLAST (http://blast.wustl.edu/) search from the seed protein followed by CLUSTALW (http://www.ebi.ac.uk /Tools/clustalw2/). They showed that the initial multiple alignments can significantly affect HMMER and SAM performance, also that SAM T-99 package generates a good quality multiple alignments. They found that SAM had better model quality than HMMER. The two systems were further evaluated by Wistrand and Sonnhammer (2005). In their work, they relied on SCOP database for high quality labeled hierarchies of protein domains. They explicitly avoided conditioning on the use of particular program to perform initial multiple alignments and instead they used Pfam (http://pfam.sanger.ac.uk/) database. They concluded that SAM's model estimation is superior, due to better usage of priors, which avoid over-fitting. On the other hand, they also showed that HMMER's model scoring is more accurate, probably due to a better null model. Bernardes *et al.* (2007) works investigate the contributions of using multiple structural alignments to build the model for remote protein homology detection by considering proteins below 30% in identity. Their experiments showed that profile HMMs derived from multiple structural alignments perform significantly better than that derived from multiple sequence alignments. They also showed that accuracy of alignment is not directly related to alignment identity. They suggested that although multiple structural alignments often present smaller identity than multiple sequence alignments, the best quality alignments based on structural information are generally considered to derive from structural superposition. In their work, they compare the performance of two pHMMs packages which are HMMER and SAM when two different kinds of alignments that are sequence and structural alignments were used. Their results showed that HMMER based on structural alignment outperforms SAM for such remote homologues.

Meanwhile, SVMs are method for constructing a rule called linear classifier in a way that it produces classifiers with theoretical guarantees of good predictive performance that is the quality of classification of unseen data. In short, SVMs are a set of related supervised learning methods used for classification and regression. Rangwala and Karypis (2006) presents an extensive evaluation of a number of methods for building SVM-based multiclass classification schemes in the context of the SCOP protein classification. Their methods are comprised of schemes that directly build a SVMs-based multiclass model, schemes that employ a second level learning approach to combine the predictions generated by a set of binary SVMs-based classifiers and also schemes that build and combine binary classifiers for various levels of the SCOP hierarchy beyond those defining the target classes. The SVM-Fisher method by Jaakkola *et al.* (1999) combines an iterative HMMs training scheme with discriminative algorithm of SVMs. For any given family of related proteins, the HMMs provide a kernel function. First, the HMMs are trained of positive members of the training set using the standard Baum-Welch (Baum *et al.*, 1970) training routine. Then, the training is iterated, adding similar sequences from a large unlabelled database to the training set at each round. After training, the gradient vector of any sequence can be computed with respects to the trained model. Lastly, SVMs are trained on a collection of positively and negatively labeled protein gradient vectors. By coupling HMMs and SVMs, this method offers an interpretable model, a means of incorporating prior knowledge and missing data and also excellent recognition performance.

In this thesis, super-families from SCOP database are used as datasets. Firstly, the performance of different multiple alignments with different classifiers are assessed. Next, a refinement algorithm is integrated to improve the multiple alignments before being classified using the classifiers. Then, the performance between the refined and unrefined multiple alignments are compared. HMMER and SAM which are two popular tools in bioinformatics for pHMMs in detecting remote protein homologies are used to provide generative classification. Meanwhile, SVM-Fold (http://svm-fold.c2b2.columbia.edu/) and SVM-Struct (http://svmlight. joachims.org/svm_struct.html) are used to provide discriminative classification.

1.2 Current Methods in Remote Protein Homology Detection

Generally, there are three basic groups of major methods in remote protein homology detection (Liao and Noble, 2003). We will discuss these methods in detail in Chapter 2.

(i) Pairwise sequence comparison algorithms which identify similarity region that may be the consequences of functional, structural or evolutionary relationships by arranging primary sequences in proteins. Examples of these algorithms include BALSA (Webb *et al.*, 2002), NdPASA (Wang and Feng, 2005), CPSA algorithm (He and Arslan, 2005) and INSPAL (Lee and Wang, 2006).

(ii) Generative models for protein families use positive examples of a protein family which can be trained iteratively using both positively labeled and unlabeled examples by pulling in close homology and adding them to the positive set. These models include HMMs (Remmert *et al.*, 2009), Naive Bayes (Nigsch *et al.*, 2008), Gaussian mixture model (Aristophanous *et al.*, 2007) and Latent Semantic Analysis (LSA: Cohen *et al.*, 2008).

(iii) Discriminative classifiers are able to gain additional accuracy by modelling the difference between positive and negative examples explicitly, providing state-of-the-art performance with appropriate kernels. Examples include SVMs (Nugent and Jones, 2009), Neural networks (NN: Rubinsky *et al.*, 2008), Linear Discriminant Analysis (LDA: Chen *et al.*, 2009) and conditional random fields (CRFs: Lafferty, *et al.,* 2001).

1.3 Challenges in Remote Protein Homology Detection

There are several challenges in remote protein homology detection which we will address in this study. Firstly, choosing multiple alignment types for protein remote homology detection can be tricky and challenging as there are two types of multiple alignments namely multiple sequence alignment and multiple structural alignments. Multiple structural alignments are often said to be more accurate than multiple sequence alignments at identifying motifs and functional residues. A study performed by Madera and Gough (2002) proved this statement to be true. However, a study by Jones and Bateman (2002) concluded that the use of structure information actually does not help to improve multiple alignment accuracy in homologue detection with pHMMs.

Secondly, the accuracy of domain identification, protein classification and reconstruction of phylogenetic history of domain families crucially depends on the quality of underlying multiple sequence alignments (Chakrabarti *et al.*, 2006). Different method has been proposed to produce a multiple sequence alignment. Some of them align all sequences simultaneously while others apply a progressive algorithm. In progressive alignment strategy, sequences are aligned in a predetermined order as dictated by the guide tree in groups with other similar sequences together with subsequent addition of more dissimilar ones. But progressive alignment has its pitfalls where misalignment made at previous stages cannot be corrected afterwards, thus can propagates into serious alignment errors. Moreover, the final alignment depends strongly on the order of the sequences being aligned. Therefore, the challenge lies in realigning the sequences in order to correct misalignments between a given sequence and the rest of the profile.

The third challenge in this study is to assess and come out with a comparative result on the performance between generative and discriminative classifiers, providing information and aid for researchers on choosing between these classifiers. Comparison on generative and discriminative classifiers has been a topic of discussion for a long time. For example, a work by Ng and Jordan (2001) compares logistic regression as discriminative classifier with naïve Bayes as generative classifier. In their work they proved that discriminative classifier works better than its generative counterpart. However, this is true only for a large number of training data. If the number of training data is limited, generative classifier can outperform the discriminative classifier. Due to this fact, several authors (Holub and Perona, 2005; Bouchard and Triggs, 2004) have proposed a hybrid of generative and discriminative classifier approaches. However, even though their procedure is heuristic, it was sometimes found that the best predictive performance is only somewhere in between the discriminative and generative limits.

1.4 Statement of the Problems

The remote protein homology detection problem to be studied can be described as follows:

"Given multiple protein sequences, the challenge is to assess the best of different combination of multiple sequence alignments and multiple structural alignments with generative and discriminative classifiers in remote protein homology detection and at the same time reducing misalignments in order to achieve higher Receiver Operating Characteristics (ROC: Beck and Schultz, 1986), Precision and Recall values"

This study will assists in the problem of selecting the best multiple alignments by comparing the performance between pHMMs and SVMs derived from multiple sequence alignments and multiple structural alignments. The factor that has to be considered in order to provide the best solution to this problem is the revelation of relationships between the proteins. This will lead to a more technical task that is analyzing the scores generated by the classifiers. Meanwhile, in order to solve the problem of misalignments in multiple alignments, a refinement algorithm will be used. To do this, iterative realignment of individual sequences with the predetermined conserved core that is the block model of a protein family will be taken as the factor which has to be considered. Misalignments resulting from the aligning process have to be reduced because the accuracy of our protein classification highly depends on the quality of the underlying alignments.

1.5 Objectives of the Study

The goal of this study is to develop a computational framework to classify proteins into each super-families and families respectively. In order to realize this goal, several objectives must be achieved:

(i) To study and investigates current remote protein homology detection methods in order to understand the processes, data and domains.

(ii) To integrate different combinations between multiple sequence alignments and multiple structural alignments with pHMMs and SVMs in order to find the best combinations in detecting remote protein homology.

(iii) To apply refining algorithm in order to reduce misalignments in multiple sequence alignments and multiple structural alignments.

(iv) To analyze results using ROC, Precision and Recall in order to evaluate the performance of the proposed computational framework.

1.6 Scope and Significance of the Study

In this study, we limit our scope of experimental datasets to SCOP database version 1.73 with identity below 30% as our work considers proteins within the *Twilight Zone* where identity between amino acids sequences is a weaker indicative of evolutionary relationships. SCOP is a manually inspected database of protein folds and it is very suitable for our study because it describes structural and evolutionary relationships between proteins including all entries in the PDB (http://www.rcsb.org) database. SCOP is an excellent dataset for assessing the performance of remote protein homology detection methods, and it has been widely used for that purpose. SCOP categorizes all protein domains of known structure into a hierarchy of four levels: class, fold, super family and family. The scope of our work will be at super-family level, in which families are grouped such that a common evolutionary origin is not obvious from sequence identity, but in the meantime probable from an analysis of structure and from functional features. We believe that this level represents remote protein homology detection the best. Throughout our study, the sequence-based multiple alignment tools that will be used are limited to: CLUSTALW, T-Coffee (http://www.tcoffee.org/), MAFFT (http://www.ebi.ac.uk/Tools/mafft/) and Prob-Cons (http://probcons.stanford.edu/). On the other hand, the structural-based multiple alignment tools will be limited to: 3DCoffee (http://www.tcoffee.org/), MAM-MOTH-mult (http://ub.cbm.uam.es/mammoth/mult/), MUSTANG (http://www.cs. mu.oz.au/~arun/mustang/) and PROMALS3D (http://prodata.swmed.edu/promals3d /). A refinement algorithm is applied on the output of the multiple alignment tools in order to reduce the misalignments. Next, the unrefined and refined multiple alignments are classified using pHMMs and SVMs. For pHMMs, HMMER and SAM are used to provide the classification. Meanwhile, SVM-Struct and SVM-Fold are used to provide SVMs classification. Lastly, an analysis on the performance of these tools which have been derived from unrefined and refined multiple alignments are conducted using ROC, Precision and Recall.

Remote protein homology detection is an important yet hard problem in computational molecular biology. A number of tools and methods have been developed towards this purpose as well as to improvise it. Therefore, the significance of this study is that it helps in improvising remote protein homology detection by providing choices in selecting the best and most appropriate multiple alignments tools. This is due to the fact that different kinds of alignments give different results. Also, the usage of different multiple alignment tools will also resulted in different level in performance due to different methods and algorithms implemented. By applying a method to reduce misalignments in protein sequences, this study will also significantly help in preventing serious alignment errors. In this study, we will also compare the performance of two different types of classifier derived from multiple alignment tools mentioned before. We will analyze all the result from these classifiers thoroughly to provide better assessments for these tools, aimed also at providing help in choices of selection.

Remote protein homology detection plays a crucial part in medicine such as in drug design and cancer genomics as well as in biotechnology such as in the design of novel enzymes. Every two years starting 1994, the performances of current methods in this field are assessed in Critical Assessment of Techniques for Protein Structure Prediction (CASP: http://predictioncenter.org/), which is a community wide experiment for protein structure prediction held by Protein Structure Prediction Center, University of California. Homology modeling has been extensively used in structure-based drug design as discussed in detail in a review by Jacobson *et al.*, (2004). Another example of using remote protein homology detection in drug design is the work by Caffrey *et al.* (2005). The main goal of their work is to compare active sites to obtain hints for drug design. They used homology model of *Schistosoma japonicum cathepsin D* to identify the structural differences between that protein and its human homolog that were responsible for differential binding of certain types of *cathepsin D* inhibitors. They used this information to design inhibitors that show greater specificity to the worm version of the protein.

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, problems, objectives, scope and significance of the study.

(ii) In Chapter 2, the basic concepts, involved phases, and raised problem in remote protein homology detection are described. Exhaustive reviews of previous related works are also presented.

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

(iv) Chapter 4 focuses on assessing the performance of pHMMs and SVMs when two different types of multiple alignments that are sequence and structural based are used.

(v) Chapter 5 describes the measuring of performances between refined and unrefined multiple alignments on pHMMs and SVMs.

(vi) In Chapter 6, the conclusion of the proposed framework and the achieved results to date is shown. Descriptions of the contributions and future works of the study are also presented.

REFERENCES

Agresti, A. (2002) *Categorical Data Analysis, 2nd Edition*. New York, USA: John Wiley & Sons, Inc.

Ahola, V., Aittokallio, T., Vihinen, M. and Uusipaikka, E. (2008) Model-based Prediction of Sequence Alignment Quality, *Bioinformatics*. 24(19): 2165-2171.

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, *Proceeding of the National Academy of Sciences of the United States of America*. 98(26): 14796-14801.

Altschul, S.F., Gertz, E.M., Agarwala, R., Schaffer, A.A. and Yu, Y.-K. (2008) PSI-BLAST Pseudocounts and the Minimum Description Length Principle, *Nucleic Acids Research*. http://nar.oxfordjournals.org/cgi/content/abstract/gkn981v981.

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.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J.P., Chothia, C. and Murzin, A.G. (2004) SCOP Database in 2004: Refinements Integrate Structure and Sequence Family Data, *Nucleic Acids Research*. 32(Database Issue): D226-229.

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S.E., Hubbard, T.J.P., Chothia, C. and Murzin, A.G. (2008) Data growth and its impact on the SCOP database: new developments, *Nucleic Acids Research*. 36(suppl_1): D419-425.

Aristophanous, M., Penney, B.C., Martel, M.K. and Pelizzari, C.A. (2007) A Gaussian Mixture Model for Definition of Lung Tumor Volumes in Positron Emission Tomography, *Medical Physics*. 34(11): 4223-4235.

Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V. and Notredame, C. (2006) Expresso: Automatic Incorporation of Structural Information in Multiple Sequence Alignments using 3D-Coffee, *Nucleic Acids Research*. 34(Suppl_2): W604-608.

Bairoch, A. and Apweiler, R. (1996) The SWISS-PROT Protein Sequence Data Bank and Its New Supplement TREMBL, *Nucleic Acids Research*. 24(1): 21-25.

Balaji, S. and Srinivasan, N. (2001) Use of A Database of Structural Alignments and Phylogenetic Trees in Investigating the Relationship between Sequence and Structural Variability among Homologous Proteins, *Protein Engineering*. 14(4): 219-226.

Balaji, S., Sujatha, S., Kumar, S.S.C. and Srinivasan, N. (2001) PALI - A Database of Phylogeny and Alignment of Homologous Protein Structures, *Nucleic Acids Research*. 29(1): 61-65.

Balakrishnan, N. (1992) *Handbook of the Logistic Distribution*. New York, USA: Marcel Dekker, Inc.

Barton, G.J. and Sternberg, M.J. (1987) A Strategy for the Rapid Multiple Alignment of Protein Sequences. Confidence Levels from Tertiary Structure Comparisons, *Journal of Molecular Biology*. 198(2): 327-337.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970) A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains, *The Annals of Mathematical Statistics*. 41(1): 164-171.

Beck, J.R. and Shultz, E.K. (1986) The Use of Relative Operating Characteristic (ROC) Curves in Test Performance Evaluation, *Archive of Pathology and Laboratory Medicine*. 110(1): 13-20.

Ben-Hur, A. and Brutlag, D. (2003) Remote Homology Detection: A Motif Based Approach, *Bioinformatics*. 19(Suppl_1): i26-33.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2008) GenBank, *Nucleic Acids Research*. 36(Database Issue): D25-30.

Benton, D. (1990) Recent Changes in the GenBank On-line Service, *Nucleic Acids Research*. 18(6): 1517-1520.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank, *Nucleic Acids Research*. 28(1): 235-242.

Bernardes, J., Davila, A., Costa, V. and Zaverucha, G. (2007) Improving Model Construction of Profile HMMs for Remote Homology Detection Through Structural Alignment, *BMC Bioinformatics*. 8(1): 435-447.

Bhadeshia, H.K.D.H. (1999) Neural Networks in Materials Science, *Iron and Steel Institute of Japan International*. 39(10): 966-979.

Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and Genomewise, *Genome Research*. 14(5): 988-995.

Birzele, F., Gewehr, J.E., Csaba, G. and Zimmer, R. (2007) Vorolign-fast Structural Alignment using Voronoi Contacts, *Bioinformatics*. 23(2): e205-211.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation, *Journal of Machine Learning Research*. 3(1): 993-1022.

Bouchard, G. and Triggs, B. (2004) The Trade-off between Generative and Discriminative Classifiers. In Antoch, J. (ed), *Proceedings of the 16th IASC International Conference on Computational Statistics*. Physica-Verlag Heidelberg, Prague, Czech Republic, 721-728.

Bourne, P. and Weissig, H. (2003) *Structural Bioinformatics*, *1st Edition*. Hoboken, New Jersey: Wiley-Liss.

Brandt, B.W. and Heringa, J. (2009) WebPRC: The Profile Comparer for Alignment-based Searching of Public Domain Databases, *Nucleic Acids Research*. http://nar.oxfordjournals.org/cgi/content/abstract/gkp279v271.

Bray, N. and Pachter, L. (2004) MAVID: Constrained Ancestral Alignment of Multiple Sequences, *Genome Research*. 14(4): 693-699.

Brejova, B., Brown, D.G., Li, M. and Vinar, T. (2005) ExonHunter: A Comprehensive Approach to Gene Finding, *Bioinformatics*. 21(Suppl_1): i57-65.

Burke, D.F., Deane, C.M., Nagarajaram, H.A., Campillo, N., Martin-Martinez, M., Mendes, J., Molina, F., Perry, J., Reddy, B.V., Soares, C.M., Steward, R.E., Williams, M., Carrondo, M.A. and Blundell, T.L.M., K. (1999) An Iterative Structure-assisted Approach to Sequence Alignment and Comparative Modeling, *Proteins*. 1(Suppl_3): 55-60.

Caffrey, C.R., Placha, L., Barinka, C., Hradilek, M., Dostál, J., Sajid, M., McKerrow, J.H., Majer, P., Konvalinka, J. and Vondrásek, J. (2005) Homology Modeling and SAR Analysis of Schistosoma japonicum Cathepsin D (SjCD) with Statin Inhibitors Identify a Unique Active Site Steric Barrier with Potential for the Design of Specific Inhibitors, *Biological Chemistry*. 386(4): 339-349.

Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X. and Chen, Y.Z. (2003) SVM-Prot: Web-based Support Vector Machine Software for Functional Classification of A Protein from Its Primary Sequence, *Nucleic Acids Research*. 31(13): 3692-3697.

Camproux, A.C. and Tufféry, P. (2005) Hidden Markov Model-derived Structural Alphabet for Proteins: The Learning of Protein Local Shapes Captures Sequence Specificity, *Biochimica et Biophysica Acta*. 1724(3): 394-403.

Chakrabarti, S., Lanczycki, C.J., Panchenko, A.R., Przytycka, T.M., Thiessen, P.A. and Bryant, S.H. (2006) Refining Multiple Sequence Alignments with Conserved Core Regions, *Nucleic Acids Research*. 34(9): 2598-2606.

Chen, X., Liang, Y.Z., Yuan, D.L. and Xu, Q.S. (2009) A Modified Uncorrelated Linear Discriminant Analysis Model Coupled with Recursive Feature Elimination for the Prediction of Bioactivity, *SAR and QSAR in Environmental Research*. 20(1): 1-26.

Chung, S.Y. and Subbiah, S. (1996) A Structural Explanation for the Twilight Zone of Protein Sequence Homology, *Structure*. 4(10): 1123-1127.

Churchill, G.A. (1989) Stochastic Models for Heterogeneous DNA Sequences, *Bulletin of Mathematical Biology*. 51(1): 79-94.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. and de Hoon, M.J.L. (2009) Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics, *Bioinformatics*. 25(11): 1422-1423.

Cohen, T., Blatter, B. and Patel, V. (2008) Simulating Expert Clinical Comprehension: Adapting Latent Semantic Analysis to Accurately Extract Clinical Concepts from Psychiatric Narrative, *Journal of Biomedical Informatics*. 41(6): 1070-1087.

Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. New York, USA: Cambridge University Press.

Dai, J. and Cheng, J. (2008) HMMEditor: A Visual Editing Tool for Profile Hidden Markov Model, *BMC Genomics*. 9(Suppl_1): S8.

Dalton, J.A.R. and Jackson, R.M. (2007) An Evaluation of Automated Homology Modelling Methods at Low Target Template Sequence Similarity, *Bioinformatics*. 23(15): 1901-1908.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1972) A Model of Evolutionary Change in Proteins. In Dayhoff, M.O. (ed), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Maryland, USA, 89-99.

Do, C.B., Mahabhashyam, M.S.P., Brudno, M. and Batzoglou, S. (2005) PROBCONS: Probabilistic Consistency-based Multiple Sequence Alignment, *Genome Research*. 15(2): 330-340.

Domingos, P. and Pazzani, M. (1997) On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, *Machine Learning*. 29(2-3): 103-130.

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

Dong, Q.W., Lin, L., Wang, X.L. and Li, M.H. (2005) A Pattern-based SVM for Protein Remote Homology Detection. *Proceedings of the 4th International Conference on Machine Learning and Cybernetics*. IEEE Computer Society, GuangZhou, China, 3363-3368.

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

Edgar, R.C. (2004) MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput, *Nucleic Acids Research*. 32(5): 1792-1797.

Edgar, R.C. and Batzoglou, S. (2006) Multiple Sequence Alignment, *Current Opinion in Structural Biology*. 16(3): 368-373.

Edgar, R.C. and Sjolander, K. (2004) COACH: Profile-profile Alignment of Protein Families using Hidden Markov Models, *Bioinformatics*. 20(8): 1309-1318.

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*. ACM, Bonn, Germany, 329-336.

Evans, M., Hastings, N. and Peacock, B. (2000) *Statistical Distributions*, *3rd Edition*. New York, USA: John Wiley & Sons, Inc.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L.L. and Bateman, A. (2008) The Pfam Protein Families Database, *Nucleic Acids Research*. 36(Suppl_1): D281-288.

Fisher, R.A. (1936) The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*. 7(1): 179-188.

Freund, Y. (1990) Boosting A Weak Learning Algorithm by Majority. *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*. Morgan Kaufmann Publishers, Inc., New York, United States, 202-216

Furmonaviciene, R. and Shakib, F. (2001) The Molecular Basis of Allergenicity: Comparative Analysis of the Three Dimensional Structures of Diverse Allergens Reveals a Common Structural Motif, *Molecular Pathology*. 54(3): 155–159.

Gribskov, M. (1994) Profile Analysis, *Methods in Molecular Biology*. 25(1): 247-266.

Gribskov, M. and Robinson, N.L. (1996) Use of Receiver Operating Characteristic (ROC) Analysis to Evaluate Sequence Matching, *Computers & Chemistry*. 20(1): 25-33.

Grundy, W.N., Bailey, T.L., Elkan, C.P. and Baker, M.E. (1997) Meta-MEME: Motif-based Hidden Markov Models of Protein Families, *Computer Applications in the Biosciences*. 13(4): 397-406.

Guda, C., Lu, S., Scheeff, E.D., Bourne, P.E. and Shindyalov, I.N. (2004) CE-MC: A Multiple Protein Structure Alignment Server, *Nucleic Acids Research*. 32(Web Server Issue): W100-103.

Guda, C., Pal, L.R. and Shindyalov, I.N. (2006) DMAPS: A Database of Multiple Alignments for Protein Structures, *Nucleic Acids Research*. 34(Database Issue): D273-276.

Havil, J. (2003) *Gamma: Exploring Euler's Constant*. New Jersey, USA: Princeton University Press.

He, D. and Arslan, A.N. (2005) A Space-Efficient Algorithm for the Constrained Pairwise Sequence Alignment Problem, *Genome Informatics*. 16(2): 237-246.

Hearst, M.A. (1998) Support Vector Machines, *IEEE Intelligent Systems*. 13(4): 18-28.

Heger, A. and Holm, L. (2001) Picasso: Generating A Covering Set of Protein Family Profiles, *Bioinformatics*. 17(3): 272-279.

Henikoff, S. and Henikoff, J.G. (1992) Amino Acid Substitution Matrices from Protein Blocks, *Proceeding of the National Academy of Sciences of the United States of America*. 89(22): 10915-10919.

Holm, L. and Park, J. (2000) DaliLite Workbench for Protein Structure Comparison, *Bioinformatics*. 16(6): 566-567.

Holm, L. and Sander, C. (1998) Removing Near-neighbour Redundancy from Large Protein Sequence Collections, *Bioinformatics*. 14(5): 423-429.

Holub, A. and Perona, P. (2005) A Discriminative Framework for Modelling Object Classes. *Proceedings of the 15th IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, San Diego, CA, USA, 664-671.

Homaeian, L., Kurgan, L.A., Ruan, J., Cios, K.J. and Chen, K. (2007) Prediction of Protein Secondary Structure Content for the Twilight Zone Sequences, *Proteins: Structure, Function, and Bioinformatics*. 69(3): 486-498.

Hosmer, D.W. and Lemeshow, S. (2000) *Applied Logistic Regression, 2nd Edition*. New York, USA: John Wiley & Sons, Inc.

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

Hou, Y., Hsu, W., Lee, M.L. and Bystroff, C. (2004) Remote Homolog Detection using Local Sequence-Structure Correlations, *Proteins: Structure, Function, and Bioinformatics*. 57(3): 518-530.

Huang, Y.M. and Bystroff, C. (2006) Improved Pairwise Alignments of Proteins in the Twilight Zone using Local Structure Predictions, *Bioinformatics*. 22(4): 413-422.

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.

Ilyin, V.A., Abyzov, A. and Leslin, C.M. (2004) Structural Alignment of Proteins by A Novel TOPOFIT Method, As A Superimposition of Common Volumes at A Topomax Point, *Protein Science*. 13(7): 1865-1874.

Ivanciuc, O. (2007) Applications of Support Vector Machines in Chemistry. In Lipkowitz, K.B. and Cundari, T.R. (eds), *Reviews in Computational Chemistry*. Wiley-VCH, Weinheim, 291-400.

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*. AAAI Press, Heidelberg, Germany, 149-158.

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.

Jacobson, M. and Sali, A. (2004) Comparative Protein Structure Modeling and Its Applications to Drug Discovery. In Doherty, A.M. (ed), *Annual Reports in Medicinal Chemistry*. Academic Press, United Kingdom, 259-276.

Jaroszewski, L., Li, W. and Godzik, A. (2002) In Search for More Accurate Alignments in the Twilight Zone, *Protein Science*. 11(7): 1702-1713.

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W. and Godzik, A. (2005) FFAS03: A Server for Profile-profile Sequence Alignments, *Nucleic Acids Research*. 33(Web Server Issue): W284-288.

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

Jones, S.G. 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.

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.

Kann, M.G., Thiessen, P.A., Panchenko, A.R., Schaffer, A.A., Altschul, S.F. and Bryant, S.H. (2005) A Structure-based Method for Protein Sequence Alignment, *Bioinformatics*. 21(8): 1451-1456.

Karplus, K. (2009) SAM-T08, HMM-based Protein Structure Prediction, *Nucleic Acids Research*. 37(Suppl_2): W492-497.

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

Katoh, K., Kuma, K., Miyata, T. and Toh, H. (2005) Improvement in the Accuracy of Multiple Sequence Alignment Program MAFFT, *Genome Informatics*. 16(1): 22-33.

Katoh, K., Kuma, K., Toh, H. and Miyata, T. (2005) MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment, *Nucleic Acids Research*. 33(2): 511-518.

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform, *Nucleic Acids Research*. 30(14): 3059-3066.

Katoh, K. and Toh, H. (2008) Improved Accuracy of Multiple NcRNA Alignment by Incorporating Structural Information into A MAFFT-based Framework, *BMC Bioinformatics*. 9(1): 212-225.

Kedem, K., Chew, L.P. and Elber, R. (1999) Unit-vector RMS (URMS) as A Tool to Analyze Molecular Dynamics Trajectories, *Proteins*. 37(4): 554-564.

Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (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., Ellrott, K. and Xu, Y. (2003) PROSPECT II: Protein Structure Prediction Program for Genome-scale Applications, *Protein Engineering*. 16(9): 641-650.

Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J. and Lesk, A.M. (2006) MUSTANG: A Multiple Structural Alignment Algorithm, *Proteins: Structure, Function, and Bioinformatics*. 64(3): 559-574.

Krogh, A., Mian, I.S. and Haussler, D. (1994) A Hidden Markov Model That Finds Genes in E.coli DNA, *Nucleic Acids Research*. 22(22): 4768-4778.

Lafferty, J.D., McCallum, A. and Pereira, F.C.N. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Brodley, C.E. and Danyluk, A.P. (eds), *Proceedings of the 18th International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, USA, 282-289.

Lampros, C., Costas, P., Exarchos, T.P., Yorgos, G. and Fotiadis, D.I. (2007) Sequence-based Protein Structure Prediction using A Reduced State-space Hidden Markov Model, *Computers in Biology and Medicine*. 37(9): 1211-1224.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007) Clustal W and Clustal X Version 2.0, *Bioinformatics*. 23(21): 2947-2948.

Lassmann, T. and Sonnhammer, E.L. (2002) Quality Assessment of Multiple Alignment Programs, *FEBS Letters*. 529(1): 126-130.

Lee, J. and Wang, X. (2006) Pairwise Sequence Analysis using Information Specific Algorithm. *Proceedings of the 6th IEEE International Conference on Computer and Information Technology*. IEEE Computer Society, Seoul, Korea, 209-209.

Lee, M.M., Chan, M.K. and Bundschuh, R. (2009) SIB-BLAST: A Web Server for Improved Delineation of True and False Positives in PSI-BLAST Searches, *Nucleic Acids Research*. http://nar.oxfordjournals.org/cgi/content/abstract/gkp301v301.

Leibowitz, N., Nussinov, R. and Wolfson, H.J. (2001) MUSTA - A General, Efficient, Automated Method for Multiple Structure Alignment and Detection of Common Motifs: Application to Proteins, *Journal of Computational Biology*. 8(2): 93-121.

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.

Leslin, C.M., Abyzov, A. and Ilyin, V.A. (2007) TOPOFIT-DB, A Database of Protein Structural Alignments Based on the TOPOFIT Method, *Nucleic Acids Research*. 35(Database Issue): D317-321.

Liang, Y.-M., Shih, S.-W., Chun-Chieh Shih, A., Liao, H.Y.M. and Lin, C.-C. (2009) Learning Atomic Human Actions Using Variable-Length Markov Models, *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*. 39(1): 268-280.

Liao, L. and Noble, W.S. (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.

Lin, K., Simossis, V.A., Taylor, W.R. and Heringa, J. (2005) A Simple and Fast Secondary Structure Prediction Method using Hidden Neural Networks, *Bioinformatics*. 21(2): 152-159.

Lo Conte, L., Brenner, S.E., Hubbard, T.J.P., Chothia, C. and Murzin, A.G. (2002) SCOP Database in 2002: Refinements Accommodate Structural Genomics, *Nucleic Acids Research*. 30(1): 264-267.

Logan, B.T., Karaoz, U., Moreno, P.J., Weng, Z. and Kasif, S. (2004) Protein Seer: A Web Server for Protein Homology Detection. *Proceedings of 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2004*. IEEE, San Francisco, California, 3064-3067.

Lu, Y. and Sze, S.H. (2008) Multiple Sequence Alignment Based on Profile Alignment of Intermediate Sequences, *Journal of Computational Biology*. 15(7): 767-777.

Lu, Y. and Sze, S.H. (2009) Improving Accuracy of Multiple Sequence Alignment Algorithms Based on Alignment of Neighboring Residues, *Nucleic Acids Research*. 37(2): 463-472.

Lupyan, D., Leo-Macias, A. and Ortiz, A.R. (2005) A New Progressive-iterative Algorithm for Multiple Structure Alignment, *Bioinformatics*. 21(15): 3255-3263.

Madera, M. (2008) Profile Comparer: A Program for Scoring and Aligning Profile Hidden Markov Models, *Bioinformatics*. 24(22): 2630-2631.

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.

Madhusudhan, M.S., Webb, B.M., Marti-Renom, M.A., Eswar, N. and Sali, A. (2009) Alignment of Multiple Protein Structures Based on Sequence and Structure Features, *Protein Engineering, Design and Selection*. gzp040.

Majoros, W.H., Pertea, M. and Salzberg, S.L. (2005) Efficient Implementation of A Generalized Pair Hidden Markov Model for Comparative Gene Finding, *Bioinformatics*. 21(9): 1782-1788.

Manohar, A. and Batzoglou, S. (2005) TreeRefiner: A Tool for Refining A Multiple Alignment on A Phylogenetic Tree. *Proceeding of the 4th International IEEE Computer Society Computational Systems Bioinformatics Conference*. IEEE, Stanford, California, 111-119.

Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Tasneem, A., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N. and Bryant, S.H. (2009) CDD: Specific Functional Annotation with the Conserved Domain Database, *Nucleic Acids Research*. 37(Suppl_1): D205-210.

Marti-Renom, M.A., Pieper, U., Madhusudhan, M.S., Rossi, A., Eswar, N., Davis, F.P., Al-Shahrour, F., Dopazo, J. and Sali, A. (2007) DBAli tools: Mining the Protein Structure Space, *Nucleic Acids Research*. 35(Suppl_2): W393-397.

McCallum, A. (2003) Efficiently Inducing Features of Conditional Random Fields. *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI03)*. Morgan Kaufmann, Acapulco, Mexico, 403-410.

McGuffin, L.J. and Jones, D.T. (2003) Improvement of the GenTHREADER Method for Genomic Fold Recognition, *Bioinformatics*. 19(7): 874-881.

McGuffin, L.J., Street, S.A., Bryson, K., Sorensen, S.-A. and Jones, D.T. (2004) The Genomic Threading Database: A Comprehensive Resource for Structural Annotations of the Genomes from Key Organisms, *Nucleic Acids Research*. 32(Suppl_1): D196-199.

McLachlan, G.J. (2004) *Discriminant Analysis and Statistical Pattern Recognition*. New York, USA: John Wiley & Sons, Inc.

McLachlan, G.J. and Peel, D. (2000) *Finite Mixture Models*. New York, USA: John Wiley & Sons, Inc.

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.

Melvin, I., Weston, J., Leslie, C. and Noble, W. (2008) Combining Classifiers for Improved Classification of Proteins from Sequence or Structure, *BMC Bioinformatics*. 9(1): 389-398.

Mendel, M. (1992) A Commercial Large-Vocabulary Discrete Speech Recognition System: Dragon Dictate, *Language Speech* 35(Pt 1-2): 237-246.

Menke, M., Berger, B. and Cowen, L. (2008) Matt: Local Flexibility aids Protein Multiple Structure Alignment, *PLoS Computational Biology*. 4(1): e10.

Mizuguchi, K. and Blundell, T. (2000) Analysis of Conservation and Substitutions of Secondary Structure Elements within Protein Superfamilies, *Bioinformatics*. 16(12): 1111-1119.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures, *Journal of Molecular Biology*. 247(4): 536-540.

Neuwald, A.F. (2009) Rapid Detection, Classification and Accurate Alignment of up to A Million or More Related Protein Sequences, *Bioinformatics*. 25(15): 1869-1875.

Neuwald, A.F. and Poleksic, A. (2000) PSI-BLAST Searches using Hidden Markov Models of Structural Repeats: Prediction of An Unusual Sliding DNA Clamp and of ß-propellers in UV-damaged DNA-binding Protein, *Nucleic Acids Research*. 28(18): 3570-3580.

Ng, A.Y. and Jordan, M.I. (2001) On Discriminative vs Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. In Dietterich, T., Becker, S. and Ghahramani, Z. (eds), *Advances in Neural Information Processing Systems (NIPS) 14*. MIT Press, Vancouver, Canada, 841-848.

Nguyen, N. and Guo, Y. (2007) Comparisons of Sequence Labeling Algorithms and Extensions. *Proceedings of the 24th International Conference on Machine Learning*. ACM, Corvalis, Oregon, 681-688.

Nigsch, F., Bender, A., Jenkins, J.L. and Mitchell, J.B.O. (2008) Ligand-target Prediction using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics, *Journal of Chemical Information and Modeling*. 48(12): 2313-2325.

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.

Notredame, C., Holm, L. and Higgins, D.G. (1998) COFFEE: An Objective Function for Multiple Sequence Alignments, *Bioinformatics*. 14(5): 407-422.

Notredame, C. and Suhre, K. (2004) Computing Multiple Sequence/Structure Alignments with the T-Coffee Package, *Current Protocols in Bioinformatics*. Chapter 3(Unit 3.8): http://mrw.interscience.wiley.com/emrw/9780471250951/cp/cpbi/article/ bi9780471250308/current/abstract.

Nugent, T. and Jones, D. (2009) Transmembrane Protein Topology Prediction using Support Vector Machines, *BMC Bioinformatics*. 10(1): 159-181.

Nuin, P., Wang, Z. and Tillier, E. (2006) The Accuracy of Several Multiple Sequence Alignment Programs for Proteins, *BMC Bioinformatics*. 7(1): 471-489.

Ochagavía, M.E. and Wodak, S. (2004) Progressive Combinatorial Algorithm for Multiple Structural Alignments: Application to Distantly Related Proteins, *Proteins*. 55(2): 436-454.

Ohlson, T. and Elofsson, A. (2005) ProfNet, A Method to Derive Profile-profile Alignment Scoring Functions that Improves the Alignments of Distantly Related Proteins, *BMC Bioinformatics*. 6(1): 253.

Oldfield, T. (2007) CAALIGN: A Program for Pairwise and Multiple Protein-structure Alignment, *Acta Crystallographica Section D*. 63(4): 514-525.

Orengo, C.A. and Taylor, W.R. (1996) SSAP: Sequential Structure Alignment Program for Protein Structure Comparison, *Methods in Enzymology*. 266(1): 617-635.

Ortiz, A.R., Strauss, C.E.M. and Olmea, O. (2002) MAMMOTH (Matching Molecular Models Obtained from Theory): An Automated Method for Model Comparison, *Protein Science*. 11(11): 2606–2621.

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.

Pavlidis, P., Wapinski, I. and Noble, W.S. (2004) Support Vector Machine Classification on the Web, *Bioinformatics*. 20(4): 586-587.

Pearson, W.R. (1990) Rapid and Sensitive Sequence Comparison with FASTP and FASTA, *Methods Enzymol*. 183(1): 63-98.

Pearson, W.R. and Lipman, D.J. (1988) Improved Tools for Biological Sequence Comparison, *Proceedings of the National Academy of Sciences of the United States of America*. 85(8): 2444-2448.

Pei, J. and Grishin, N.V. (2006) MUMMALS: Multiple Sequence Alignment Improved by using Hidden Markov Models with Local Structural Information, *Nucleic Acids Research*. 34(16): 4364-4374.

Pei, J. and Grishin, N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins, *Bioinformatics*. 23(7): 802–808.

Pei, J., Kim, B.-H. and Grishin, N.V. (2008) PROMALS3D: A Tool for Multiple Protein Sequence and Structure Alignments, *Nucleic Acids Research*. 36(7): 2295-2300.

Peng, J., Zhang, P. and Riedel, N. (2008) Discriminant Learning Analysis, *IEEE Transactions on Systems, Man and Cybernetics*, *Part B: Cybernetics*. 38(6): 1614-1625.

Petrey, D., Xiang, Z., Tang, C.L., Xie, L., Gimpelev, M., Mitros, T., Soto, C.S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., Koh, I.Y., Alexov, E. and Honig, B. (2003) Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling, *Proteins*. 53(430-5).

Pirooznia, M. and Deng, Y. (2006) SVM Classifier - A Comprehensive Java Interface for Support Vector Machine Classification of Microarray Data, *BMC Bioinformatics*. 7(Suppl_4): S25.

Poirot, O., O'Toole, E. and Notredame, C. (2003) Tcoffee@igs: A Web Server for Computing, Evaluating and Combining Multiple Sequence Alignments, *Nucleic Acids Research*. 31(13): 3503-3506.

Poirot, O., Suhre, K., Abergel, C., O'Toole, E. and Notredame, C. (2004) 3DCoffee@igs: A Web Server for Combining Sequences and Structures Into A Multiple Sequence Alignment, *Nucleic Acids Research*. 32(Web Server Issue): W37-40.

Rabiner, L.R. (1990) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In Waibel, A. and Lee, K.-F. (eds), *Readings in Speech Recognition*. Morgan Kaufmann Publishers, Inc, San Francisco, USA, 267-296.

Raghava, G.P.S., Searle, S., Audley, P., Barber, J. and Barton, G. (2003) OXBench: A Benchmark for Evaluation of Protein Multiple Sequence Alignment Accuracy, *BMC Bioinformatics*. 4(1): 47-70.

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): 455-471.

Rausch, T., Emde, A.-K., Weese, D., Doring, A., Notredame, C. and Reinert, K. (2008) Segment-based Multiple Sequence Alignment, *Bioinformatics*. 24(16): i187-192.

Reid, A.J., Yeats, C. and Orengo, C.A. (2007) Methods of Remote Homology Detection Can Be Combined to Increase Coverage by 10% in the Midnight Zone, *Bioinformatics*. 23(18): 2353-2360.

Remmert, M., Linke, D., Lupas, A.N. and Soding, J. (2009) HHomp-prediction and Classification of Outer Membrane Proteins, *Nucleic Acids Research*. 37(Web Server Issue): W446-451.

Riccardo, C., Gianni, P., Eugenio, U. and Humberto, G.-D. (2009) Computational Chemistry Study of 3D-structure-function Relationships for Enzymes Based on Markov Models for Protein Electrostatic, HINT, and Van der Waals Potentials, *Journal of Computational Chemistry*. 30(9): 1510-1520.

Rish, I. (2001) An Empirical Study of the Naive Bayes Classifier. *Proceedings of IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., Seattle, USA.

Rodrigues, A.P., Grant, B.J., Godzik, A. and Friedberg, I. (2007) The 2006 Automated Function Prediction Meeting, *BMC Bioinformatics*. 8(Suppl_4): S1-4.

Rost, B. (1999) Twilight Zone of Protein Sequence Alignments, *Protein Engineering*. 12(2): 85-94.

Rubinsky, L., Raichman, N., Lavee, J., Frenk, H. and Ben-Jacob, E. (2008) Spatio-temporal Motifs 'Remembered' in Neuronal Networks Following Profound Hypothermia, *Neural Networks*. 21(9): 1232-1237.

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.

Sadreyev, R.I., Baker, D. and Grishin, N.V. (2003) Profile-profile Comparisons by COMPASS Predict Intricate Homologies Between Protein Families, *Protein Science*. 12(10): 2262-2272.

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

Sato, K., Morita, K. and Sakakibara, Y. (2008) PSSMTS: Position Specific Scoring Matrices on Tree Structures, *Journal of Mathematical Biology*. 56(1-2): 201-214.

Schapire, R.E. (1990) The Strength of Weak Learnability, *Machine Learning*. 5(2): 197-227.

Shao, X., Tian, Y., Wu, L., Wang, Y., Jing, L. and Deng, N. (2009) Predicting DNA and RNA-binding Proteins from Sequences with Kernel Methods, *Journal of Theoretical Biology*. 258(2): 289-293.

Shatsky, M., Nussinov, R. and Wolfson, H.J. (2004) A Method for Simultaneous Alignment of Multiple Protein Structures, *Proteins*. 56(1): 143-156.

Sheinerman, F.B., Al-Lazikani, B. and Honig, B. (2003) Sequence, Structure and Energetic Determinants of Phosphopeptide Selectivity of SH2 Domains, *Journal of Molecular Biology*. 334(4): 823-841.

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

Soding, J. (2005) Protein Homology Detection by HMM-HMM Comparison, *Bioinformatics*. 21(7): 951-960.

Sonego, P., Kocsor, A. and Pongor, S. (2008) ROC Analysis: Applications to the Classification of Biological Sequences and 3D Structures, *Briefings in Bioinformatics*. 9(3): 198-209.

Stebbings, L.A. and Mizuguchi, K. (2004) HOMSTRAD: Recent Developments of the Homologous Protein Structure Alignment Database, *Nucleic Acids Research*. 32(Database Issue): D203-207.

Subramanian, A., Kaufmann, M. and Morgenstern, B. (2008) DIALIGN-TX: Greedy and Progressive Approaches for Segment-based Multiple Sequence Alignment, *Algorithms for Molecular Biology*. 3(1): 6-17.

Subramanian, A., Weyer-Menkhoff, J., Kaufmann, M. and Morgenstern, B. (2005) DIALIGN-T: An Improved Algorithm for Segment-based Multiple Sequence Alignment, *BMC Bioinformatics*. 6(1): 66-79.

Suchard, M.A. and Redelings, B.D. (2006) BAli-Phy: Simultaneous Bayesian Inference of Alignment and Phylogeny, *Bioinformatics*. 22(16): 2047-2048.

Supper, J., Spangenberg, L., Planatscher, H., Draeger, A., Schroeder, A. and Zell, A. (2009) BowTieBuilder: Modeling Signal Transduction Pathways, *BMC Systems Biology*. 3(1): 67.

Tang, C.L., Xie, L., Koh, I.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.

Taylor, W.R. (1988) A Flexible Method to Align Large Numbers of Biological Sequences, *Journal of Molecular Evolution*. 28(1-2): 161-169.

Taylor, W.R., Flores, T.P. and Orengo, C.A. (1994) Multiple Protein Structure Alignment, *Protein Science*. 3(10): 1858-1870.

Taylor, W.R. and Orengo, C.A. (1989) Protein Structure Alignment, *Journal of Molecular Biology*. 208(1): 1-22.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: 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.

Titterington, D., Smith, A. and Makov, U. (1985) *Statistical Analysis of Finite Mixture Distributions*. New York, USA: John Wiley & Sons, Inc.

Tsochantaridis, I., Homann, T., Joachims, T. and Altun, Y. (2004) Support Vector Machine Learning for Interdependent and Structured Output Spaces. *Proceedings of the 21st International Conference on Machine Learning*. ACM, Alberta, Canada, 104-112.

Van Walle, I., Lasters, I. and Wyns, L. (2005) SABmark - A Benchmark for Sequence Alignment That Covers the Entire Known Fold Space, *Bioinformatics*. 21(7): 1267-1268.

Wallace, I.M., O'Sullivan, O. and Higgins, D.G. (2005) Evaluation of Iterative Alignment Algorithms for Multiple Alignment, *Bioinformatics*. 21(1): 1408-1414.

Walraven, J.M., Trent, J.O. and Hein, D.W. (2007) Computational and Experimental Analyses of Mammalian Arylamine N-Acetyltransferase Structure and Function, *Drug Metabolism and Disposition*. 35(6): 1001-1007.

Wang, J. and Feng, J.A. (2005) NdPASA: A Novel Pairwise Protein Sequence Alignment Algorithm that Incorporates Neighbor-dependent Amino Acid Propensities, *Proteins: Structure, Function, and Bioinformatics*. 58(3): 628-637.

Wang, L. and Sauer, U.H. (2008) OnD-CRF: Predicting Order and Disorder in Proteins Conditional Random Fields, *Bioinformatics*. 24(11): 1401-1402.

Wang, L., Xue, P. and Chan, K.L. (2008) Two Criteria for Model Selection in Multiclass Support Vector Machines, *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*. 38(6): 1432-1448.

Wang, Q., Song, E., Jin, R., Han, P., Wang, X., Zhou, Y. and Zeng, J. (2009) Segmentation of Lung Nodules in Computed Tomography Images using Dynamic Programming and Multidirection Fusion Techniques, *Academic Radiology*. 16(6): 678-688.

Webb, B.J.M., Liu, J.S. and Lawrence, C.E. (2002) BALSA: Bayesian Algorithm for Local Sequence Alignment, *Nucleic Acids Research*. 30(5): 1268-1277.

Weston, J., Leslie, C., Ie, E., Zhou, D., Elisseeff, A. and Noble, W.S. (2005) Semi-supervised Protein Classification using Cluster Kernels, *Bioinformatics*. 21(15): 3241-3247.

Wilm, A., Higgins, D.G. and Notredame, C. (2008) R-Coffee: A Method for Multiple Alignment of Non-coding RNA, *Nucleic Acids Research*. 36(9): e52-62.

Wistrand, M. and Sonnhammer, E. (2005) Improved Profile HMM Performance by Assessment of Critical Algorithmic Features in SAM and HMMER, *BMC Bioinformatics*. 6(1): 99-109.

Won, K.J., Hamelryck, T., Prugel-Bennett, A. and Krogh, A. (2007) An Evolutionary Method for Learning HMM Structure: Prediction of Protein Secondary Structure, *BMC Bioinformatics*. 8(1): 357-370.

Wu, C.H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z.-Z., Ledley, R.S., Lewis, K.C., Mewes, H.-W., Orcutt, B.C., Suzek, B.E., Tsugita, A., Vinayaka, C.R., Yeh, L.-S.L., Zhang, J. and Barker, W.C. (2002) The Protein Information Resource: An Integrated Public Resource of Functional Annotation of Proteins, *Nucleic Acids Research*. 30(1): 35-37.

Wu, C.H. and Nebert, D.W. (2004) Update on Human Genome Completion and Annotations: Protein Information Resource, *Human Genomics*. 1(3): 229-233.

Xia, X., Zhang, S., Su, Y. and Sun, Z. (2009) MICAlign: A Sequence-to-structure Alignment Tool integrating Multiple Sources of Information in Conditional Random Fields, *Bioinformatics*. 25(11): 1433-1434.

Xu, Y. and Xu, D. (2000) Protein Threading using PROSPECT: Design and Evaluation, *Proteins: Structure, Function, and Genetics*. 40(3): 343-354.

Xu, Y., Xu, D., Crawford, O.H., Einstein, J.r., Larimer, F., Uberbacher, E., Unseren, M.A. and Zhang, G. (1999) Protein Threading by PROSPECT: A Prediction Experiment in CASP3, *Protein Engineering*. 12(11): 899-907.

Yan, R.X., Si, J.N., Wang, C. and Zhang, Z. (2009) DescFold: A Web Server for Protein Fold Recognition, *BMC Bioinformatics*. 10(1): 416.

Yang, Y., Tantoso, E. and Li, K.B. (2008) Remote Protein Homology Detection using Recurrence Quantification Analysis and Amino Acid Physicochemical Properties, *Journal of Theoretical Biology*. 252(1): 145-154.

Ye, Y. and Godzik, A. (2005) Multiple Flexible Structure Alignment using Partial Order Graphs, *Bioinformatics*. 21(10): 2362-2369.

Zaki, N.M. and Deris, S. (2004) *Learning Enhancements Of Support Vector Machine For Detecting Remote Protein Homology*. Universiti Teknologi Malaysia, Skudai.

Zhou, H. and Zhou, Y. (2005) SPARKS 2 and SP3 Servers in CASP6, *Proteins: Structure, Function, and Bioinformatics*. 61(S7): 152-156.