PROTEIN SECONDARY STRUCTURE PREDICTION FROM AMINO ACID SEQUENCES USING A NEURAL NETWORK CLASSIFIER BASED ON THE DEMPSTER-SHAFER THEORY

SATYA NANDA VEL ARJUNAN

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

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

MARCH 2003

To Motorcycle Appa

ACKNOWLEDGMENTS

This thesis is an outcome of the contributions of many people. Professor Safaai Deris was always available to support my research effort during my years at UTM. Despite of his busy schedule, he was constantly able to offer intelligent advice, comments, criticisms and suggestions whenever I had consulted him. I have also particularly appreciated his immediate and very thoughtful responses to my problems. The enthusiasm of Professor Rosli Illias in molecular biology has inspired me to delve deeper into the field of computational biology. He has given me valuable insights in aspects relating to structural biology, which have helped me to gain a correct perspective of my research in protein secondary structure.

Professor Thierry Denoeux has been very helpful and kind with my trivial queries relating to his proposed network and the field of pattern classification in general. Nazar Zaki has given me much advice about research and graduate work. My sincere appreciation also goes to all the people that have answered my email queries so promptly, the open source and Slashdot community, the maintainers of Google, CiteSeer and DBLP, and to all those authors who have generously made their publications and lecture notes available online.

On a personal note, I would like to thank my parents, brothers and sister, and Gracy for their love, support and understanding throughout this endeavour.

ABSTRACT

Recently, Denoeux proposed a novel neural network classifier based on the Dempster-Shafer theory. Several of his preliminary experiments in some typical problems demonstrated that the classifier has an excellent performance when compared to other statistical and machine learning approaches. However, up to now there has been little further work reported pertaining to its improvements or applications. As a result, this research extends the initial work by examining its potential improvements and applicability in a new real world task such as the protein secondary structure prediction. In order to reduce the computational demand when training with large data of proteins, an interface was developed using the data parallel approach to parallelize the training phase of the classifier and other accompanying methods such as data clustering algorithms. The parallelized classifier also permitted rigorous experiments to be conducted in two other benchmark problems with disparate dimensions to determine the classifier's inherent attributes and drawbacks. The experiments showed that although the classifier performed better than some of the best methods such as Support Vector Machines and Kernel Fisher Discriminants in the small dimensional problem (dimension size = 9), its performance deteriorated significantly in the higher dimensional problem (dimension size = 60). This presented a substantial challenge because the secondary structure prediction exhibits high dimensionality as well. An improved version of the classifier was designed by introducing Multilayer Perceptrons to replace the distance measure of the classifier, which appeared to be impaired in high dimensions. The results of the secondary structure prediction demonstrated that the new classifier performed better than the original one. Moreover, at the level of sequence-to-structure prediction, its performance was comparable to the PHD (Profile network from Heidelberg) method, which is one of the best secondary structure prediction schemes.

ABSTRAK

Baru-baru ini, Denoeux telah mengutarakan sebuah pengkelas rangkaian neural berasaskan teori Dempster-Shafer. Beberapa eksperimen awal beliau dalam masalah tipikal telah menunjukkan pengkelas tersebut mempunyai prestasi yang cemerlang apabila dibandingkan dengan teknik-teknik statistik dan pembelajaran mesin lain. Namun demikian, sehingga kini hampir tiada kerja lanjutan dilaporkan mengenai perkembangan atau aplikasi teknik tersebut. Oleh yang demikian, kajian ini menyambung hasil kerja awal tersebut dengan meneliti kebolehgunaan dan prestasinya dalam masalah dunia sebenar lain seperti peramalan struktur sekunder protein dan menyelidiki sama ada sebarang pembaikan dapat dilakukan terhadap teknik pengkelasan tersebut. Untuk menangani beban komputasi hasil daripada data protein yang besar, sebuah antaramuka menggunakan kaedah penselarian data telah dibangunkan untuk menselarikan fasa latihan pengkelas dan teknik-teknik komputasi yang mengiringinya seperti algoritma pengkelompok data. Pengkelas yang diselarikan tersebut juga telah membolehkan banyak eksperimen dijalankan dalam dua masalah lain yang berdimensi berbeza untuk mengenalpasti ciri-ciri serta kelemahannya. Hasil eksperimen tersebut menunjukkan sungguhpun keputusan pengkelas tersebut menandingi beberapa teknik terbaik seperti Mesin Vektor Sokongan dan Pengdiskriminasi Kernel Fisher dalam masalah berdimensi kecil (saiz dimensi = 9), namun prestasinya jatuh mendadak dalam masalah berdimensi besar (saiz dimensi = 60). Sebuah pengkelas yang lebih baik telah direka dengan memperkenalkan Perseptron Berbilang Aras untuk menggantikan kaedah ukuran jarak yang didapati terjejas dalam dimensi besar. Hasil peramalan struktur sekunder menunjukkan prestasi pengkelas yang telah dibangunkan menandingi yang sedia ada. Lebih-lebih lagi, pada aras peramalan jujukan-ke-struktur, prestasinya adalah setara dengan kaedah PHD (*Profile network from Heidelberg*), iaitu salah satu teknik peramalan struktur sekunder yang terbaik.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

a	–	current dimension under consideration $(1,\cdots,nDimension)$
A_b	–	net input of a hidden unit
α^j	–	parameter denoting the relative importance of a prototype
b	–	current hidden unit under consideration $(1,\cdots,nHidden)$
B_c	–	net input of a output unit
β_q^j	–	class membership strength of a prototype
c	–	current output unit under consideration $(1,\cdots,nOutput)$
colBlk	–	column block size
contxt	–	context of the process grid
$\vec{d}^j$	–	distance vector between a prototype and a data sample
$E_{\alpha MLPDS}$	–	empirical error of the αMLPDS classifier
E_{DBNN}	–	empirical error of the DBNN classifier
$E_{\gamma DBNN}$	–	empirical error of the γDBNN classifier
E_{MLP}	–	empirical error of the MLP classifier
E_{MLPDS}	–	empirical error of the MLPDS classifier
ε	–	desired empirical error limit
η_0	–	initial learning rate
F	–	number of subsets used for a cross validation
γ^j	–	receptive field parameter of a prototype (DBNN)
$\vec{\gamma}^j$	–	receptive field parameter vector of a prototype (γDBNN)
gblCols	–	global columns
gblRows	–	global rows

j	–	current prototype under consideration $(1,\cdots,nProtos)$
k	–	number of clusters
λ	–	cost function to distribute the output ignorance mass
lclCols	–	local columns
lclRows	–	local rows
lower	–	learning rate decrement factor
maxIter	–	maximum number of iterations
$\vec{m}^j$	–	basic belief assignment
m_q^j	–	belief mass of the basic belief assignment
$m_{nClasses+1}^j$	–	ignorance mass of the basic belief assignment
mk_q	–	output of the classifier
$mk_{nClasses+1}$	–	output ignorance mass of the classifier
μ	–	regularization parameter
n	–	current sample under consideration $(1,\cdots,nSamples)$
nClasses	–	number of classes
nDimension	–	dimension size
nHidden	–	number of hidden units
nOutput	–	number of output units
nProcs	–	total number of processes
nSamples	–	number of data samples
O_q	–	normalized output of the classifier
$O_{nClasses+1}$	–	normalized output ignorance mass of the classifier
$\vec{p}^j$	–	weight vector of a prototype
prCol	–	process column coordinate of the process grid
prCols	–	total process columns of the process grid
prRow	–	process row coordinate of the process grid
prRows	–	total process rows of the process grid
q	–	current class under consideration $(1,\cdots,nClasses)$
Q_q	–	output of the classifier in terms of pignistic probability
rowBlk	–	row block size

s^j	–	activation function of a prototype
startPrCol	–	starting process column
startPrRow	–	starting process row
T_q^n	–	target output class of the classifier
u_q^j	–	normalized class membership strength of a prototype
upper	–	learning rate increment factor
$\vec{V}_b$	–	input weight vector
$\vec{V}_b^j$	–	input weight vector of a prototype-MLP
$\vec{W}_c$	–	output weight vector
$\vec{W}^j$	–	output weight vector of a prototype-MLP
$\vec{X}^n$	–	data sample vector
Y_c	–	activation function of a output unit
Z_b	–	activation function of a hidden unit

LIST OF ABBREVIATIONS

AB	–	AdaBoost
ABR	–	Regularized AdaBoost
ATLAS	–	Automatically Tuned Linear Algebra Software
BBA	–	Basic Belief Assignment
BLAS	–	Basic Linear Algebra Subprograms
DBNN	–	Denoeux Belief Neural Network
DDPI	–	Distributed Data Partitioning Interface
dgemm	–	double precision generalized matrix multiply
DNA	–	Deoxyribonucleic Acid
D-S	–	Dempster-Shafer
DSSP	–	Database of Secondary Structure in Proteins
EVA	–	EValuation of Automatic protein structure prediction
HMM	–	Hidden Markov Model
KFD	–	Kernel Fisher Discriminant
k-NN	–	k-Nearest Neighbour
LHS	–	Left-Hand Side
MLP	–	Multilayer Perceptron
MLPDS	–	Multilayer Perceptrons with Dempster-Shafer Theory
MPI	–	Message Passing Interface
NMR	–	Nuclear Magnetic Resonance
PBLAS	–	Parallel Basic Linear Algebra Subprograms
PDB	–	Protein Data Bank
pdgemm	–	parallel double precision generalized matrix multiply
PHD	–	Profile network from HeiDelberg
RBF	–	Radial Basis Function Classifier
RNA	–	Ribonucleic Acid

RS126	–	126 Non-Homologous Proteins by Rost and Sander
ScaLAPACK	–	Scalable Linear Algebra Package
SMP	–	Symmetrical Multi-Processor
SVM	–	Support Vector Machine
UCI	–	University of California at Irvine
αMLPDS	–	Collaborative MLPs with Dempster-Shafer theory
γDBNN	–	Weighted Prototype Vector Element DBNN

LIST OF APPENDICES

CHAPTER I

INTRODUCTION

1.1 Introduction

The purpose of this research is to investigate the properties of a recently introduced neural network classifier based on the Dempster-Shafer theory of evidence and its applicability in protein secondary structure prediction. This chapter begins by introducing the context and motivations of protein secondary structure prediction. Since it assumes some familiarity with protein structure, a review of protein structure is given in Appendix A. For quick references, a glossary of structural genomic terms is also provided in Appendix B. A review of existing secondary structure prediction methods is given in this chapter before describing the challenges posed by the prediction problem. The basic concepts underlying the Dempster-Shafer theory of evidence in relation to the neural network classifier is described in this chapter as well. The motivations and objectives of the research are presented in the final sections of the chapter.

1.2 Context and Motivations of Protein Secondary Structure Prediction

Proteins, the fundamental molecules of all organisms have three-dimensional structures that are specified by sequences of amino acids. The functional properties of a protein are determined by the conformation of its structure, through which

interactions with various substrates such as DNA, RNA and other proteins take place. Thus, elucidating the structure of a protein becomes a prerequisite to understanding its function. Some of these functions include serving as catalysts, transport agents and as building blocks of life. The structure of a protein relays the function through grooves and compartments into which other proteins and molecules fit like keys into a keyhole. Rational drug design is an approach which exploits this property to manufacture targeted drugs that can dock into these keyholes, inactivating them or turning them on, much like other proteins and molecules do. With this approach, drugs of the future will be custom designed not only to cater for specific diseases such as cystic fibrosis, Mad Cow disease, Alzheimer's disease, and cancer, but for specific people as well.

Rational drug design is one of the strongest proponents for the inception of two new disciplines in molecular biology: protein engineering, where protein functions are altered by mutating genes of existing proteins, and protein design, where novel proteins are contrived from scratch (Branden and Tooze, 1999). These two fields and other disciplines that are heavily dependent on the function and thus, the structure of proteins, are facing a stumbling block in determining the native structure of a target sequence of amino acids due to the protein folding problem. The NP-complete problem of protein folding is described by the Levinthal's paradox (Levinthal, 1969): for a 100-residue protein with three possible conformations per residue, it would take about 10^{27} years to fully search through all of its $3^{100} \cong 10^{48}$ distinct conformations; yet, naturally-occurring proteins fold reliably to their native state on a time scale of seconds to minutes. As a result of this complex problem, several ways are attempted to predict the structure instead using computational methods. Experimental approaches such as X-ray crystallographic or Nuclear Magnetic Resonance (NMR) on the other hand, are not favourable because in addition to posing restrictions, they are also expensive, laborious and time consuming, taking months or even years to complete. This in fact is illustrated by the total number of experimentally determined structures being only 16,500 (Berman et al., 2002) whereas there are over a million known protein sequences (Wu et al., 2002). Figure 1.1 depicts the four principal techniques that can be employed to predict the structure.

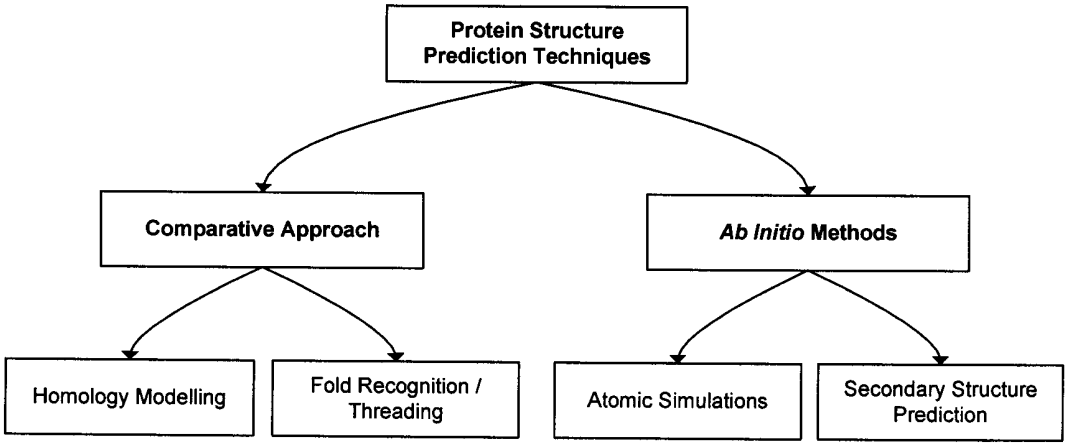

Figure 1.1: Computational methods of protein structure prediction.

Among these techniques, the comparative based approaches, which predict protein structure by referring to sequences of known structure, are the most productive as they can increase the number of known structures by a factor of 10 (Dodge et al., 1998). In homology modelling, the known structures, determined experimentally using the aforementioned approaches, are scrutinized for locations at which significant sequence identity $(> 25\%)$ with the target sequence exists. By assembling the structural information at these homologous locations from various proteins and optimizing it under spatial and conventional energy restraints, the full three-dimensional structure of the target sequence can be predicted. The MODELLER tool (Marti-Renom et al., 2000) for example, predicts the three-dimensional structure by satisfying the spatial restraints, whereas the SWISS-MODEL server (Guex and Peitsch, 1997) adopts the energy based optimization method.

Despite of having limited pairwise sequence identity, two naturally evolved proteins can still fold into homologous structures. The other comparative based approach, fold recognition addresses such sequences by attempting to find structural similarities that are not accompanied by any observable sequence similarity. This is done by threading through known structures in a database to detect and align remote homologues with a scoring function that assesses the fit of the target sequence. Consequently, an efficient threading tool such as PROSPECT (Xu and Xu, 2000) is preferred over homology modelling based tools when the target sequence has less

than 25% pairwise sequence identity. Despite of their success in increasing the number of known structures, both comparative based approaches are impractical when there are no homologues of known structures. Finding the homology between dissimilar sequences as in threading is also very difficult because it is NP-complete (Lathrop, 1994). For these reasons, *ab initio* prediction of structure, using only the information of the amino acid sequences, is considered.

In response to the Levithal's paradox, there are numerous attempts to simulate protein folds at atomistic level from the first principles but with simplified approaches. These approaches include molecular mechanics (MacKerell et al., 1998) and Monte Carlo simulations (Bonneau et al., 2001) that search the energy space for a global minimum, and exhaustive and semi-exhaustive lattice based studies (Skolnick and Kolinski, 1990) which explore all or large amounts of conformational space using an initial approximation of the protein fold. Genetic algorithms (Day et al., 2002) and deterministic global optimization schemes (Klepeis and Floudas, 2003) are also used to minimize the protein potential energy function. These techniques rely on thermodynamic hypothesis of protein folding (Anfinsen, 1973) which suggests that the native structure of a protein sequence corresponds to its global free energy minimum state and accordingly, search for the global minimum. It was found however that the conformation taken by *in vivo* proteins may not necessarily be the global minimum, but the one that is most accessible (Branden and Tooze, 1999). Furthermore, in addition to exhibiting inconsistencies of performance in different proteins, the current best techniques are not able to predict proteins which are more than 150 residues in length (Bonneau and Baker, 2001).

Much of the reason the protein is able to fold quickly to its native state as described by the Levinthal's paradox is that it is directed through pathways (Figure 1.2) rather than random conformational searches. The folds in the pathways follow the structural hierarchy illustrated in Figure 1.3. Hence, an alternative approach to predicting the three-dimensional structure would be to initially reduce the complexity by decomposing it to a one-dimensional problem in the hierarchy, and to extract the crucial features from the amino acid sequence. As such, the secondary structure of proteins, which succeeds the amino acid sequence in the hierarchy, is predicted first as a stepping stone towards the three-dimensional structure. The prediction of

Step 1: **Formation of local secondary structure**
- Development of alpha-helices and beta-sheets which are connected through loops.
- Duration: 5 ms.

Step 2: **Hydrophobic collapse**
- Much of the secondary structure has been formed.
- Transition from secondary to tertiary structure.
- Collapsed state known as a molten globule.

Step 3: **Formation of tertiary structure**
- Packing of alpha-helices and beta-sheets.
- Duration: 5-1000 ms.

Step 4: **Rigidification of protein structure**
- Hydrogen bonding.
- Expulsion of water from protein's interior.
- Duration: 1-100 s.

Figure 1.2: Steps of the protein folding pathways.

secondary structure, considerably less complicated than full tertiary structure prediction (three-dimensional structure), elucidates the type, number and order of secondary structure elements (alpha-helix, beta-sheet, or loop) as shown in Figure 1.4. The goal is to predict which one of these elements is constituted by each residue of the protein.

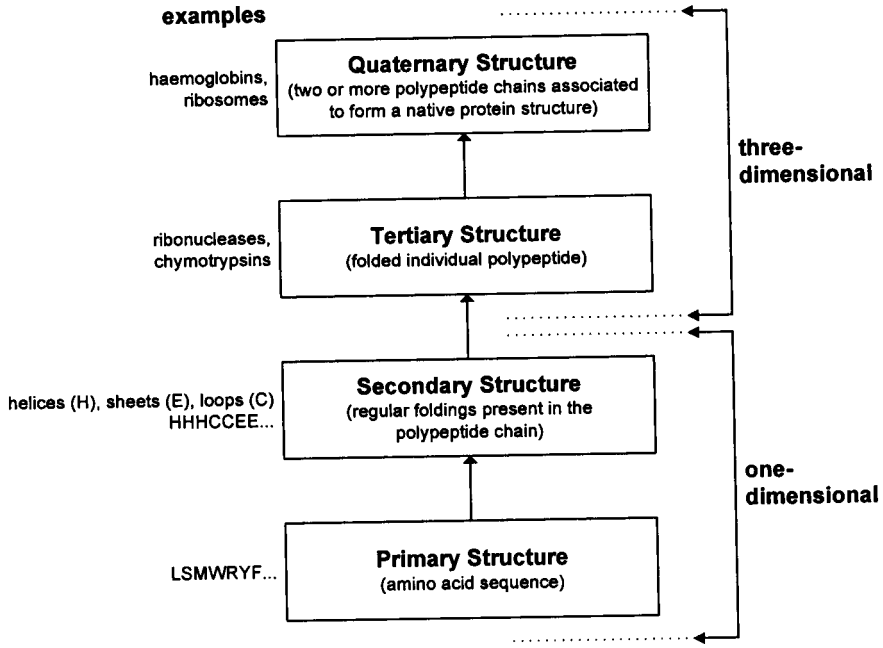

Figure 1.3: Structural hierarchy of proteins.

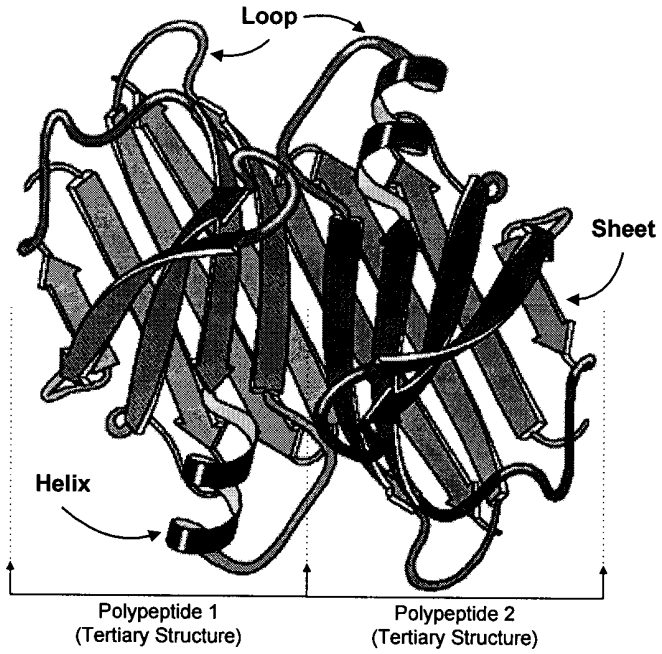

Figure 1.4: Secondary structure elements in the quaternary structure of the protein transthyretin (Richardson and Richardson, 1992).

In addition to three dimensional structure prediction via threading (Xu and Xu, 2000) or atomic simulation (Bonneau et al., 2001; Xia et al., 2000; Eyrich et al., 1999; Chen et el., 1999), predicted secondary structures can also aid in immediate predictions of protein functions (Stawiski et al., 2000), classification of proteins for genome analysis (Przytycka et al., 1999) and in identification of regions of the protein that will likely to undergo structural change (Young et al., 1999). Figure 1.5 displays the significance of protein secondary structure prediction.

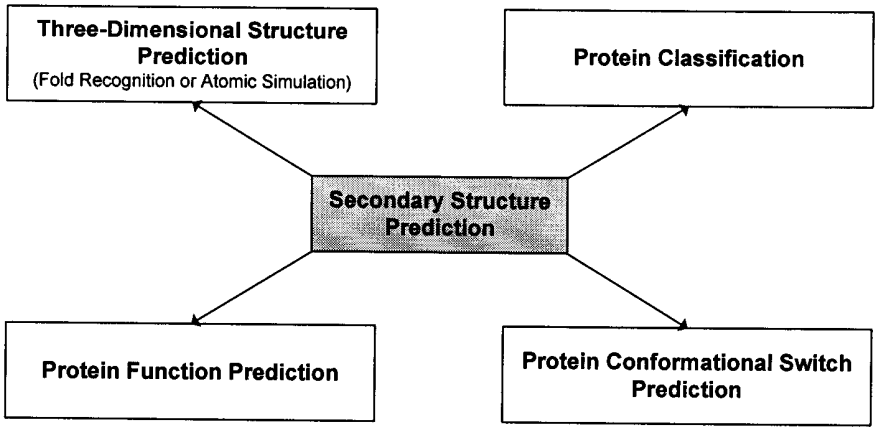

Figure 1.5: Significance of protein secondary structure prediction.

1.3 Review of Protein Secondary Structure Prediction Methods

The significance of secondary structure has been noted for more than four decades. It is evident when the first attempt to correlate amino acids in the protein sequence with the three structural elements was conducted by Blout et al. (1960) about a decade after Pauling and colleagues first suggested the existence of alpha-helices (Pauling and Corey, 1951) and beta-sheets (Pauling et al., 1951). Following that, the field of secondary structure prediction has evolved from statistical methods to neural networks, nearest neighbours, hidden Markov models and recently to kernel methods, correlating with the developments in the machine learning discipline. In this section, the various methods that have been used for the prediction problem will be described.

The most widely used accuracy index for secondary structure prediction is the three-state per residue accuracy Q_3, which gives the percentage of correctly predicted residues in either of the three classes of structural elements, (Rost and Sander, 2000):

$$Q_3 = \left(\frac{P_{helix} + P_{sheet} + P_{loop}}{\text{Total Residues}} \right) \times 100 \qquad (1.1)$$

where P_{helix}, P_{sheet} and P_{loop} are number of residues predicted correctly in class helix, sheet and loop respectively. The three-state per residue accuracy will be used in the following sections to describe the performance of the prediction models.

1.3.1 Statistical Methods

Most approaches in the early era of secondary structure prediction were based on residue statistics. From the limited protein databases in 1960-70s, evidence was obtained for the preference of each amino acid for a particular secondary structure element. For example, in an algorithm developed by Chou and Fasman (1974), the probability of each amino acid to form a helix or a sheet is computed using the

frequencies with which it appears for the corresponding structure element of experimentally derived protein structures. The algorithm predicts a structure element based on the contiguous regions of sequences which collaboratively demonstrated to have the highest probability of forming the structure element. The Chou and Fasman method can accurately predict for about 49.9% of the residues.

The GOR method by Garnier, Osguthorpe and Robson (1978) is another statistical based algorithm which computes the probability for a particular residue based on the adjacent residues in a window like manner. Even though the first version demonstrated only 55.9% correct prediction, the latest (Kloczkowski et al., 2002) which includes a variable window size and more importantly, the evolutionary information, can correctly predict the structure of up to 74.2% of the residues. The evolutionary information is represented by a profile of multiple sequences of amino acids that are derived from known homologues of the target sequence. The advantage of evolutionary information is that it offers long range effects of the residues to the prediction scheme.

1.3.2 Neural Network Methods

Neural networks were first used to predict secondary structure by Qian and Sejnowski (1988). A fully connected Multilayer Perceptron with a single hidden layer of 40 units was used for this purpose. A sliding window consisting of 13 consecutive residues was used as the input to the network to predict the secondary structure of the residue in the middle of the window. The window is used to incorporate neigbourhood influence into the prediction. The network employed three output nodes, each representing a class of the secondary structure. The 20 distinct residues were represented using orthogonal encoding in which each residue is assigned a unique binary vector i.e. (1,0,0...) or (0,1,0...). Therefore, for the network, the input dimension was of size (20 binary bits) × (13 residues) = 260. After training the network with the standard backpropagation algorithm, it had 64.3% correct predictions. The Qian and Sejnowski's work set the platform for numerous neural network based approaches to predict secondary structure. Maclin and Shalvik

(1993) for example, incorporated the Chou and Fasman (1978) residue statistics into the design of their network.

Despite of other attempts, the next significant improvement in neural network based prediction came only when Rost and Sander (1993) incorporated evolutionary information into the neural network. It was the pioneering work which introduced the long range effects via a profile of evolutionary information. The profile, which contains the target sequence along with multiple aligned sequences of homologues of other proteins, is represented by orthogonal encoding similar to that of Qian and Sejnowski's. As such, the values across the multiple sequences are averaged at each residue to form a single input vector to the network. To effectively represent the profile, additional bits were introduced to the input vector at each residue to constitute empty position of the sliding window at the end of the sequence, as well as insertions and deletions in the aligned sequences. The network, referred to as PHD (Profile network from HeiDelberg), has 71.9% accuracy (Rost and Sander, 1994). The significant increase of performance attributed to the evolutionary information has motivated its adoption in almost all ensuing techniques after PHD.

Recently, another neural network based predictor with an alternative approach to incorporating long range influences was reported by Baldi and colleagues (Baldi et al., 1999). Their server, SSpro uses 11 bidirectional recurrent neural networks in different architectural setups to capture this information without overfitting by 'rolling' them along the multiple aligned sequences in both directions until they reach the residue under consideration. The final prediction is computed by using a simple averaging scheme to form an ensemble of all the networks. For the training procedure, SSpro used very large data (over 200,000 samples) to ascertain its performance. To address the computational demand arising from the large data, each of the 11 recurrent neural networks was trained on separate machines concurrently using error backpropagation algorithm. It took about two months for the authors to complete the training procedure. According to the EVA (EValuation of Automatic protein structure prediction) server (Rost and Eyrich, 2001), the latest version of the SSpro server (Pollastri et al., 2002) has an accuracy of 74.5%.

1.3.3 Nearest Neighbour Methods

Nearest neighbour based methods differ from other approaches such that they predict the secondary structure of a target protein using local sequence similarity to segments of known proteins (usually through a sliding window) even when the overall target protein sequence differ substantially from the reference proteins. This approach benefits from availability of numerous similarity matches or from several highly identical matches of known structures. A method by Salamov and Solovyev (1995) which was implemented through the NNSSP server is the most successful of this approach (Rost, 2001). It is an improvement of the work initiated by Yi and Lander (1993) which employs a neural network and the nearest neighbour algorithm. The enhancement of the original work was done by incorporating the N and C-terminal position of helices, sheets and loops as distinctive classes of the secondary structure. An inherent drawback of nearest neighbour algorithm, the computational time is reduced by limiting the reference database with a smaller portion of proteins that are similar with the target sequence. The server, which includes evolutionary information through multiple sequence alignments, has 72.2% correct predictions.

The PREDATOR server (Frishman and Argos, 1997), an alternative nearest neighbour based predictor, adopts local pairwise alignment of the target sequence as opposed to the multiple sequence alignment. The carefully selected alignment is derived from known structures. According to the authors, the technique achieves 75% accuracy using jack-knife validations.

1.3.4 Hidden Markov Model Methods

Karplus and colleagues (Karplus et al., 1998) have developed a secondary structure prediction algorithm based on the hidden Markov model (HMM) using the same concept as the nearest neighbour technique. Starting from a target sequence, their technique, SAM-T98 iteratively builds and refines a HMM from a set of potential homologues found in a non-redundant protein database. The resulting model is then used to search the Protein Data Bank (Berman et al., 2002) for similar

proteins. The multiple sequence alignments arising from the similar proteins are used to predict the secondary structure of the target sequence. Based on the results from the EVA server, a newer version of the tool, SAM-T99 has an accuracy of 74.9%.

A second HMM method by Bystroff and colleagues (Bystroff et al., 2000) is implemented through the HMMSTR server. Initially, a database of motifs common to all protein families and containing 3 to 19 residues is built from known structures. Each Markov state contains information about the sequence and structure of a single position in the derived motifs. The adjacent positions are represented by transitions from one state to the next. The HMM is generated by hierarchically merging the linear chains of the states based on sequence and structural similarity. Each state in the resulting HMM produces the predicted secondary structure according to the probability distribution specific to that state. The accuracy of this method was reported to be 74.3%.

1.3.5 Kernel Methods

Support Vector Machine (SVM) is one of the powerful kernel based learning machines which has favourable properties such as structural risk minimization and effective avoidance of overfitting (Müller et al., 2001). Hua and Sun (2001) was the first to introduce SVMs for the secondary structure prediction problem. Since SVMs can effectively predict two classes at a time, they have used six binary classifiers, H/~H, E/~E, C/~C, H/E, E/C and C/H, and combined them to produce a tertiary classifier for the three secondary structure classes. Although Hua and Sun did not use evolutionary information for their prediction scheme, they have reported to achieve 73.5% accuracy.

Despite of Hua and Sun's very impressive results, Casbon (2002) was not able to reproduce their work in his Masters research. He was only able to correctly predict for 71.5% of the residues using the same data sets and setups even after using multiple sequence alignments. He found their results to be unusual because as reported by Rost (2002), no other methods are able reach above 70% accuracy when

evolutionary information was not considered. He was also skeptical of their results because they were unwilling to further discuss their prediction scheme although they have not published their procedure for optimizing the SVM parameters. He conjectures that they may have overstated the accuracy by optimizing the parameters with respect to the test set. Additionally, it was difficult for him to verify their results because they have not made their system available for public like other prediction servers.

1.3.6 Summary of Prediction Methods

Table 1.1 summarizes the performance of the prediction techniques. It is worth noting that the accuracy claimed by authors may sometimes be overestimated. In order to address such cases, an evaluation server such as EVA can be used to objectively compare and verify the performance of the methods. As can be observed

Table 1.1: Summary of performance of various prediction methods.

Type	Method	Year	Generation*	Q_3 % claimed	Q_3 % verified
Statistical	Chou and Fasman	1974	First	77.0	49.9
	Garnier et al. (GOR I)	1978	First	57.0	55.9
	Kloczkowski et al. (GOR V)	2002	Third	74.2	
Neural Networks	Qian and Sejnowski	1988	Second	64.3	
	Rost and Sander (PHD)	1994	Third	71.6	70.5
	Pollastri et al. (SSpro)	2002	Third	76.0	74.5
Nearest Neighbours	Salamov and Solovyev (NNSSP)	1995	Third	72.2	
	Frishman and Argos (PREDATOR)	1997	Third	75.0	
Hidden Markov Models	Karplus et al. (SAM-T99)	1998	Third	–	74.9
	Bystroff et al. (HMMSTR)	2000	Third	74.3	
Support Vector Machines	Hua and Sun	2001	Second	73.5	
	Casbon	2002	Third	71.5	

Note:
*Generation (Rost and Sander, 2000):
First : using only residue statistics
Second : using sliding windows and large database
Third : introduction of long range influences through evolutionary information

in Table 1.1, the verified performances are always lower than the performances reported by the authors. An alternative way to objectively compare the efficiency of different methods is to use the same data set that was used for training and testing.

1.4 Challenges of Protein Secondary Structure Prediction

Machine learning methods, have been used for the last 15 years to predict secondary structure of proteins and have consistently demonstrated to be the best approach. Through the years, as a result of increased training data and innovative techniques, the performance has steadily improved by about one percentage point per year (Baldi and Pollastri, 2002). However, according to Rost (2001), since the secondary structure assignments of the training data resulting from the Database of Secondary Structure in Proteins (DSSP) (Kabsch and Sander, 1983) differ by about 12 percentage points between structural homologues, the maximum accuracy one could expect from secondary structure predictions is 88 percentage points. The variation is attributed to the threshold used by the DSSP to convert the experimentally derived three-dimensional structures into secondary structures and the dynamics of the three-dimensional structure itself, since some regions of protein are more mutable than others. Additionally, a protein structure formation is not only affected by local residue interactions but also by influence from distant residues of the amino acid sequence during folding (Rost, 2001). Consequently, identical sequences of up to five amino acids have been found to exist for two distinct structure elements. Hence, a challenge is to formulate an approach to either include these long range influences in the prediction scheme or to discriminate cleverly between two classes when they are both trained with identical input sequences. A further problem is the effect of the folding environment (Krittanai and Johnson, 2000) that is neglected altogether in secondary structure prediction. From the computational perspective, the secondary structure prediction problem also presents a substantial challenge since it deals with very large data.

1.5 Evidence-Theoretic Neural Network Classifier

A novel classifier based on Dempster-Shafer (D-S) theory of evidence (Shafer, 1976) has been introduced by Denoeux (2000) recently. The neural network-like architecture, consisting of an input layer, two hidden layers and an output layer, summarizes the training samples into several similar groups of prototypes with each having its own degree of class memberships. When an input sample is considered for classification, its similarity to each prototype and consequently its degrees of class memberships are used as items of evidence. Adopting the ideas of evidence theory, the items of evidence acquired from the prototypes are represented by basic belief assignments (BBAs) which are accumulated using the Dempster's rule of combination to derive the aggregated class membership of the input sample. The weight vector, the receptive field, the relative importance and the class membership of each prototype, which are the adjustable parameters of the classifier, are computed by minimizing its error function.

Based on Denoeux's experiments on simulated and typical real world data, the classifier demonstrated an excellent performance when compared to other statistical and machine learning techniques. The favourable performance of the classifier is largely attributed to the D-S theory which offers a simple procedure for collecting the evidences from all prototypes in order to compute the uncertainty attached to either simple or compound class memberships of the input sample under consideration. This approach provides a way of moderating the importance of prototypes in the decision, depending on the closeness to the sample to be classified. It also permits combination of the outputs of complementary evidences from different prototypes, which could have been trained at different levels of abstraction (Denoeux, 1995). For example, given three classes H, E and C, one prototype may discriminate between class H and the other two, while another one may help to ascertain E and C. By combining the BBAs produced by each of these prototypes, Dempster's rule provides a method to assess the reliability of the resulting classification. In this work, the classifier will be referred to as Denoeux belief neural network (DBNN).

1.6 Motivations of Research

As described in Section 1.3, secondary structure prediction has benefited from improvements in machine learning techniques (Rost, 2002). Additionally, neural network based methods appeared to perform relatively well in the prediction problem. Therefore, following the favourable results of DBNN in Denoeux's experimental studies, the prediction of secondary structure may potentially benefit from the classifier's performance as well. Furthermore, certain features of DBNN, such as its ability to combine complementary evidences and to compute simple or compound class memberships of the input sample, support the specific challenges imposed by the prediction problem. More importantly, despite of DBNN's excellent performance in the experiments conducted by Denoeux, up to now there has been little further work reported pertaining to its applications in other classification problems or about any improvements to the original classification scheme. Taking these factors into consideration, the secondary structure prediction, which poses several challenges, would present a good platform to further investigate and possibly improve the DBNN classifier.

1.7 Objectives of Research

The goal of this research is to investigate the applicability of the recently introduced DBNN classifier in a challenging real world task such as the protein secondary structure prediction while looking for areas of potential improvements. In order to realize the goal, several objectives need to be achieved:

- To develop a generic tool for parallelizing computationally intensive algorithms and techniques investigated in this research such that the computational load arising from the immensity of the protein data can be addressed.
- To determine the features and drawbacks of the DBNN classifier such that they can be taken into account when the classifier is used to predict the protein structure.

- To design and develop an improved neural network classifier based on the D-S theory to address the challenges posed by the protein secondary structure prediction.
- To predict the protein secondary structure from amino acid sequences using the improved classifier and to see how it compares with existing methods.

1.8 Scope of Research

Since the goal of this research is to evaluate the applicability of the recently introduced DBNN classifier in the secondary structure prediction, the prediction can be confined to sequence-to-structure prediction without the inclusion of evolutionary information. This is because the result of the sequence-to-structure prediction without the incorporation of evolutionary information would be sufficient to gauge the performance of the DBNN classifier when it is compared to the results of other existing methods at the same level.

1.9 Overview of the Thesis

Figure 1.6 illustrates the four major steps that are adopted in this thesis to achieve the objectives of the research. After developing a parallelizing interface, the interface is used to parallelize the DBNN classifier and accompanying techniques which would become computationally intensive as a result of the large protein data. The parallelized DBNN classifier is used to conduct rigorous experiments on several benchmark data to determine its attributes and possible limitations. Following that, the possible causes of its limitations can be analyzed and several potential solutions can be formulated based on the analysis. The proposed solutions can be rigorously evaluated on the same benchmark data of the original DBNN to determine their effectiveness. The parallelization interface would support such computationally demanding evaluations. The results of the evaluation can be used to determine the

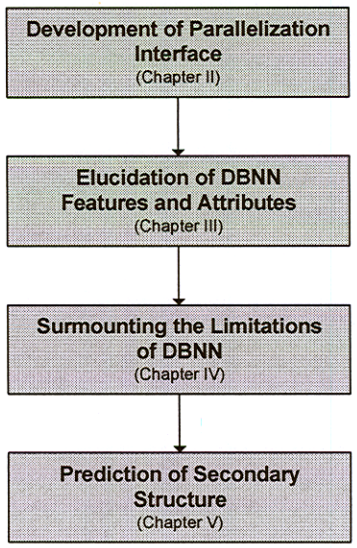

Figure 1.6: Major steps adopted to achieve the research objectives.

best solution among the proposed techniques to address the challenges of protein secondary structure prediction. The best of the proposed solutions along with the original DBNN classifier can be employed to predict the secondary structure using the same data set adopted by the existing prediction methods. This way, it would be possible to objectively compare the performances of the predictors and to see whether the proposed solution addresses the limitations of the original DBNN classifier in the prediction. From the results, the extent of which both of the classifiers, which are based on the D-S theory of evidence, are applicable in secondary structure prediction can be determined.

The chapters in this thesis follow the steps illustrated in Figure 1.6. A general description of the contents of the chapters is given as follows:

- Chapter II presents the design, implementation and evaluation of an interface developed to parallelize computationally intensive algorithms and methods investigated in this research.
- Chapter III provides a description of DBNN classifier and its classification strategy. The chapter also empirically examines the attributes and also the possible limitations of DBNN.

- Chapter IV attempts to analyze the possible limitations of DBNN, proposes several potential solutions and then evaluates them.
- Chapter V presents the secondary structure prediction method employed in this research and evaluates both DBNN and the best of the proposed solutions in the secondary structure problem.
- Chapter VI concludes the thesis and provides suggestions for future research.

RELATED PUBLICATIONS

Title	Author	Journal/Proceedings	Related Chapter
Prediction of Protein Secondary Structure	Arjunan, S. N. V., Deris, S. and Illias, R. M.	Jurnal Teknologi Volume: 35(C) Pages: 81-90 Year: 2001	I
A Parallelization Interface for Large Biological Data Problems	Arjunan, S. N. V., Deris, S. and Illias, R. M.	Parallel Processing Letters (submitted)	II
Assignment of Protein Sequence to Functional Family Using Neural Network and Dempster-Shafer Theory	Zaki, N. M., Deris, S., Arjunan, S. N. V. and Illias, R. M.	Journal of Theoretics Volume: 5 (1) Year: 2003	III
A Neural Network Classifier based on the Dempster-Shafer Theory for Protein Secondary Structure Prediction	Arjunan, S. N. V., Deris, S. and Illias, R. M.	Pattern Recognition Letters (submitted)	V
Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network	Arjunan, S. N. V., Deris, S. and Illias, R. M.	Proceedings of the International Conference on Artificial Intelligence in Engineering and Technology 2002. Sabah, Malaysia Pages: 554-564 Year: 2002	V
UTMPred: Ab Initio Eight-state Protein Secondary Structure Predictor	Arjunan, S. N. V., Deris, S. and Illias, R. M.	12th Malaysian Society for Molecular Biology and Biotechnology Scientific Meeting Year: 2002	V

REFERENCES

Aberdeen, D. and Baxter, J. (2001). Emmerald: a fast matrix-matrix multiply using Intel's SSE instructions. *Concurrency and Computation: Practice and Experience*. 13(2). 103 – 119.

Agarwal, A., Kranz, D. A. and Natarajan, V. (1995). Automatic Partitioning of Parallel Loops and Data Arrays for Distributed Shared-Memory Multiprocessors. *IEEE Transactions on Parallel and Distributed Systems*. 6(9). 943 – 962.

Albrecht, M., Hanisch, D., Zimmer, R. and Lengauer, R. (2002). Improving fold recognition of protein threading by experimental distance constraints. *In Silico Biology*. 2. 0030.

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. and Schaffer, A. A. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*. 25(17). 3389 – 3402.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*. 181. 223 – 238.

Arjunan, S. N. V., Deris, S. and Illias, R. M. (2002). Protein Secondary Structure Prediction based on Denoeux Belief Neural Network. *Proceedings of the International Conference on Artificial Intelligence in Engineering and Technology 2002*. Sabah, Malaysia: Universiti Malaysia Sabah Publications. 554 – 560.

Baldi, P. and Pollastri, G. (2002). A Machine Learning Strategy for Protein Analysis. *IEEE Intelligent Systems*. 17(2). 28 – 35.

Baldi, P., Brunak, S., Frasconi, P., Pollastri, G. and Soda, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*.

15(11). 937 – 946.

Bellman, R. E. (1961). *Adaptive Control Processes*. New Jersey, USA: Princeton University Press.

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., Ravichandran, V., Schneider, B., Thanki, N., Padilla, D., Weissig, H., Westbrook, J. D. and Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallographica Section D*. 58(6). 899 – 907.

Blackford, L. S., Choi, J., Cleary, A., D'Azevedo, E., Demmel, J., Dhillon, I., Dongarra, J. J., Hammarling, S., Henry, G., Petitet, A., Stanley, K., Walker, D. and Whaley, R. C. (1997). *ScaLAPACK Users' Guide*. Philadelphia, USA: SIAM Publications.

Blake, C.L. and Merz, C.J. (1998). *UCI Repository of machine learning databases*. Irvine, CA: University of California, Department of Information and Computer Science. UCI Repository is available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Blout, E. R., de Lozé, C., Bloom, S. M. and Fasman, G. D. (1960). The dependence of the conformations of synthetic polypeptides on amino acid composition. *Journal of the American Chemical Society*. 79(14). 3787 – 3789.

Bohte, S. M., La Poutre, H. and Kok, J. N. (2002). Unsupervised clustering with spiking neurons by sparse temporal coding and multilayer RBF networks. *IEEE Transactions on Neural Networks*. 13(2). 426 – 435.

Boniface, Y., Alexandre, F. and Vialle, S. (1999). A Library to Implement Neural Networks on MIMD Machines. *Proceedings of the 5th International Euro-Par Conference on Parallel Processing*. Toulouse, France. 935 – 938.

Bonneau, R. and Baker, D. (2001). Ab Initio Protein Structure Prediction: Progress and Prospects. *Annual Review of Biophysics and Biomolecular Structure*. 30. 173 – 189.

Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, M. and Baker, D. (2001). Rosetta in CASP4: Progress in Ab Initio Protein Structure Prediction. *Proteins: Structure, Function, and Genetics*. 45(S5). 119 – 126.

Branden, C. and Tooze, J. (1999). *Introduction to Protein Structure*. 2nd Edition. New York, USA: Garland Publishing Inc.

Bystroff, C., Thorsson, V. and Baker, D. (2000). HMMSTR: A hidden Markov model for sequence-structure correlations in proteins. *Journal of Molecular Biology*. 301(1). 173 – 190.

Carpenter, B., Zhang, B. and Wen, Y. (1997). *NPAC PCRC Runtime Kernel Definition*. Technical Report CRPC-TR97726. Center for Research on Parallel Computation, Rice University, USA.

Casbon, J. (2002). *Protein Secondary Structure Prediction with Support Vector Machines*. The University of Sussex: Masters Thesis.

Chatterjee, S., Lebeck, A. R., Patnala, P. K. and Thottethodi, M. (1999). Recursive Array Layouts and Fast Parallel Matrix Multiplication. *Proceedings of the 11th ACM Symposium on Parallel Algorithms and Architectures*. Saint-Malo, France. 222 – 231.

Chen, C. C., Singh, J. P. and Altman, R. B. (1999). Using imperfect secondary structure predictions to improve molecular structure computations. *Bioinformatics*. 15(1). 53 – 65.

Chen, J. and Taylor, V. E. (2002). Mesh Partitioning for Efficient Use of Distributed Systems. *IEEE Transactions on Parallel and Distributed Systems*. 13(1). 67 – 79.

Chou, P.Y. and Fasman, G.D. (1974). Prediction of protein conformation. *Biochemistry*. 13(2). 222 – 245.

Chou, P.Y. and Fasman, G.D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology and Related Areas of Molecular Biology*. 47. 45 – 48.

Comino, C. and Narasimhan, V. L. (2002). A Novel Data Distribution Technique for Host-Client Type Parallel Applications. *IEEE Transactions on Parallel and Distributed Systems*. 13(2). 97 – 110.

Cuisenaire, O., Duay, V., Solanas, E. and Thiran, J.-P. (2001). Relative anatomical location for statistical non-parametric brain tissue classification in MR images. *Proceedings of the 2001 IEEE International Conference on Image Processing*. 2. 885 – 888.

Day, R. O., Zydallis, J. B., Lamont, G. B. and Pachter, R. (2002). Solving the Protein Structure Prediction Problem through a Multiobjective Genetic Algorithm. *Technical Proceedings of the 2002 International Conference on Computational*

Nanoscience and Nanotechnology. Cambridge, USA: Computational Publications. 32 – 35.

Delichère, M and Memmi, D. (2002). Neural dimensionality reduction for document processing. *Proceedings of the 10th European Symposium on Artificial Neural Networks*. Bruges, Belgium. 211 – 216.

Denoeux, T. (1995). A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics A*. 25(5). 804 – 813.

Denoeux, T. (2000). A neural network classifier based on Dempster-Shafer theory. *IEEE transactions on Systems, Man and Cybernetics A*. 30(2). 131 – 150.

Dettling, M. and Bühlmann, P. (2002). Supervised clustering of genes. *Genome Biology*. 3(12). research0069.1 – 0069.15.

Dhillon, I. S. and Modha, D. S. (1999). A Data-Clustering Algorithm on Distributed Memory Multiprocessors. Large-Scale Parallel Data Mining. *Lecture Notes in Computer Science*. 1759. 245 – 260.

Dodge, C., Schneider, R. and Sander, C. (1998). The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Research*. 26(1). 313 – 315.

Domeniconi, C. and Gunopulos, D. (2002). Adaptive Nearest Neighbor Classification Using Support Vector Machines. *Advances in Neural Information Processing Systems*. 14. Cambridge, USA: MIT Press.

Dongarra, J. J. (2002). *Performance of Various Computers Using Standard Linear Equations Software*. Technical Report CS-89-85. University of Tennessee, USA.

Dongarra, J. J., Croz, J. D., Hammarling, S. and Duff, I. S. (1990). A Set of Level 3 Basic Linear Algebra Subprograms. *ACM Transactions on Mathematical Software*. 16(1). 1 – 17.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001). *Pattern Classification*. New York, USA: John Wiley and Sons.

Estivill-Castro, V. and Houle, M. E. (2001). Robust Distance-Based Clustering with Applications to Spatial Data Mining. *Algorithmica*. 30(2). 216 – 242.

Eyrich, V. A., Standley, D. M. and Friesner, R. A. (1999). Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *Journal of Molecular Biology*. 288(4). 725 – 742.

Frishman, D. and Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Genetics*. 23(4). 566 – 579.

Frishman, D. and Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. *Proteins: Structure, Function, and Genetics*. 27(3). 329 – 335.

Garey, M. R., Johnson, D. S. and Witsenhausen, H. S. (1982). The Complexity of the Generalized Lloyd-Max Problem. *IEEE Transactions on Information Theory*. 28(2). 255 – 256.

Garnier, J., Osguthorpe, D. J. and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*. 120. 97 – 120.

Guex, N. and Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. *Electrophoresis*. 18(15). 2714 – 2723.

Gunnels, J. A., Henry, G. M. and van de Geijn, R. A. (2001). A Family of High-Performance Matrix Algorithms. Part 1, Computational Science – 2001. *Lecture Notes in Computer Science*. 2073. 51 – 60.

Hamerly, G. and Elkan, C. (2002). Alternatives to the k-means algorithm that find better clusterings. *Proceedings of the 11th ACM International Conference on Information and Knowledge Management*. McLean, USA. 600 – 607.

Hendrickson, B. and Leland, R. (1994). *The Chaco User's Guide: Version 2.0*. Technical Report SAND94-2692. Sandia National Laboratory, USA.

High Performance Fortran Forum. (1997). *High Performance Fortran language specification, Version 2.0*. Center for Research on Parallel Computation, Rice University, USA.

Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992). Selection of representative data sets. *Protein Science*. 1(3). 409 – 417.

Holden, M., Hill, D. L. G., Denton, E. R. E., Jarosz, J. M., Cox, T. C. S., Rohlfing, T., Goodey, J. and Hawkes, D. J. (2000). Voxel Similarity Measures for 3D Serial MR Brain Image Registration. *IEEE Transactions on Medical Imaging*. 19(2). 94 – 102.

Hua, S. and Sun, Z. (2001). A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine

Approach. *Journal of Molecular Biology*. 308(2). 397 – 407.

Jolliffe, I. T. (1986). *Principal Component Analysis*. New York, USA: Springer Verlag.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22(12). 2577 – 2637.

Kantabutra, S. and Couch, A. L. (2000). Parallel K-means Clustering Algorithm on NOWs. *NECTEC Technical Journal*. 1(6).

Karplus, K., Barrett, C. and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*. 14(10). 846 – 856.

Karypis, G and Han, E. H. (2000). Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. *Proceedings of the 9th ACM International Conference on Information and Knowledge Management*. Washington, USA. 12 – 19.

Karypis, G. and Kumar, V. (1998). A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM Journal on Scientific Computing*. 20(1). 359 – 392.

Klepeis, J. L. and Floudas, C.A. (2003). Ab initio Tertiary Structure Prediction of Proteins. *Journal of Global Optimization*. 25(1). 113 – 140.

Kloczkowski, A., Ting, K.-L., Jernigan, R. L. and Garnier, J. (2002). Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins: Structure, Function, and Genetics*. 49(2). 154 – 166.

Krittanai, C. and Johnson, W. C. (2000). The relative order of helical propensity of amino acids changes with solvent environment. *Proteins: Structure, Function, and Genetics*. 39(2). 132 – 141.

Kumar, V., Grama, A., Gupta, A. and Karypis, G. (1994). *Introduction to Parallel Computing*. Redwood City, USA: Benjamin/Cummings.

Lathrop, R. H. (1994). The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Engineering*. 7(9). 1059 – 1068.

LeCun, Y., Bottou, L., Orr, G. B. and Müller, K.-R. (1998). *Efficient BackProp. Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer-Verlag. 9 – 50.

Levinthal, C. (1969). How to fold graciously. In DeBrunner, P., Tsibris, J. and

Munck, E. *Mossbauer spectroscopy in biological systems*. Urbana, USA: University of Illinois Press. 22 – 24.

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D. and Karplus, M. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *Journal of Physical Chemistry B*. 102(18). 3586 – 3616.

Maclin, R. and Shavlik, J. W. (1993). Using knowledge-based neural networks to improve algorithms: Refining the Chou-Fasman algorithm for protein folding. *Machine Learning*. 11(2 – 3). 195 – 216.

Magoulas, G, D., Vrahatis, M. N. and Androulakis, G. S. (1999). Improving the Convergence of the Backpropagation Algorithm Using Learning Rate Adaptation Methods. *Neural Computation*. 11(7). 1769 – 1796.

Marti-Renom, M. A., Stuart, A., Fiser, A., Sánchez, R., Melo, F. and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*. 29. 291 – 325.

Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (1994). Machine Learning, Neural and Statistical Classification. New York, USA: Ellis Horwood. STATLOG data is available at ftp.ncc.up.pt/pub/statlog/.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Smola, A. J. and Müller, K.-R. (2000). Invariant feature extraction and classification in kernel spaces. *Advances in Neural Information Processing Systems*. 12. Massachusetts, USA: MIT Press. 526 – 532.

Møller, M. (1993). Supervised learning on large redundant training sets. *International Journal of Neural Systems*. 4(1). 15 – 25.

MPI Forum. (1998). Special Issue: MPI2: A Message-Passing Interface Standard. *The International Journal of High Performance Computing Applications*. 12(1 – 2). 1 – 299.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K. and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*. 12(2). 181 – 201.

Nagl, S. B., Freeman, J. and Smith, T. F. (1999). Evolutionary Constraint Networks in Ligand-Binding Domains: An Information-Theoretic Approach. *Pacific Symposium on Biocomputing*. 4. 90 – 101.

Nebauer, C. (1998) Evaluation of convolutional neural networks for visual recognition. *IEEE Transactions on Neural Networks*. 9(4). 685 – 696.

Ng, M. K. (2000). K-Means-Type Algorithms on Distributed Memory Computer. *International Journal of High Speed Computing*. 11(2). 75 – 91.

Pauling, L. and Corey, R. B. (1951). Configurations of Polypeptide Chains with Favoured Orientations Around Single Bonds: Two New Pleated Sheets. *Proceedings of the National Academy of Sciences USA*. 37. 729 – 740.

Pauling, L., Corey, R. B., and Branson, H. R. (1951). The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *Proceedings of the National Academy of Sciences USA*. 37. 235 – 240.

Peña, J. M., Lozano, J. A. and Larrañaga, P. (1999). An Empirical Comparison of Four Initialization Methods for the K-Means Algorithm. *Pattern Recognition Letters*. 20(10). 1027 – 1040.

Pollastri, G., Przybylski, D., Rost, B. and Baldi, P. (2002). Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. *Proteins: Structure, Function, and Genetics*. 47(2). 228 – 235.

Prechelt, L. and Hänßgen, S. U. (2002). Efficient Parallel Execution of Irregular Recursive Programs. *IEEE Transactions on Parallel and Distributed Systems*. 13(2). 167 – 178.

Przytycka, T., Aurora, R. and Rose, G. D. (1999). A protein taxonomy based on secondary structure. *Nature Structural Biology*. 6(7). 672 – 682.

Qian, N. and Sejnowski, T. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*. 202(4). 865 – 884.

Rasmussen, C. E., Neal, R. M., Hinton, G. E., van Camp, D., Revow, M., Ghahramani, Z., Kustra, R. and Tibshirani, R. (1996). *The DELVE manual, Version 1.1*. DELVE data is available at http://www.cs.utoronto.ca/~delve/data/datasets.html.

Rätsch, G., Onoda, T. and Müller, K.-R. (1998). *Soft margins for AdaBoost.*

NeuroCOLT Technical Report NC-TR-1998-021, Department of Computer Science, University of London, UK. IDA Benchmark repository is available at http://ida.first.gmd.de/~raetsch/data/benchmarks.htm.

Rätsch, G., Onoda, T. and Müller, K.-R. (2001). Soft Margins for AdaBoost. *Machine Learning*. 42(3). 287 – 320.

Richards, F. M. and Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins: Structure, Function, and Genetics*. 3. 71 – 84.

Richardson, D. C. and Richardson, J. S. (1992). The kinemage: A tool for scientific communication. *Protein Science*. 1(1). 3 – 9.

Riis, S. K. and Krogh, A. (1996). Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *Journal of Computational Biology*. 3. 163 – 183.

Rogers, R.O. and Skillicorn, D.B. (1998). Using the BSP Cost Model to Optimize Parallel Neural Network Training. *Future Generation Computer Systems*. 14(5 – 6). 409 – 424.

Rost, B. (2001). Review: Protein secondary structure prediction continues to rise. *Journal of Structural Biology*. 134(2 – 3). 204 – 218.

Rost, B. (2002). Rising accuracy of protein secondary structure prediction. In Chasman, D. *Protein structure determination, analysis and modeling for drug discovery*. New York, USA: Dekker. 207 – 249.

Rost, B. and Eyrich, V. A. (2001). EVA: Large-scale analysis of secondary structure prediction. *Proteins: Structure, Function, and Genetics*. 45(S5). 192 – 199.

Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*. 232(2). 584 – 599. RS126 data set is available at http://antheprot-pbil.ibcp.fr/Rost.html.

Rost, B. and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Structure, Function, and Genetics*. 19(1). 55 – 72.

Rost, B. and Sander, C. (2000). Third generation prediction of secondary structure. In Webster, D. *Protein structure prediction: methods and protocols*. Totowa, USA: Humana Press. 71 – 95.

Rost, B., Sander, C. and Schneider, R. (1994). Redefining the goals of protein

secondary structure prediction. *Journal of Molecular Biology*. 235(1). 13 – 26.

Salamov, A. A. and Solovyev V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of Molecular Biology*. 247(1). 11 – 15.

Schikuta, E. and Weidmann, C. (1997). Data Parallel Simulation of Self-organizing Maps on Hypercube Architectures. *Proceedings of the Workshop on Self-Organizing Maps 1997*. Helsinki, Finland. 142 – 147.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. New Jersey, USA: Princeton University Press.

Silva, F. M., and Almeida, L. B. (1990). Speeding up backpropagation. *Advanced Neural Computers*. Amsterdam, Netherlands: North-Holland. 151 – 158.

Skolnick, J. and Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science*. 250. 1121 – 1125.

Smets, P. (1990). The combination of evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 12(5). 447 – 458.

Stawiski, E. W., Baucom, A. E., Lohr, S. C. and Gregoret, L. M. (2000). Predicting protein function from structure: Unique structural features of proteases. *Proceedings of the National Academy of Sciences USA*. 97(8). 3954 – 3958.

Sundararajan, N. and Saratchandran, P. (1998). *Parallel Architectures for Artificial Neural Networks*. Los Alamitos, USA: IEEE Computer Society Press.

Valsalam, V. and Skjellum, A. (2002). A framework for high-performance matrix multiplication based on hierarchical abstractions, algorithms and optimized low-level kernels. *Concurrency and Computation: Practice and Experience*. 14(10). 805 – 839.

van de Geijn, R. A. (1997). *Using PLAPACK: Parallel Linear Algebra Package*. Massachusetts, USA: MIT Press.

Whaley, R. C., Petitet, A. and Dongarra, J. J. (2001). Automated Empirical Optimization of Software and the ATLAS Project. *Parallel Computing*. 27(1 – 2). 3 – 25.

Wu, C. H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Ledley, R. S., Lewis, K. C., Mewes, H-W., Orcutt, B. C., Suzek, B. E., Tsugita, A., Vinayaka, C. R., Yeh, L-S., 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.

Xia, Y., Huang, E. S., Levitt, M. and Samudrala, R. (2000). Ab initio construction of protein tertiary structures using a hierarchical approach. *Journal of Molecular Biology*. 300(1). 171 – 185.

Xu, Y. and Xu, D. (2000). Protein threading using PROSPECT: Design and evaluation. *Proteins: Structure, Function, and Genetics*. 40(3). 343 – 354.

Yi, T. M. and Lander E. S. (1993). Protein secondary structure prediction using nearest-neighbor methods. *Journal of Molecular Biology*. 232(4). 1117 – 1129.

Young, M. M., Kirshenbaum, K., Dill, K. A. and Highsmith, S. (1999). Predicting Conformational Switches in Proteins. *Protein Science*. 8(9). 1752 – 1764.

Zhang, B. (2001). Generalized K-Harmonic Means – Boosting in Unsupervised Learning. *Proceedings of the 1st SIAM International Conference on Data Mining*. Chicago, USA.

Zhang, B., Hsu, M. and Forman, G. (2000). Accurate Recasting of Parameter Estimation Algorithms Using Sufficient Statistics for Efficient Parallel Speed-Up: Demonstrated for Center-Based Data Clustering Algorithms. *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*. Lyon, France. 243 – 254.