11-27-2007

Prediction of Oxidation States of Cysteines and Disulphide Connectivity

Aiguo Du

PREDICTION OF OXIDATION STATES OF CYSTEINES AND DISULFIDE BRIDGES IN PROTEINS

by

AIGUO DU

Under the Direction of Yi Pan

ABSTRACT

Knowledge on cysteine oxidation state and disulfide bond connectivity is of great importance to protein chemistry and 3-D structures. This research is aimed at finding the most relevant features in prediction of cysteines oxidation states and the disulfide bonds connectivity of proteins. Models predicting the oxidation states of cysteines are developed with machine learning techniques such as Support Vector Machines (SVMs) and Associative Neural Networks (ASNNs). A record high prediction accuracy of oxidation state, 95%, is achieved by incorporating the oxidation states of N-terminus cysteines, flanking sequences of cysteines and global information on the protein chain (number of cysteines, length of the chain and amino acids composition of the chain etc.) into the SVM encoding. This is 5% higher than the current methods. This indicates to us that the oxidation states of amino terminal cysteines infer the oxidation states of other cysteines in the same protein chain. Satisfactory prediction results are also obtained with the newer and more inclusive SPX dataset, especially for chains with higher number of cysteines. Compared to literature methods, our approach is a one-step prediction system, which is easier to implement and use. A side by side comparison of SVM and ASNN is conducted. Results indicated that SVM outperform ASNN on this particular problem.

For the prediction of correct pairings of cysteines to form disulfide bonds, we first study disulfide connectivity by calculating the local interaction potentials between the flanking

sequences of the cysteine pairs. The obtained interaction potential is further adjusted by the coefficients related to the binding motif of enzymes during disulfide formation and also by the linear distance between the cysteine pairs. Finally, maximized weight matching algorithm is applied and performance of the interaction potentials evaluated. Overall prediction accuracy is unsatisfactory compared with the literature.

SVM is used to predict the disulfide connectivity with the assumption that oxidation states of cysteines on the protein are known. Information on binding region during disulfide formation, distance between cysteine pairs, global information of the protein chain and the flanking sequences around the cysteine pairs are included in the SVM encoding. Prediction results illustrate the advantage of using possible anchor region information.

INDEX WORDS: cysteines, oxidation states, connectivity, disulphide bonds, support vector machine (SVM), Associative neural network (ASNN)

PREDICTION OF OXIDATION STATES OF CYSTEINES AND DISULFIDE BRIDGES IN PROTEINS

by

AIGUO DU

A Dissertation Submitted in Partial Fulfillment of Requirements for the Degree of

Doctor of Philosophy

In the College of Arts and Sciences

Georgia Stage University

2007

PREDICTION OF OXIDATION STATES OF CYSTEINES AND DISULRIDE BRIDGES IN PROTEINS

by

AIGUO DU

Major Professor: Yi Pan
Committee: Anu Bourgeois
Alex Zelikovsky
Gengsheng Qin

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2007

ACKOWLEDGEMENTS

The dissertation would not have been possible without the help of so many people. I would like to take this opportunity to express my deep appreciation to all those who help and encourage me in this hard but extremely rewarding process.

First and foremost, I would like to thank my thesis advisor, Professor Yi Pan, for all his help, advice, inspiration, support, guidance, and patience. During my graduate study at GSU, I have times of feeling confused and frustrated and even do not know what to do, but I always believe that Dr. Pan is there to consult, to guide, and to seek strength and direction from. He has always been helpful whenever I needed an intelligent discussion. During all these years, he is not only a patient advisor in academics, but also a true mentor for career development and attitude toward life. I believe it is the best luck of me to have the chance to work with him on both my master and Ph.D degrees.

I am also very grateful to Dr. Anu Bougeois, Dr. Alex Zelikovsky, and Dr. Gengsheng Qin for serving on my thesis committee, and for their valuable time and co-operation in reviewing this work. I would like to thank Dr. Qin specially for providing valuable opinions on the statistics related to my research.

In the past years, I have been really fortunate to study at GSU where faculty members are friendly and classmates/labmates are smart and fun. I thank Dr. Raj Sunderraman for his great patience and support during my study. Special thanks go to Dr. Hai Deng, Dr. Wei Zhong for valuable inputs on my research topic. I would also like to thank Hai-Jin Hu wholeheartedly for helping me out in times of needs.

Also I am very lucky to have supporting families, who have always been standing behind me with great patience, love and inspiration. I am so grateful to have my parents' constant support and motivation. Special appreciation goes to my husband for the encouragement during my study. I would also like to thank my baby for his tender love and care which gave me tremendous motivation for life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

4-Fluoro-7-sulfoamoylbenzofurazan	ABD-F
Accuracy	Q2
Amino Acid	AA
Analine (one of the amino acids)	A
Arginine (one of the amino acids)	R
Artificial Neural Networks	ANN
Asparagine (one of the amino acids)	N
Aspartic acid	D
Associative Neural networks	ASNN
Basic Local Alignment and Search Tool	BLAST
Cysteine (one of the amino acids)	C
Cysteine in oxidized form	SS
Cysteine in reduced form	SH
Endoplasmic Reticulum	ER
False Negative	FN
False Positive	FP
Glutamic acid (one of the amino acids)	E
Glutamine (one of the amino acids)	Q
Glycine (one of the amino acids)	G
High-performance liquid chromatography	HPLC
Histidine (one of the amino acids)	H

Homology-derived secondary structure of proteins	HSSP
Isoleucine (one of the amino acids)	I
k-nearest-neighbors	KNN
Leucine (one of the amino acids)	L
Lysine (one of the amino acids)	K
Matthews correlation coefficient	MCC
Messenger Ribonucleos Neuclotide Acid	mRNA
Methionine (one of the amino acids)	M
multilayer perceptron	MLP
Pheneylanaline (one of the amino acids)	F
Point specific sequence profile	PSSP
Proline (one of the amino acids)	P
Protein Data Bank	PDB
Protein Disulfide Isomerase	PDI
Protein Disulfide Isomerase from pancreas	PDIp
Serine (one of the amino acids)	S
Support Vector Machine	SVM
Threonine (one of the amino acids)	T
True Negative	TN
True Positive	TP
Tryptophen (one of the amino acids)	W
Tyrosine (one of the amino acids)	Y
University of California, Irvine	UCI

Valine (one of the amino acids)	V
Virtual Computational Chemistry Lab	VCCLAB

CHAPTER 1 BACKGROUND

1.1 Cysteine residue and its oxidation states

Cysteine is one of the few amino acids that contain sulfur. This allows cysteine to bond in a special way though disulfide bond and help to maintain the structure of proteins. In addition, cysteines play very important roles in the function[1, 2] properties [3] as well as aging process of proteins[4]. As was shown in Figure 1, cysteine residues have two possible chemical states in proteins: oxidized state and reduced state. These two states are interchangeable when proper conditions are met. In its reduced form, cysteine undergoes chemical reactions such as alkylation[5, 6], oxidation[7] or forming complex compounds with metal ions[8]. These chemical reactions play critical biological roles such as activation, deactivation of the active sites of enzymes and altering the local environment of the proteins. In their oxidized form, two cysteine residues from the same protein chain form intra-chain disulfide bond, which links distant portion of a protein chain and provides strong structural constraints in the form of long-range interactions[9]. Two cysteines from different proteins form inter-chain disulfide bond and enable more complicated protein structures[10,11] and functions[12,13]. Interchange between the reduced form and oxidized form of cysteines also serves as regulation switches for enzymes[14] and other proteins.[15] Therefore, accurately predicting the oxidation states of cysteine is essential to the studies of protein stability[16], protein function[17,18] and three-dimensional structure[19] of proteins.

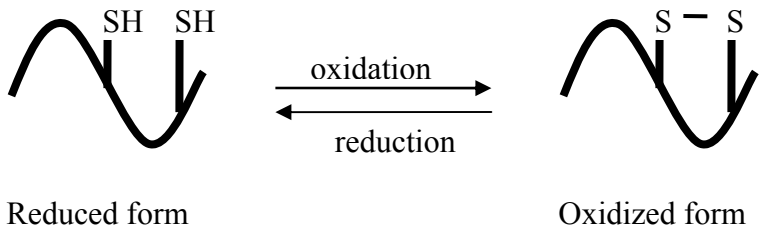


Figure 1 Reduced form and oxidized form of cysteine residues.

Experimentally, the determination of bonding states of cysteines involves use of costly biomarker and thiol reagent such as 4-Fluoro-7-sulfoamoylbenzofurazan (ABD-F) in combination with techniques such as electrophoresis[20] and HPLC[21] for quantitative amino acid analysis, mass spectroscopy etc[22], which are time consuming lab procedures and use costly bio-grade chemicals as well as expensive equipments. On the other hand, prediction of bonding states via computational approaches takes advantage of available knowledge and extends it to unknown systems, thus providing us a fast and relatively low cost way of understanding biological molecules. Therefore, prediction of the bonding states of cysteines by computational methods is of great value and importance.

1.2 Disulfide bonds

Disulphide bond is the covalent bond formed between cysteine residues on protein chains. They are usually found in proteins that are secreted from cells. Disulfide bonds are considered one of the elements of protein tertiary structure and directly contribute to the stability of the

protein [23]. Disulfide bonds play very important roles in the folding and stability of many proteins, particularly those secreted to the extracellular medium. Figure 2 is an illustration of protein in 3D with the disulfide bonds denoted in light green. The disulfide bonds stabilize protein folding in several mechanisms: a) They form a protein complex by firmly holding two or more distal portions and contribute to stabilize folding topology. b) The disulfide bonds serve as nucleus of hydrophobic cores of the folded proteins. c) Disulfide bonds link two or more regions on the protein chains increasing the effective local concentration of protein residues and lowering the effective local concentration of water molecules. d) The disulfide bonds stabilize alpha helical and beta sheet secondary structures and their nearby random loops by repelling water solvable molecules that attack residues' hydrogen bonds and break up secondary structures. e) In prokaryotes, disulfide bonds regulate a protein coding gene as protections of bacteria as a reversible on-off switch when bacterial cells are exposed to oxidation reactions. f) In eukaryotes, disulfide bonds play much more diverse roles and are important to reproductive systems in males. The prediction of disulfide bonds is thus an important but difficult task.

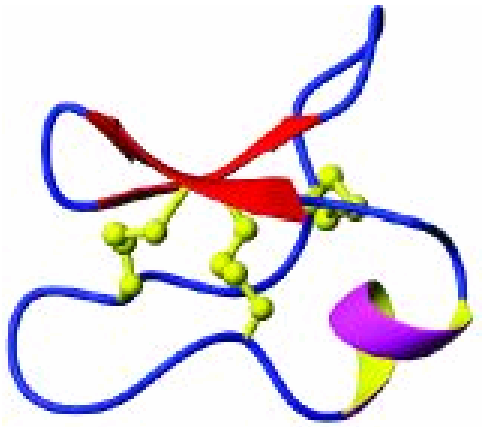


Figure 2 Illustration of disulfide bonds (shown in light green) on proteins

There are two types of disulfide bonds present in proteins, inter-chain and intra-chain disulfide bonds [24]. Both are formed by oxidation of two cysteines on the protein chain. As illustrated in Figure 3, the inter-chain disulfide bonds are formed between two cysteines from two different protein chains and the intra-chain disulfide bonds are formed between two cysteines from the same chain. Currently, most of the research efforts in the prediction of disulfide bond are focused on the prediction of intra-chain disulfide bonds. In this research only the inter-chain disulfide bonds are considered and cysteines that form intra-chain disulfide bonds are treated as non-bonded.

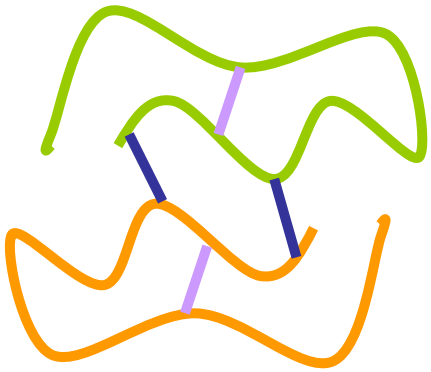


Figure 3 Inter-chain disulfide bonds (shown in blue) and intra chain disulfide bonds (purple)

As was observed in many proteins, disulfide bonds link distant portions of protein chains and provide strong structural constraints in the form of long-range interactions[25]. Prediction and knowledge of disulfide bond formation (connectivity) is important in reducing the search space

of protein conformation. Disulfide bond prediction has great value in predicting the protein three-dimensional structure, and in understanding the function and evolution process of proteins. With the rapid development in protein modeling and engineering, the ability to accurately predict the biologically stable disulfide bridges may even be more useful in introducing extra disulfide bonds to protein structures for better stability and controlled folding pathway [26], [27].

Prediction of disulphide bonds on a protein chain involves solving following problems[28]: i. Chain classification which distinguishes disulfide containing chains from non-disulfide containing chains; ii. Prediction of the oxidation state of cysteines, which identifies the candidate cysteines that actually involve in disulfide bonds (oxidation state prediction); iii. Prediction of a given pair of cysteines is linked or not (bridge classification) and iv. Prediction of the connectivity pattern of all oxidized cysteines on the chain. (connectivity prediction in the level of whole protein). Among these, connectivity prediction is the most challenging step.

1.3 Formation of disulfide bonds and role of Protein Disulfide Isomerase (PDI)

Formation of disulfide bonds in proteins involves de-novo formation and exchange[29]**.** This process is facilitated by the enzymes named as protein disulphide isomerases (PDI). In the next two sections, disulfide bond formation and the catalyst for the reaction are introduced in detail.

1.3.1 Disulfide bond formation

The reaction scheme of disulphide bond formation is shown in Figure 4. In the living organisms, the formation and dissociation of disulfide bonds are complicated. It involves not

only the direct formation via oxidation of the cysteine pairs, but also the process of error-correction. In cases where the wrong pair of cysteines is oxidized to form disulfide bond, a mechanism called isomerization[30] will be activated. Enzyme will scramble the mis-matched pair and reform the disulfide bond. PDI catalyzes the formation, rearrangement and breakage of disulfide bonds[31]. Although known for more than 30 years, its large size, multi-domain architecture and multiple reactive sites of PDI made it difficult to understand its structure and reaction mechanism.

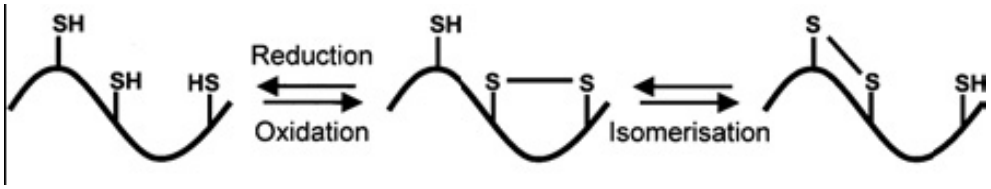


Figure 4: Formation of disulfide bonds

1.3.2 PDI--- Enzyme that catalyze the formation of disulfide bonds

Enzymes are defined as a class of multifunctional, multivalent macromolecules with the ability to bind small molecules and more importantly, subsequently to affect the reaction[32,33]. Proper binding between the enzyme and the substrate is the first step of the reaction and crucial for the enzyme-catalyzed reaction to occur. Figure 5 illustrates the well-established lock and key theory [34] of enzyme-substrate binding. The binding of a substrate to the active site of an enzyme

is a very specific interaction. The specificity of binding is usually realized by hydrogen bonds, ionic bonds, and hydrophobic interactions[35]. Once the substrate is bound to the active site of an enzyme, multiple mechanisms can accelerate its conversion to the product of the reaction.

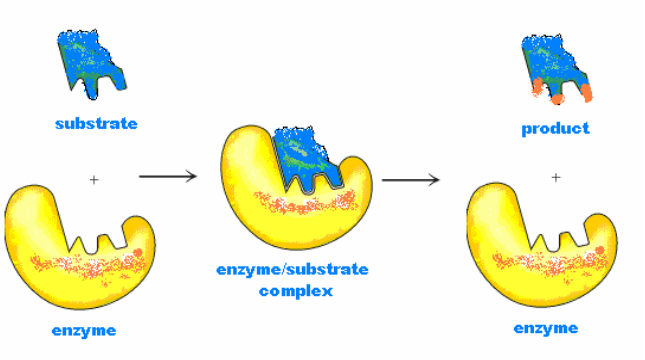


Figure 5: Illustration of mechanism of enzyme based on lock and key model [34, 36].

PDIs are a group of enzymes that facilitate and catalyze the formation, rearrangement and cleavage of disulfide bonds in secreted proteins. They are versatile in the catalysis of formation of new disulfide bonds, rearrangement of existing disulfide bonds as well as dissociation of undesired disulfide bonds. The multi-domain structure of PDI is believed to be the reason for this versatility. PDI consists of four major domains denoted as a, b, b' and a'. There is a common characteristic[37] about the reactive sites of PDI domains that assisted in forming disulfide bonds: a universal Cys-X-X-Cys pattern in the reactive site. (Different sub-

families tend to have different amino acids between the two cysteines, however, in one sub-family, the amino acids between the two reactive cysteines tend to be very well preserved).

Ruddock et al. [38] found that PDIp, a glycosylated PDI-related protein pancreas, has a special binding motif to its substrate: a tyrosyl side chain with a free phenolic hydroxyl group. Experiments indicated that if the sequence in the binding region is altered or blocked, the binding affinity of the PDIp and protein chain was greatly reduced; hence disulfide bonds were formed in a much less efficient manner. Pirneskoski et al. [39] identified the primary substrate binding region of PDI by modeling the tertiary structure of the b' domain and by mutation experiments within the site.

Based on the identified binding motifs of PDIp and the substrate binding region obtained by modeling the tertiary structure of some domains on the reactive sites of enzyme, it is believed that identification of the bonding sites between enzyme and substrate protein would gave valuable information on the actual pairings of cysteines. Similar to the situation when passengers try to buckle up the seatbelt when riding in a vehicle, the task performer (hand) has to hold one end of the belt to get it clicked in the other end. The assumption is that finding the binding sites of PDIs would provide some additional information on the location of disulfide bonds that were catalyzed by PDIs.

1.4 Influencing Factors for the purpose of prediction

The formation of disulfide bonds involves bringing the two cysteines into very close proximity and the process has to be facilitated by PDIs. Therefore the flanking sequences, the physical-chemical property of the whole protein, as well as other biological factors such as secondary structure and solvent accessibility play important roles. These factors have been accounted for in

the prediction of disulfide connectivity and the oxidation states of cysteines in literature. Here these factors are briefly introduced and their effects on the predictions are explained.

1.4.1 Flanking sequences around cysteines

Flanking sequences refer to the neighboring amino acids around cysteine. The polarity, hydrophobicity and charge of these neighboring amino acids provide physical-chemical environment for the formation of disulfide bonds. Because the formation of disulfide bond brings the flanking sequences around the cysteines to close proximity, the composition of the flanking amino acids directly contribute to the stability of the disulfide bonds formed. Figure 6 is an illustration of flanking sequences.

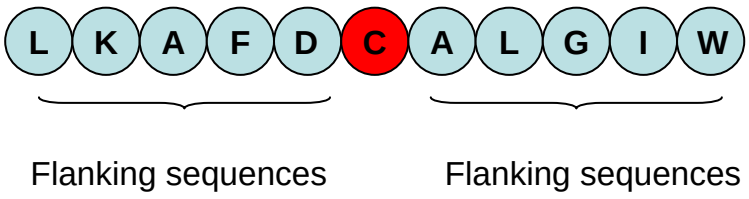


Figure 6: Flanking sequences around cysteine

Flanking sequences can be either represented by the order relative to the cysteine or by the frequency of each amino acid in the neighborhood. When represented in order, the interaction between the flanking sequences of two cysteines that form disulfide bonds is emphasized.

1.4.2 Global sequence information

Global sequence information includes number of cysteines in the chain, chain length and amino acid composition. This information indirectly indicates the size and the polarity of the protein chain.

1.4.3 Other biological factors

Except flanking sequences and the global sequence information described above, other biological factors such as secondary structure and solvent accessibility may also be useful. For an example, research found that formation of disulfide bonds greatly accelerate secondary structure formation in the folding of proteins[40]. And there is slightly more probability to find oxidized cysteines on β sheets than other secondary structures[41]. Solvent accessibility sheds light on how accessible a cysteine is to the reaction and stability to outside attacks.

1.4.4 Interactions between flanking sequences

This is specifically for the prediction of paring of cysteines to form disulfide bonds. When two cysteines form disulfide bond, the flanking sequences of both cysteines were brought to close proximity and inevitably interact with each other. The nature of the interaction contributes to the stability of the disulfide bond. Due to the critical roles of disulfide bonds in maintaining the proper 3D structures, the stability of the disulfide bonds formed is very much preferred. Therefore, contact potential or interactions between flanking sequences around

cysteines has been used as a criterion for actual cysteine pairing[99]. Figure 7 illustrates the interaction between flanking sequences.

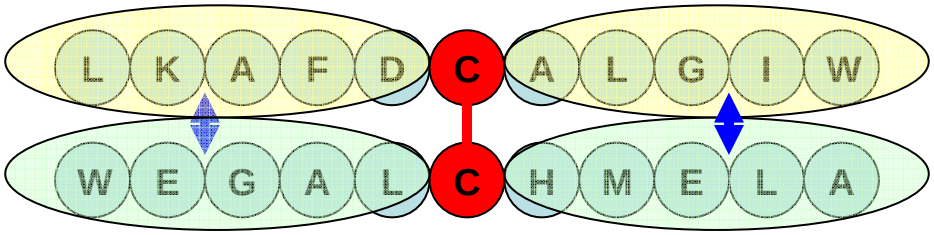


Figure 7: Interaction between flanking sequences.

The interaction between flanking sequences of cysteine i and cysteine j can be computed by $\omega_{i,j} = \sum_{k \in S_i} \sum_{l \in S_j} U(k,l)$, where S_i and S_j are the sets containing the nearest neighboring residues of cysteine i an cysteine j.

1.4.5 Distance between cysteine pairs

Distance between cysteine pairs can indicate the size of the protein in 3D structures and has been used for predicting the probability of cysteine pairing in lliterature[54]. Statistical analysis on 1018 protein chains with disulfide bonds indicate that close to 85% of disulfide bonds occur between cysteine pairs of 60 amino acid apart or less. Only 50% of the non-bonded

pairs have distance less than 60 amino acids. Therefore, when combined with other influencing factors, the distance between cysteine pairs can be used as a useful indicator in predictions.

1.5 Signal hypothesis on secreted and membrane proteins

Biochemistry textbook[23] showed that disulfide bonds are most commonly found in secreted proteins. A Nobel prize-winning procedure called pulse-chase experiment allows researchers to track the pathway of proteins from its synthesis to the final stage when the protein reaches its cellular destination[42, 43, 44, 45]. Nobel laureate George Palade and his coworkers illustrated that endoplasmic Reticulum(ER) (the irregular structure seen in Figure 8) is the essential organelle in the cells where membrane components are synthesized. Hence ER is a place of very active protein synthesis and after synthesis.

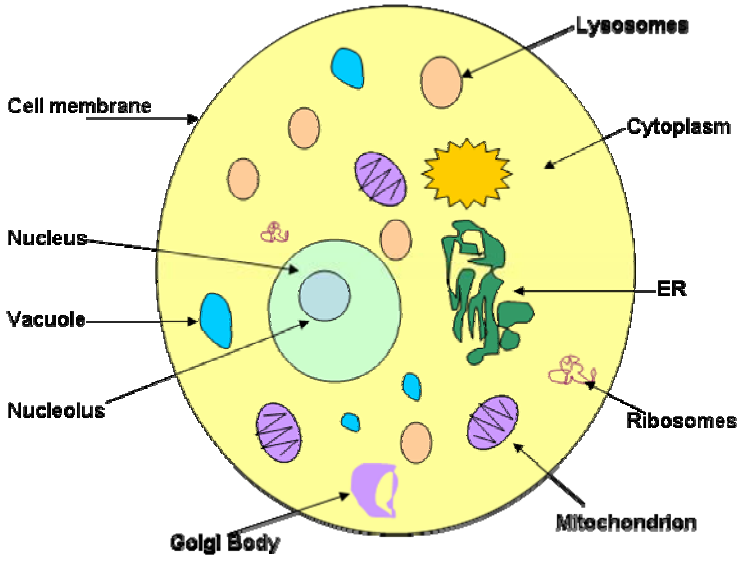


Figure 8: Cross section of animal cell

After synthesis in the rough ER area, the new protein has to leave the ER and travel through the membrane and reach its destination. It can be envisioned that it is difficult to transport a long protein chain into or through a membrane (lipid in composition) once it had been made in an aqueous environment, primarily because proteins behave differently in water and in lipid. In water, a protein tends to exposes its hydrophilic portion to the environment and keeps the hydrophobic core to itself. And in lipid, the hydrophilic portion of the chain is hidden and hydrophobic part exposed. Results of pulse chase experiments confirmed that that secreted proteins were synthesized in the rough ER, not on free ribosome in the cytosol where the secreted proteins are found to be present. This indicates that membrane and secreted proteins are transported through the membrane as they were being made.

One of the famous hypotheses on how the secreted and membrane proteins are transported out of ER after synthesis is the signal hypothesis. It was first proposed by Blobel and Sabatini[46] in 1975. Based on experimental observations, Blobel and Sabatini believed that the first part (N-terminus) of any membrane or secretary protein contains a special amino acids sequence that binds to some receptor in the ER membrane. This binding peptide contributes in two ways. 1. It attaches the mRNA, ribosome, and newly synthesized protein to the ER. 2. It threads the protein through the membrane to get out of ER and obtain the correct fold (Figure 9). After the transportation, the signal sequence may be nibbled off by an enzyme called signal peptidase. The signal hypothesis is supported by experiments and it is now generally believed to be essentially true, although scientists in the field agree that what is really happening in the cell is more complicated than this simple model suggested.

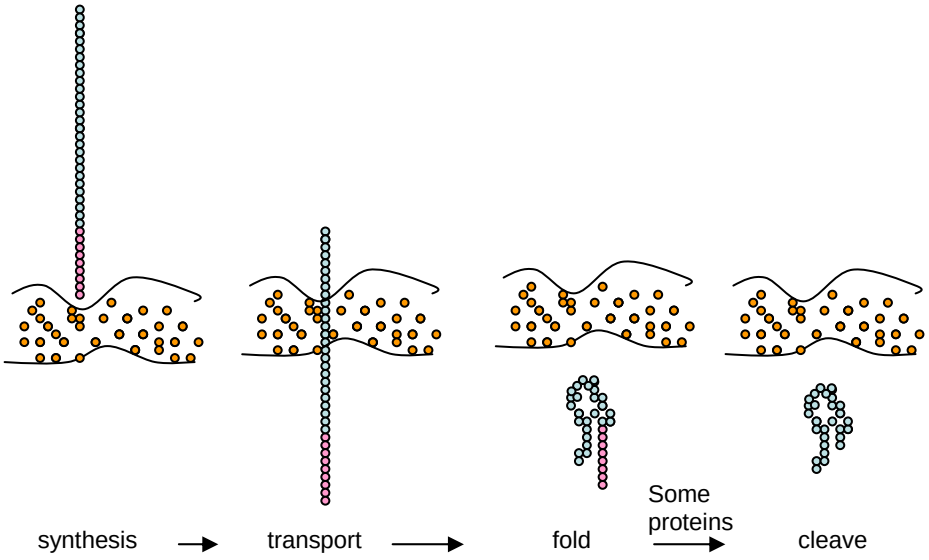


Figure 9: Protein transport according to signal hypothesis

From the biological evidences supporting signal hypothesis, it is clear that not all part of protein is critical for prediction purposes. Protein folding occurs after the protein's synthesis and secretion and the folding process requires the assistance of many enzymes and co-factors. Binding sites and recognizing sites are crucial for these enzymes and co-factors to work properly. As the first part of the protein to be transported outside of the membranes and obtain its correct folds, the N-terminus of the protein can not be neglected. And it is a reasonable suspect that the N-terminus of a protein carries more information than other part of the chain. Effectively capture of the most relevant features such as the N-terminus of the protein will greatly help with the predictions.

CHAPTER 2 INTRODUCTION

As it is seen in the background introduction in the previous chapter, there are a number of factors that could influence the oxidation states of cysteines as well as the connectivity of oxidized cysteines on proteins. Except the biological factors, various prediction methods such as artificial neural networks, support vector machines, statistical methods were used to capture the most of the characteristics of the oxidation and connectivity patterns. And because of the different ways influence factors were represented and the model was built as well as the way prediction error was assessed and minimized, the prediction methods to be used for prediction also have an impact on the final accuracy. In this chapter, the motivation for the way in which prediction done were first briefly introduced and then the organization of this thesis is outlined for the prediction of oxidation states of cysteines and for the prediction of connectivity pattern of disulfide bonds.

2.1 Prediction of oxidation states of cysteines

As noted in the previous chapter, knowledge of the oxidation states of cysteines infer a lot of information about the protein such as local sequence environment, the possible three dimensional structure of protein and in some cases, the function and working mechanisms of the protein. For the prediction of oxidation states of cysteines, flanking sequences around the cysteines in consideration and the characteristic of the whole protein were among the most commonly used. And sequences other than the flanking sequences of the protein were given equal amount of consideration. This might not necessarily be a fair decision because from signal hypothesis, it is apparent that certain part of protein can be more informative than other parts in certain aspects. Based on these thoughts, it might be more reasonable to give more emphasis on

the one end of protein chains that carry more information during transportation and folding process. Therefore in this research, except the commonly used characters such as flanking sequences and global protein information, the oxidation state information of the amino terminus of protein were also incorporated in the prediction. It is believed that his assumption is more experimentally achievable than assuming the knowledge of the transition probability of the oxidations states of protein chains in the particular testing sets.

As far as prediction methods are concerned, machine learning techniques were adopted mainly because of their proven effectiveness in extracting information from existing data[47]. Improved artificial neural networks and support vector machines were both adopted in this research and comparison was made.

2.2 Prediction of connectivity pattern of cysteines

Connectivity pattern prediction is a challenging and yet very biological meaningful task. It is challenging because there are two many possibilities of disulfide bonding for a given protein and many factors can influence the final connection patter. Even in the living organism, sometimes the Protein Disulfide Isomerases would have to provide an extra function to dissociate the wrongly formed disulfide bonds and reform the correct foldings. However, correctly predicting the disulfide connectivity can also be very rewarding, mainly because of the critical roles disulfide bonds provide in order to have a stabilized three dimensional protein structure. In this research, a different approach from past research efforts (will be detailed in Chapter 5) was proposed and tested. Multiple sequence alignments were first conducted on a number of proteins to confirm the existence of the conserved binding motifs on protein followed by either incorporating these features into calculation of flanking sequence interaction or integrating the binding motif information into the encodings of machine learning techniques.

Then two different prediction approaches in evaluating the disulfide formation probability of cysteine pairs. For the first approach, the interaction between the flanking amino acids between cysteine pairs was first calculated. Then the interaction was further adjusted according to whether the binding motif was located and by checking the distance between cysteine pairs. Based on the adjusted interactions for all possible pairs, a maximum weighted graph matching algorithm was applied to the completely connected graph constructed from all possible pairs of cysteines with the weight of each edge to be the adjusted interaction. The prediction of disulfide bridges via calculated interaction between flanking sequences is described in detail in Chapter 5.

In the second approach, the binding motif, flanking sequences around both cysteines, global information of the protein chain and the distance between cysteine pairs were incorporated into the encoding for machine learning techniques. Through machine learning, the probability of disulfide formation to each possible pair of cysteines was assigned according to the model built by machine learning techniques. And this probability was used as the weight for the connection between the cysteine pair. Then maximum weighted graph matching algorithm is applied to the graph obtained and top n connections in the resulted graph are selected to be disulfide bridges. Chapter 6 covers the details of the prediction by machine learning approach based on the binding motif of PDIs.

2.3 Organization

This research is organized as follows: in chapter 3, related works on the prediction of cysteine oxidation state were first reviewed, followed by prediction of oxidation states of cysteines using new set of SVM encoding which accounted for the N-terminus cysteines. Different influencing factors were discussed in details. Chapter 4 is prediction of oxidation states of cysteines using Associative Neural Networks (ASNN). The performance of both methods was

compared. The next two chapters mainly focus on the prediction of connectivity of cysteines. In chapter 5, related literatures were reviewed on the computation and prediction of disulfide connectivity. Then prediction of disulfide connectives by calculating the direct interactions between the flanking sequences of cysteine pairs was discussed. In chapter 6, based on the results from multiple sequence alignments, new predictive method based on searching for enzyme anchor sites is introduced and influencing factors discussed. Predictions based on the interaction of flanking sequences, global protein features, distance of the cysteine pairs as well as binding motifs to PDIs were discussed and prediction performance compared. Chapter 7 summarizes this work and future improvements were proposed.

CHAPTER 3 PREDICTING THE OXIDATION STATES OF CYSTEINES USING SUPPORT VECTOR MACHINES (SVMs)

3.1 Related works on prediction of oxidation states of cysteines

Prediction of bonding states via computational approaches takes advantage of available knowledge and extends it to unknown systems, thus providing us a fast and cost-effective way of understanding biological molecules. Therefore, prediction of the bonding states of cysteines by computational methods is of great value and importance.

There have been a number of studies in the prediction of the bonding states of cysteines via computational approaches. Muskal et al[48] obtained 81% accuracy by using neural network with the sequence information around the cysteines as inputs. Fiser et al.[49] performed statistical analysis on the amino acid frequencies in the sequence environment of cysteines and achieved 71% accuracy on a different testing set. Mucchielli-Giorgi et al.[50] assessed the relative efficiency of different descriptors in predicting the cysteine disulfide bonding states and concluded that the amino acid content of the whole protein is more informative than the flanking sequences. This approach yielded 84% accuracy. Martelli et al,[51, 52] implemented a new hybrid system that combines a neural network and a hidden Markov model (hidden neural network) and obtained an overall accuracy of 88%. Frasconi et al[53] introduced support vector machines (SVM) based predictor that operated in two stages which incorporated information at the protein level and local sequences and achieved an accuracy of 83.6%.

By using the SVM based on multiple feature vectors and statistical results on cysteine state sequences on the dataset, Chen *et al.*[54] achieved the record high accuracy of 90%. This is the current record for the prediction of oxidation state of cysteines. In this work, it was pointed out that cysteine state sequence has significant influence on the prediction accuracy. Most recently, Ceroni et al[55] employed SVM binary classifier to predict the bonding states of each cysteine, followed by a refinement step that takes the overall bonding state assignment of the entire chain into consideration and achieved an accuracy of 88%. The literature was summarized in Table 1.

Table I: Summary of past efforts on prediction of oxidation states of cysteines

year	author	methods	input	accuracy (%)
1990	Muskal et al[48]	neural networks	sliding window of flanking sequences	81
1992	Fiser A et al[49]	statistical analysis	sequence around cysteines	71
1999	Fariselli et al [56]	multiple sequence alignments	evolutionary information	81
2000	Fiser A et al[49]	multiple sequence alignments	statistical analysis on flanking sequences	82
2002	Mucchielli et al[50]	logistic functions	subsets of proteins homogeneous in terms of their amino acid content	84
2002	Muartelli et al[52]	hidden neural networks	both local and global characteristics of proteins	88
2004	chen et al[54]	support vector machine	flanking sequences, global protein info, cys-oxidation state pattern	90

3.2 Introduction to Support Vector Machines

Support Vector Machine (SVM) is a learning system that uses a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from

optimization theory[57]. This learning scheme was designed by Vapnik and his co-workers[58] and proved to be a very powerful machine learning technique since its introduction.

Before proceeding to introduce SVM, how the prediction errors (risks of predictions) are assessed in machine learning will be briefly discussed. For any single data point (x,y), where y is predicted to be f(x), the prediction error can be evaluated by simply comparing the predicted result f(x) with the actual value y. If they match, the prediction error is 0, otherwise, prediction error is 1.

$$E(y, f(x)) = \begin{cases} 0 & if \quad y = f(x) \\ 1 & otherwise \end{cases}$$

For learning machines, the error associated with the training set of the learning machine is called empirical risk. Suppose a learning machine with adjustable parameter λ, and input data pairs are in the form of (x_i,y_i), the empirical risk Rem is evaluated as the mean error over the training data:

$$R\text{em}(\lambda) \quad = \frac{1}{l}\sum_{i=1}^{l} E(y, f(x, \lambda))$$

where l is the size of the training data.

The expected error for testing is called the actual risk and it can be calculated by

$$R(\lambda) = \int \mathrm{E}(\mathrm{y}, \mathrm{f}(\mathrm{x}, \lambda))\, d(P(x, y))$$

where P(x,y) is some distribution probability from which the testing data are drawn.

Conventional machine learning techniques such as neural networks emphasize the minimization of empirical risk, which is the average error over the training set and does not take much effort to reduce the prediction error. Different from these conventional machine learning

techniques, SVM engages in Structural Risk Minimization (SRM). The goal of SRM is to achieve a trade-off between empirical risk and the capacity of the learning machine. In the next paragraph, how SVM computes the classification models is very briefly introduced and the advantages of SVM offers are summarized.

3.2.1 Support Vector Machines (SVMs) for linearly separable data

As shown in Figure 10, for a group of linearly separable data, there exists a hyper plane (solid dark line) that separate the two classes. w is a vector normal to the separation hyper plane and b is the distance from the hyper plane to the reference point. This hyper plane can be denoted as:

$$f(x, \lambda) = \text{sgn}\,(wx + b), \text{where x, w} \in R^N$$

And the two classes satisfy the following inequalities:

$$x_i w + b \geq +1 \quad for \quad y_i = +1$$
$$x_i w + b \leq -1 \quad for \quad y_i = -1$$

The inequalities can be generalized to be:

$$y_i (w x_i + b) - 1 \geq 0 \qquad \forall i$$

Now consider the data points that are nearest to the separation hyper plane and these data points defines two new hyper planes (shown in dotted line in Figure 10) that are parallel to the separation hyper plane. Intuitively, the larger the distance between these two new hyper planes, the better chance it is for a better classification model. This distance between the new hyper planes is called margin. Therefore, finding the data points that define hyper planes with maximum margin between the two classes will result in minimized prediction error. The margin can be calculated as

$$\text{margin} = \frac{2}{||\text{w}||} \;\; ; \;\; ||\text{w}|| \text{is the Euclidean norm of w}$$

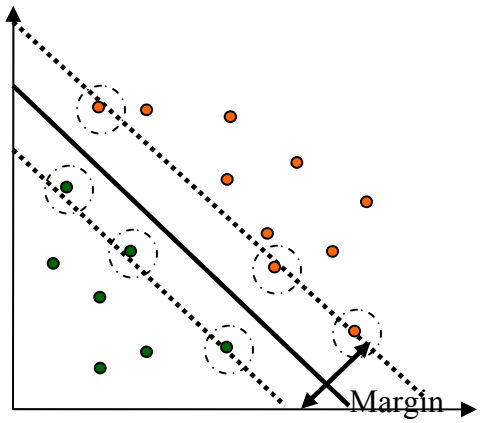


Figure 10: Calculation of support vectors for simple classification problem

Finding minimized ||w|| can be done through Lagragian formulation and subsequently linear programming. Those points that define the hyper planes with maximized margin are the support vectors.

3.2.2 Advantages of SVM

Compared with other machine learning techniques, SVM has a number of superior properties, such as effective avoidance of over-fitting, the ability to handle large feature spaces and information condensing of the given data set etc. It has been successfully applied to a wide range of pattern recognition problems, including isolated handwritten digit recognition, object recognition, speaker identification, and text categorization, etc. Figure 11 illustrates the process of information discovery and where SVM fits in the process. Before applying SVM, it is important to know that data selection and sampling is crucial to the prediction accuracy. SVM

provides an efficient tool for successful classification. Without meaningful feature selections, SVM itself is limited in giving meaningful predictions.

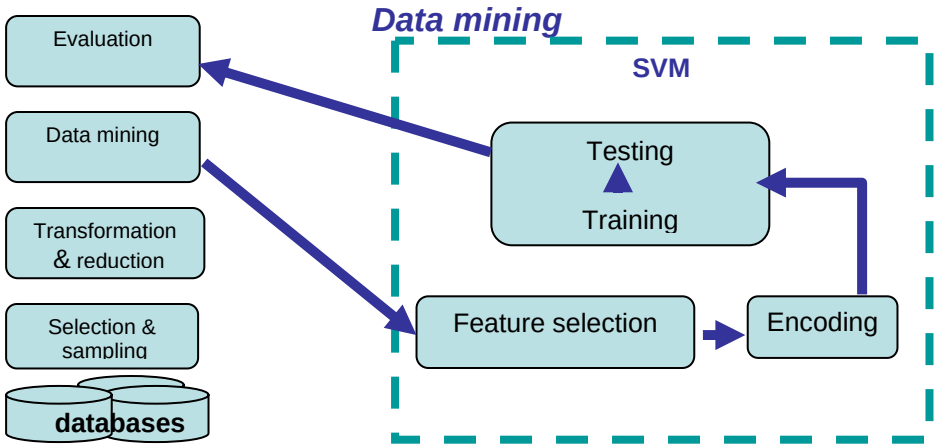
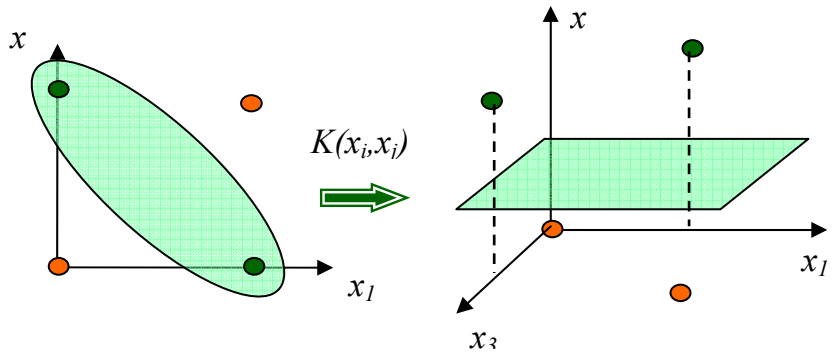


Figure 11: Process of information discovery from data bases

3.2.3 Support Vector Machines (SVMs) for non-linearly separable data

In the previous section, how support vectors were obtained for linearly separable was briefly illustrated. It is very similar for the case of non-linearly separable data. For complicated data that can not be separated linearly, kernel functions can be applied to make them linearly separable in a higher dimension (shown in Figure 12). Then these data would be able to be treated as linearly separable and apply the fore-mentioned algorithm to calculate the support vectors.



(a) Non linearly separable (b) Linearly separable

Figure 12 Transform non-linearly separable data to linearly separable in higher dimension by kernel method

3.2.4 SVM encodings

Given data $x_1, \ldots, x_1$, the label (also called target) for these data was set as +1 if x_i is in class 1 (in our case, +1 if the cysteine residue is bonded with another cysteine in the same protein chain (intra-chain bonded)) and set as -1 if x_i belongs to class 2 (in our case, -1 if the cysteine residue is non-bounded or is bonded with another cysteine from a different protein chain (inter-chain bonded)). Then with these training data, SVM solves an optimization problem for binary classification:

The SVM encodings are comprised of target value followed by multiple feature:value vectors of the following format[59]:

<line> .=. <target> <feature>:<value> <feature>:<value> ... <feature>:<value> # <info>

Where the target, feature and value are defined as the following:

<target> .=. +1 | -1 | 0 | <float>
<feature> .=. <integer> | "qid"
<value> .=. <float>
<info> .=. <string>

In applications, SVM is usually first trained to recognize the input patterns. Based on the training dataset, SVM computes support vectors and returns a value for each line of input in the testing dataset. The decision was made based on the magnitude of value returned by SVM. Usually a +1 decision is assigned if the returned value was larger than 0. Otherwise, -1 is assigned.

3.3 Motivation of this research

As described in the previous section, SVM techniques has been applied to the prediction of oxidation states of cysteines in previous research efforts [54,55] and it has been shown that SVM is effective in improving the prediction accuracy over the past research that utilized other techniques such as statistical analysis and neural networks. However, both Chen[54] and Ceroni's [55] works involve two steps. The first step is to use the flanking sequences around the cysteine residue considered and global information on the whole protein chain as the input to the SVM and obtain a rough estimation of the bonding states of cysteines. Then a second refining step is applied to assign the bonding states of all cysteines on the protein chain.

These research efforts demonstrated to us that flanking sequence and global environment parameters are crucial to the predictive accuracy. However, after studying these research methods, question arises on if SVM can also handle the second step used in previous research. For an example, in Chen et al's work, the second step calculates the probability of the oxidation states of cysteines based on the oxidation states of other cysteines on the chain. And this calculation helps to improve the predictive accuracy from 88% to 90%, which indicate that oxidation states of one cysteine on the protein chain does affect the oxidation state of other cysteines. As it was generally known, SVM solves classification and regression problems by

"simultaneously minimizing the empirical classification error and maximize the geometric margin"[60]. SVM can take a large number of vectors in each training/classification example without compromising too much as far as computation time is concerned. Theoretically, every relevant factor can be encoded into the input to SVM, although initially it may not be clear which of these factors are more directly related to the prediction than the others.

Here we propose the use of a single step of SVM classification, which incorporate all relevant information into the SVM classification and minimizes the chance of mis-classification caused by a two step process. The idea is to incorporate all possible combinations of influencing factors into the input vectors and let the SVM build a model and determine which factor or combination of factors gave the best prediction results.

In our consideration of possible influencing factors, we take information such as flanking sequence, global information into consideration. Examining the 90% accuracy paper in a greater detail (Figure 13), it can be found that the 90% accuracy is actually obtained by assuming the statistics of the oxidations states for protein chains in the dataset was known before any prediction was performed. In actual situation, we may not be so lucky to know the statistics of the oxidations states for proteins with certain number of cysteines in the testing dataset. And it is very possible that the protein chain we are trying to predict is so different from the statistical data we used in the prediction that it leads to totally inaccurate prediction.

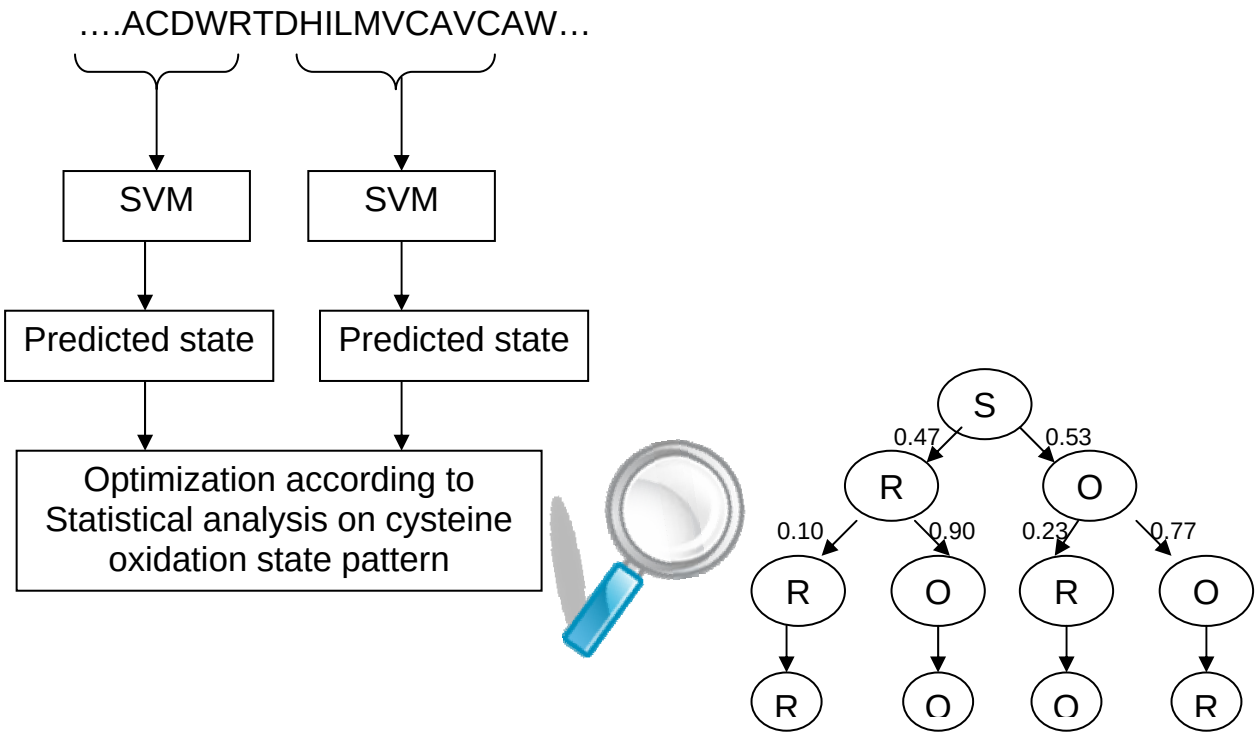


Figure 13: A closer look at the two step process in the current record holder[54].

Based on these thoughts, and inspired by the signal hypothesis for secreted and membrane proteins, as well as the biological findings that the terminals of proteins usually carry critical signals for the whole protein chain[61, 62, 63, 64,65,66, 67], we think it is more reasonable to assume the knowledge of the cysteines on the one end of protein chains. We believe this assumption is experimentally achievable than assuming the pre-knowledge of the bonding probability of the oxidations states of protein chains in the particular testing sets. Encodings to SVM were developed by assuming the knowledge of the bonding states of the first and second cysteines on the amino terminal of the protein chain. In combination of flanking sequence information, and global information such as amino acid composition and chain length etc, the

developed encoding is expected to establish the relationship between the determining factors about the oxidation states and the overall accuracy of prediction.

Also to push the assumption of known oxidation states of all other cysteines to the upper limit, we tried adding a new set of vectors that uniquely identify each individual cysteine residue and its reduction/oxidation state environment (oxidation states of all its peers in the same chain) and found it does not outperform in performance where the amino terminus cysteines states were assumed. This confirms our assumption that the first couple of cysteines are crucial for determination of oxidation states of other cysteines on the chain.

Also the last effort seen in prediction of oxidation states of cysteine using neural networks was in 2002. Since then, more powerful and efficient neural nets has been developed such as ASNN, which aims at correcting the bias caused by traditional neural networks and at the same time provides a more efficient way in training with new datasets. Therefore, we also adopted ASNN in our prediction in chapter 4 and studied the effectiveness of this technique in the prediction of oxidation states of cysteines. A side by side comparison of SVM with ASNN was also done in next chapter.

3.4 Prediction of oxidation states of cysteines by SVM

In this section, SVM was used to predict the oxidation states of cysteines. Multiple sequence features were encoded for the purpose of training.

3.4.1 Encodings with pre-knowledge of oxi-states of N-terminal cysteines

Multiple feature vectors composed of information on flanking sequences, global information of the protein chain, oxidation states of up to two cysteines on the amino end of protein chain and ordinal number of the cysteine being considered were used in this study. The

overall encoding is arranged in the following order: target(1/-1), vectors for global information, vector for ordinal number of cysteine in consideration, vectors for flanking sequences and vectors for oxidation states pattern of up to two cysteines on the amino terminal of the interested protein chain. The overall coding layout is illustrated in Figure 14. The encoding details of all available information are described in detail in following paragraphs.

1/-1	Vectors for global info	Vector for ordinal number of cysteine in consideration	Vectors for flanking sequences	Vectors for oxidation state pattern of up to two cysteines on N terminus

Figure 14 Coding layout of the multiple feature vectors (from left to right)

3.4.1.1 Global information vector:

The global information vector is composed of two parts. One corresponds to the general information of the protein such as total number of cysteines and the total number of amino acids in the protein chain. The other part corresponds to the actual amino acid composition of the protein chain. In previous research by Chen *et al.*[54], it was found that by inclusion of global information, the predicting accuracy was greatly improved. To improve the accuracy of the prediction, we have the global information vectors included in some of our tests. The features for these two sets of vectors were randomly chosen. "1" was used as the feature for total number of cysteines and "10" was used as the feature for the protein length (number of amino acids on

chain). And from 350 to 370 indicates the actual percentage of certain amino acids in the whole chain. (AA% =total number of AA/total number of amino acids on the chain x 100).

3.4.1.2 Oxidation pattern vectors:

The oxidation pattern vectors are composed of two parts: the ordinal number vector and the oxidation pattern of up to two cysteines on the amino end of the protein chain, which we believed that combined with the flanking sequences, form a unique representation for each of the cysteines on the same protein chain.

3.4.1.3 Ordinal number vector:

The ordinal number vector locates where the cysteine in consideration is located relative to the other cysteines in the protein. We used 5 as the feature value for this vector because we put it in the front part of the encoding. Other feature value should also work fine as long as the entire encoding for all cysteines is consistent. To make the value of this vector have the similar magnitude as the amino acid representation used in the flanking sequences, we multiplied this ordinal number by 10. For example, if the sixth cysteine is considered, the vector is designated as 5:60. This vector is the key difference of the encoding used in this study from previous SVM encodings. By indicating the ordinal number of the cysteine in consideration, the encodings of cysteines on the same protein chain became distinctive from each other.

3.4.1.4 Oxidation states pattern of N terminal cysteines

The vectors for oxidation state pattern of amino terminal cysteines describe the oxidation states of up to two cysteines on the amino end of the chain. The features of these vectors can be any number, as long as it is consistent with the whole encoding and manifest the order of the cysteines in the protein chain. Number 10 was used as the value of the vector if the cysteine

belongs to an intra-chain disulfide bond and -10 if it does not present in intra-chain disulfide bonds.

3.4.1.5 Flanking sequences vectors.

The flanking sequence vectors are determined by the n adjacent amino acids of cysteines, where n is the window size. In this study, the feature for each amino acid manifested how close the residue is to the cysteine in consideration. The value of the vector for the amino acid is determined by the physical-chemical properties of the flanking amino acids. Representations (See appendix I) developed by Meiler et al.[68] were used in this study, which accounted for five parameters of the 20 amino acids. The five parameters include graph shape index, polarizability, volume, hydrophobicity and isoelectric point of amino acids. Two encoding schemes for the flanking sequences were used: by order and by frequency.

Flanking sequence By order: The features of flanking sequence vectors is encoded as how far away flanking amino acid is close to the cysteine. The value of the vector is the amino acid representation previously mentioned. To accommodate the situation when 50 flanking amino acids were encoded the feature of the oxidation pattern vectors for cysteines are correspondingly incremented. It was tested that as long as the encoding is consistent in the while dataset, there is no difference in the testing results.

***Flanking sequence by frequency*:** The feature of the flanking sequence vectors is the amino acid representations mentioned above. The value of the vector is the frequency the residue appeared in the considered flanking sequence window.

Table II: Example encodings for a protein chain that has six cysteines with the oxidation patter as ORROOO.

Cys being encoded	target	Global Vectors	Ordinal vector	Flanking sequence vectors	Vectors describing oxidation pattern of other cysteines
1st cys	1	XX..X	5:10	AA..A	101:-10
2nd cys	-1	XX..X	5:20	BB...B	101:10
3rd cys	-1	XX..X	5:30	CC..C	101:10 102:-10
4th cys	1	XX..X	5:40	DD..D	101:10 102:-10
5th cys	1	XX..X	5:50	EE…E	101:10 102:-10
6th cys	1	XX..X	5:60	FF…F	101:10 102:-10

Note: (O: oxidized form, R: reduced form; XX…X denotes the vectors for global information. And AA…A to FF…F denote the flanking sequence of the cysteine in consideration. The encoding details of the global information and flanking sequences were described in the encoding section and not shown in table I.)

Table II is an illustration of the encoding used for a protein chain with six cysteines. The sample protein has a cysteine oxidation pattern of ORROOO.

3.4.2 Extending the assumption - with known oxidation pattern of whole chain

To evaluate how the amino terminus cysteines' oxidation pattern influence the oxidation states of other cysteines on the protein chain, we extended our assumption to upper limit—assuming the oxidation pattern is known and see how much this assumption outperform the N-terminus cysteines' oxidation states. Multiple feature vectors composed of information on flanking sequences, global information of the protein chain, oxidation states pattern of cysteines and ordinal number of the cysteine being considered were used in this study. The overall coding is arranged in the following order: target(1/-1), vectors for global information, vector for ordinal number of cysteine in consideration, vectors for flanking sequences and vectors for oxidation

states pattern of other cysteines. The overall coding layout is illustrated in Figure 15. The encoding details of the oxidation pattern is described in detail and illustrated in Table II.

1/-1	Vectors for global info	Vector for ordinal number of cysteine in consideration	Vectors for flanking sequences	Vectors for oxidation state pattern of other cysteines

Figure 15: Coding layout of the multiple feature vectors (from left to right)

The oxidation pattern vectors are composed of two parts: the ordinal number vector and the oxidation pattern of all other cysteines on the protein chain form a unique representation of the cysteines on the same protein chain. The ordinal number vector was described in detail in previous section. These oxidation pattern vectors (abbreviated as pattern) is believed to be the most contributing factor in training the SVM to differentiate the bonded cysteines from those that are not bonded.

The vectors for oxidation state pattern of other cysteines describe the oxidation states of all other cysteines in the chain. These vectors, paired with the ordinal number of the cysteine in consideration, gave a unique tag for each of the cysteine in the chain. The features of these vectors can be any number, as long as it is consistent with the whole encoding and manifest the order of the cysteines in the protein chain. Number 10 was used as the value of the vector if the cysteine belongs to an intra-chain disulfide bond and -10 if it does not present in intra-chain

disulfide bonds. Note that the cysteine being considered was not included in the coding of the pattern since its ordinal number has already represented its actual location in the chain. Leaving out the oxidation state of the cysteine in consideration provided a distinctive set of vectors for each cysteine on the chain. For example, if we denote O as oxidized cysteines and R as cysteines in the reduced form, a chain with 6 cysteines in the order of ORROOO would give the encoding for all cysteines in the chain as the encodings in Table III.

Table III: Example encodings for a protein chain that has six cysteines with the oxidation patter as ORROOO.

Cys being encoded	target	Global Vectors	Ordinal vector	Flanking sequence vectors	Vectors describing oxidation pattern of other cysteines
1st cys	1	XX..X	5:10	AA..A	101:-10 102:-10 103:10 104:10 105:10
2nd cys	-1	XX..X	5:20	BB...B	101:10 102:-10 103:10 104:10 105:10
3rd cys	-1	XX..X	5:30	CC..C	101:10 102:-10 103:10 104:10 105:10
4th cys	1	XX..X	5:40	DD..D	101:10 102:-10 103:-10 104:10 105:10
5th cys	1	XX..X	5:50	EE…E	101:10 102:-10 103:-10 104:10 105:10
6th cys	1	XX..X	5:60	FF…F	101:10 102:-10 103:-10 104:10 105:10

Note: O: oxidized form, R: reduced form; XX…X denotes the vectors for global information. And AA…A to FF…F denotes the flanking sequence of the cysteine in consideration. The encoding details of the global information and flanking sequences were described in the encoding section and not shown here.

3.4.3 Encodings with more emphasis on the N-terminus information but no assumptions

Because of the signal hypothesis on secreted and membrane proteins and biological literatures that indicates the N-terminus of a protein contains more structural information that

other part of the protein, it is tempting to simply include the N-terminus sequences and check if the N-terminus sequences, rather than the oxidation states of cysteines on the N-terminus, will contribute to the prediction of cysteines' oxidation states. Encoded in this part were global information vectors, flanking sequence vectors of the cysteine in consideration and the amino acid sequences around the first two cysteines. The detailed encoding are very similar to what was described for section 2.3.1 except no assumption of cysteine states on N terminus and the flanking sequences around the first two cysteines were included in the encoding.

3.5 Data sets

Two dataset are used in this study. The first one is the same as the one used in research done by Martelli et al[51]. The dataset is comprised of 4136 cysteine-containing segments. There are 1446 segments in the disulfide bonded states and 2690 segments in the non-bonded states. The segments were extracted from 969 non homologous proteins from Protein Data Bank (PDB)[69]. The sequence identity of the set is less than 25% and there is no chain breaks. Table IV lists the number of sequences in the dataset divided according to the number of cysteines on each protein chain. In this dataset, cysteines that form inter-chain disulfide bonds were classified as non-bonded or free cysteines. The results reported in this work were from 20-fold cross-validation.

Table IV: statistics of dataset used for cysteine oxidation states prediction, divided according to the number of cysteines on each protein chain.

Number of cysteines per chain	Number of sequences in dataset
1	187
2	207
3	106
4	144
5	71
6	80
7	35
8	55
9	18
10	16
11	4
12	16
14	7
15	1
16	8
17	2
18	4
22	1
23	1
24	3
30	1
34	1
35	1
TOTAL	969

The second data set was compiled by Chen et al[28]. This dataset was first used in 2006[28] and consist of more new sequences from PDB. The dataset was obtained by downloading all proteins from the PDB (sequences up to May 17, 2004). Among the 25,465 proteins, 26.8% contain at least one disulfide bond, 89% of which contain exclusively intra-chain disulfide bonds. SPX was obtained by further filtering based on the homology-derived secondary structure of Proteins (HSSP) distance[70] (cutoff distance was set to 10) and 1018 chains were retained in the

dataset. The SPX dataset has a total of 5983 cysteines and 15% (901) of them are in reduced form. The statistics of this dataset is shown in the following table (Table V).

Table V: the number of sequences in the dataset, divided according to the number of cysteines on each protein chain.

n: number of cysteines on chain	number of chains with n cys
2	193
3	71
4	204
5	53
6	223
7	31
8	92
9	18
10	55
11	5
12	25
13	2
14	13
15	1
16	8
17	5
18	4
19	1
20	3
21	1
24	3
28	1
32	2
34	1
35	1
50	1
54	1
Total	1018

Except the SPX contains more and newer sequences in the dataset, another difference between the SPX dataset and the previous dataset was that SPX dataset does not contain inter chain disulfide bonds and cysteines that are capable of forming inter chain disulfide bonds.

3.6 Evaluation of predictive accuracy

The prediction accuracy was calculated by following the standard conventions[71]: accuracy for prediction:

$$Q_2 = N_c/N_0$$

where N_c is the total number of correctly predicted cysteines and N_0 is the total number of cysteines. Specificity of the prediction

$$\text{Spec} = \frac{TN_x}{TN_x + FP_x}$$

where x denotes the bonded cysteines or non-bonded cysteines, FP_x is the number of false negatives in the prediction and TP_x is the number of true positive predictions for bonding state x.

Sensitivity of the prediction Sens was calculated as:

$$\text{Sens.} = \frac{TP_x}{TP_x + FN_x}$$

where FN_x is the number of false negatives for bonding state x. The Matthews correlation coefficient (MCC)[72] is calculated as:

$$\text{MCC} = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

TN_x is the true negatives of bonding state x. The value of MCC is an indication of how good is the prediction. The closer the MCC is to 1, the closer the prediction is to a perfect prediction.

Another measure of accuracy used in this study is the Youden Index[73]. Youden's index is one way to attempt summarizing test accuracy into a single numeric value.

$$\text{Youden's index} = 1 - ((\text{false positive rate}) + (\text{false negative rate}))$$
$$= 1 - ((1 - (\text{sensitivity})) + (1 - (\text{specificity})))$$
$$= (\text{sensitivity}) + (\text{specificity}) - 1$$

3.7 Result and discussion on the SVM prediction

Based on the different encodings described in the previous section, prediction results are evaluated. The following paragraphs discuss the prediction results in detail.

3.7.1 Overall testing results

The performance of different combinations of input vectors was compared. The testing results are shown in Table VI. Except the listed vectors in the table, all of these testing also included the general information about the chain such as total number of cysteines and total number of amino acids in the chain. Also included in each of these testings are the ordinal numbers of the cysteine in consideration. The SVM prediction using the combination of known bonding states of first and second cysteines from the amino end of the protein chain and the global information yielded the best accuracy among all the tests (94.8%). This is 4.8% higher in accuracy than the current best result in literature, which assumes the knowledge of bonding state pattern of the whole training dataset being tested. Adding the flanking sequence information does not improve the prediction result.

Table VI: Testing results with up to 2 N-terminus cysteine oxidation states known.

	SS.sens.	SS spec.	SH sens	SH spec	accuracy
1 st cys known	0.721	0.898	0.956	0.865	0.874
1,2 cys known	0.904	0.942	0.970	0.949	0.947
Global info +1st ,2nd cys known	0.907	0.942	0.970	0.951	**0.948**
flanking seq. and 1st, 2nd cys	0.901	0.941	0.970	0.948	0.946
1st ,2nd cys and flanking seq. and global info	0.909	0.941	0.969	0.952	**0.948**

Notes: Also included in the encoding are global information and flanking sequences. (every testing also included general information about the chain such as number of cysteines and number amino acids on the chain)

3.7.2 Influencing factors

Between the 1st and 2nd cysteine on the amino end of the chain, the oxidation state of second cysteine is more influential to the oxidation states of other cysteines. As is shown in Table VII, knowledge of the oxidation state of the second cysteine alone combined with the global information and flanking sequence information give a 90.8% accuracy.

Table VII: The relative importance of first 2 cysteines on N-terminus.

	SS.sens.	SS spec.	SH sens	SH spec	accuracy
1 st cys only	0.721	0.898	0.956	0.865	0.874
2nd cys only	0.784	0.906	0.957	0.892	0.896
2nd cys + flanking	0.799	0.908	0.957	0.898	0.901
2nd cys+flanking+global	0.808	0.918	0.961	0.903	0.908
1st and 2nd cys+flanking+global	0.909	0.941	0.969	0.952	**0.948**

The way how flanking sequences was represented took a role too. As is shown in Table VIII, when represented by frequency, the accuracy is about 1.5 % higher when represented by the order they appear.

Table VIII: Influence of flanking sequence representations (with all other information the same)

	SS.sens.	SS spec.	SH sens	SH spec	accuracy
flanking by frequency	0.909	0.941	0.969	0.952	0.948
Flanking by order	0.864	0.936	0.968	0.930	0.932

3.7.3 Testing results given all other cysteines' oxidation states

To understand how the amino side of the cysteines influences the whole chain prediction accuracy, we extend the assumption to all the cysteines on the chain. This is an extreme assumption that tests the recognition power of SVM. As shown in Table IX, assuming all the other cysteines' oxidation states were known expect the one that is being predicted, the SVM prediction using the combination of pattern of bonding states of cysteines and the flanking sequences represented in frequency yielded an overall accuracy of 94.5%. This is not as high as the prediction accuracy we obtained earlier—the one with the assumption of only the first and second cysteines' on the amino terminus known. Adding the global information does not seem to improve the prediction result based purely on the oxidation pattern of cysteines.

Table IX: Prediction accuracy if oxidation states pattern of chain is known.

	Overall accuracy	SS-sensi	SS-spec	SH-sensi	SH-spec
pattern only	0.939	0.927	0.895	0.945	0.962
pattern+global	0.918	0.922	0.836	0.916	0.962
pattern+flanking seq in order (10aas)	0.937	0.939	0.878	0.937	0.969
pattern+flanking seq in frequency(10aa)	***0.945***	0.928	0.912	0.953	0.962

Notes**:** (Pattern refers to the oxidation pattern of all cysteine residues, including the ordinal number of the cysteine in consideration and the oxidation pattern of all other cysteines. "aa" refers to the flanking amino acids around cysteine in consideration. The number before "aa" means the window size of the flanking sequence used in the encoding. "global" refers to the global information vectors. "SS" refers to the oxidized cysteines and "SH" refers to the cysteine residues in the reduced form.)

Table X illustrates the comparison of prediction accuracy by assuming the amino terminus cysteines' oxidation states and by assuming the whole oxidation pattern known in the whole chain, categorized by number of cysteines on protein chains.

From Table X, it can be seen that knowledge of the amino terminus oxidation state (if the first or the second cysteine is the one to be predicted, then the oxidation state of that cysteine is not encoded), combined with the flanking sequence information and global information about the protein chain, we could obtain an accuracy of 94.8%. This is higher than the prediction model built by assuming all other cysteines' oxidation known (an overall accuracy of 94.5%).

Table X: Summary of Prediction accuracy on oxidation states of cysteines

number of Cys on chain	accuracy with N terminus cysteine known	accuracy with all other cysteines oxidation states known
1	0.95	1.000
2	0.949	0.884
3	0.943	0.849
4	0.957	0.932
5	0.944	0.938
6	0.938	0.948
7	0.931	0.959
8	0.961	0.948
9	0.963	0.907
10	0.963	0.988
11	0.934	0.977
12	0.923	0.990
14	1.000	1.000
15	0.667	1.000
16	0.930	1.000
17	0.941	0.971
18	0.972	0.972
22	1.000	1.000
23	0.913	1.000
24	1.000	1.000
30	1.000	0.833
34	0.971	0.706
35	0.400	0.971
overall	0.948	0.939

3.7.4 Testing results without any assumptions

To study what exactly make the N-terminus cysteines to have inference of oxidation states of other cysteines in the same chain, we designed SVM encodings to study if it was the overall N-terminus sequences, or the flanking sequences of N-terminus cysteines, or it has to be the hard coded oxidation states of N-terminus cysteines itself.

Table XI: Prediction with N-terminus sequences but no N-terminus cysteine's oxidation states is assumed

	accu.	SS sens.	SS spec.	SH sens	SH spec	MCC	SS Youden Index	SH Youden Index
5 flanking aas for 2 N term. Cys	0.78	0.55	0.75	0.90	0.79	0.49	0.30	0.69
10 flanking aas for 2 N term. cys	0.76	0.22	0.73	0.90	0.77	0.44	-0.05	63
seq. up to 10aas past 2nd N term. cys	0.74	0.44	0.71	0.90	0.75	0.39	0.15	0.61
no N-terminus info.	0.74	0.71	0.44	0.75	0.90	0.40	0.15	0.65

Notes: Except the flanking sequences in the first column, the encoding also included global sequence information and flanking sequences of the interested cysteines.

From Table XI, it can be seen that N-terminus sequence characteristics only are not as effective in giving the correct prediction of oxidation states of other cysteines as the case where cysteine oxidation states of the N-terminus cysteines were directly assumed. Although the N-terminus sequence characteristics help improve the accuracy, the overall prediction is not nearly as comparable to the prediction results obtained by direct inclusion of the oxidation states of cysteines on the N-terminus. Therefore, there might be some other unknown factors in play.

3.7.5 Summary on the predicting cysteines' oxidation states by SVM

In this research, models to predict the oxidation states of cysteines on protein chains were built by using SVM. We incorporated the flanking sequences of the cysteine in consideration, the general information about the protein chain (number of cysteines and number of amino acids on chain, as well as the amino acid compositions), flanking sequences around the interested cysteine, ordinal number of cysteine in consideration, the global amino acid composition of protein, as

well as assumed knowledge of amino terminus cysteines' oxidation states. The effects of different influencing factors were discussed.

It was found that oxidation states of amino terminus cysteines infer the oxidation states of all other cysteines on the same protein chain, when combined with other information such as flanking sequences and general information of the protein chain. We have obtained a model that predict the oxidation states of cysteines at 95% accuracy, which is 5% higher than the current record. Compared with the literature, this approach is a one-step prediction system and easier to implement and use. Moreover our system made fewer assumptions than the record literature.

Our study shows that the bonding state of cysteines on protein chains are closely related to the bonding pattern of the amino terminal cysteines' oxidation states. And to be specific, the first two cysteines on the amino terminal of protein shed more light case where the oxidation states of all cysteines except the one to be predicted is known. Other contributing factor such as the flanking sequence was also helpful when used properly. And also, by using only the flanking sequences of the N-terminus cysteines, it is not adequate to give a decent prediction of cysteine oxidation states. The N-terminus cysteines' oxidation states contribute the most to the accurate prediction of other cysteines.

3.8 Validation the conclusion from SVM models by SPX dataset

To further confirm what the amino terminus cysteines infer the oxidation states of other cysteines (section 2.3), another dataset (SPX dataset) was used for prediction. SPX is a newer and bigger compared the one we have been using . The SPX dataset is quite different from the one we have been using in the percentage of oxidized cysteines (85% for SPX and 35% for the

other dataset) and also in that SPX dataset excluded chains that are capable of forming inter-chain disulfide bonds. Table XII and XIII showed the testing results for the new dataset.

Table XII is the prediction accuracy when global information, flanking sequences and up to 2 cysteines on the N-terminus were included in the encoding to SVM. Because of the huge difference in population for oxidized and reduced cysteines, we had to manually adjust the relative weight (denoted as j in the table) of the positive and negative data in the dataset. When the ratio is adjusted so that the positive (oxidized)/negative(reduced) <= 0.5, the performance of the model stabilizes at an overall accuracy of 88%. From the testing results, we could see that the sensitivity and specificity for the oxidized cysteines are quite good. While the sensitivity and specificity for cysteines in reduced form are not as well. These are dragging the accuracy for the whole dataset down to about 88%. This might be caused by the insufficient representation of cysteines in the reduced form.

Table XII: Testing result on SPX with adjusted positive/negative ratio j

	SS. sens.	SS spec.	SH sens	SH spec	accu.	MCC	Youden-SS	Youden-SH
J=1	0.95	0.90	0.41	0.61	0.87	0.43	0.85	0.02
J=0.5	0.94	0.92	0.53	0.60	0.88	0.49	0.86	0.13
J=0.2	0.94	0.92	0.53	0.60	0.88	0.49	0.86	0.13
J=0.1	0.94	0.92	0.53	0.60	0.88	0.49	0.86	0.13

Notes: The included features were global information, flanking sequences and up to 2 cysteines on the N-terminus, which are the same as the encodings which has the best prediction accuracy for the other dataset.

Table XIII is the prediction accuracy in the chain level. From these result, we could see that when the number of cysteines on chain is small, the prediction accuracy for chains with even number of cysteines are significantly higher than those with odd number of cysteines. With the increase in the number of cysteines, this phenomenon becomes less obvious. This might be related to the fact that oxidized cysteines have to appear in pairs.

Table XIII: Prediction accuracy on oxidation states of cysteines for SPX dataset.

n: number of cysteines on chain	Accuracy
2	1.00
3	0.41
4	0.89
5	0.56
6	0.95
7	0.69
8	0.90
9	0.70
10	0.91
11	0.87
12	0.92
13	0.77
14	0.95
15	0.87
16	0.97
17	0.93
18	0.94
19	0.89
20	1.00
21	0.81
24	0.97
28	1.00
32	1.00
34	1.00
35	0.43
50	1.00
54	0.96

In an effort to improve the accuracy for protein chains with odd number of cysteines, a simple ranking mechanism is used to further process the prediction result from the SVM run. The

rule of the ranking is: for protein chains with odd number of cysteines, if the prediction from SVM resulted odd number of oxidized cysteines (which, in reality, is not a legitimate prediction since oxidized cysteines only show up in pairs), then the cysteine with the highest possibility (or the cysteine with the largest SVM output) is removed from the list of cysteines that are predicted to be oxidized.

By applying this simple ranking mechanism, an overall improvement of 0.4% was observed for the SPX dataset. Chain level wise, the accuracy for chains with three cysteines has been improved from 41% to 59%. For other chains with odd number of cysteines, no change in accuracy or a slight drop is observed. Table XIV shows the actual changes before and after ranking for the SPX dataset. A more complicated ranking system could improve the performance further for the SPX dataset.

Table XIV Accuracy for chains with odd number of cysteines after a simple ranking mechanism is applied for the SVM prediction

n: number of cysteines on chain	Accuracy (%)	afterOddRanking-accu.%
2	1	1.00
3	0.41	0.59
4	0.89	0.89
5	0.56	0.54
6	0.95	0.95
7	0.69	0.68
8	0.9	0.90
9	0.7	0.70
10	0.91	0.91
11	0.87	0.85
12	0.92	0.92
13	0.77	0.69
14	0.95	0.95
15	0.87	0.87
16	0.97	0.97
17	0.93	0.92
18	0.94	0.94
19	0.89	0.89
20	1	1.00
21	0.81	0.81
24	0.97	0.97
28	1	1.00
32	1	1.00
34	1	1.00
35	0.43	0.43
50	1	1.00
54	0.96	0.96

3.9 Conclusions

In this section, models to predict the oxidation states of cysteines were built using SVM. We found that the oxidation states of amino terminal cysteines infer the oxidation states of other cysteines in the same protein chain. By including the oxidations states of up to two cysteines on the amino terminus, accuracy of 95% is achieved. This is 5% higher than the current record. Satisfactory prediction results were also achieved with the newer and more inclusive SPX dataset,

especially for chains with higher number of cysteines. Compared with the literature, our approach is a one-step prediction system which is easier to implement and use. Moreover this system makes fewer and more manageable assumptions than the record literature. We incorporated the flanking sequences of the cysteine in consideration, the general information about the protein chain (number of cysteines and amino acids composition of the chain), flanking sequences around the interested cysteine, as well selective cysteine's oxidation states from the amino terminus of chain. The effects of different influencing factors were discussed.

CHAPTER 4 PREDICTING OXIDATION STATES OF CYSTEINES BY ASSOCIATIVE NEURAL NETWORKS (ASNN)

4.1 Artificial neural networks (ANNs)

Artificial neural networks (ANNs) are models of intelligent systems. They consist of large number of processing units (neurons) which collectively perform complex pattern matching tasks[74]. These mathematical neural networks are designed to simulate biological neural networks in the human brain (Figure 16).

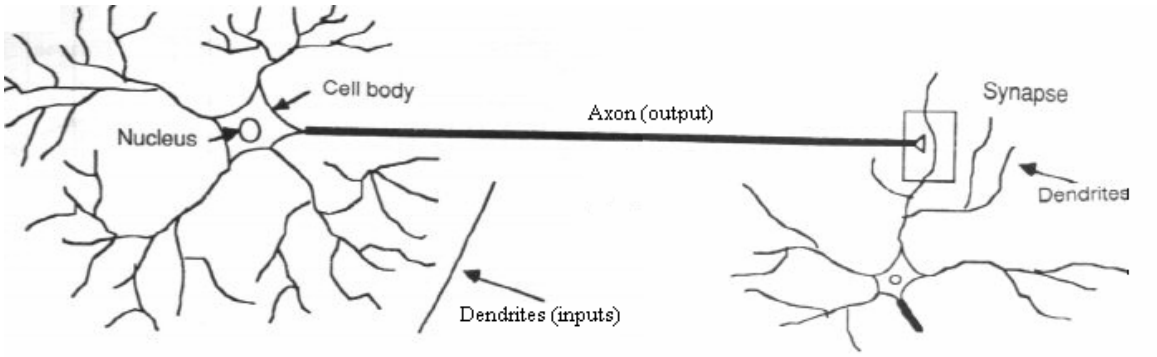


Figure 16: How the human brain learns[75]

As is shown in Figure 16, in the human brain, communication between neurons often involves an electrochemical process. Neurons collect signals through a host of fine structures called *dendrites* (signal input from other neurons). Upon receipt of signals from its dendrites, the neuron sends out spikes of electrical signal to other neurons through a long, thin stand known as an *axon (signal output to other neurons)*, which splits into thousands of branches. At the end of each of these branches, a structure called a *synapse* converts the signal from the axon into

electrical effects that either inhibit or excite the connected neurons. When a neuron receives excitatory input that is sufficiently large compared with its inhibitory input, it sends a spike of electrical response down its axon. Learning occurs by properly weighing the input signals collected by the dendrites and sending out reliable outputs signals.

When modeling artificial neurons, the complexity of real neurons is highly abstracted. As shown in Figure 17, an artificial neuron basically consist of *inputs* (like synapses), which are multiplied by *weights* (strength of the respective signals), and then computed by a mathematical function which determines the *activation* of the neuron. Another function (which may be the identity) computes the *output* of the artificial neuron (sometimes in dependence of a certain *threshold*). Artificial neural networks (ANNs) combine artificial neurons in order to process information. As is shown in Figure 18, ANN can be simply viewed as a black box. Input were feed in and by applying proper weights, output were computed in the black box. Figure 18 shows a single hidden layer in the black box. The most common neural network model is the so called multilayer perceptron (MLP)[76], where there are multiple hidden layers in the black box. This type of neural network is known as a supervised network because it requires a desired output in order to learn. The purpose of this type of network is to create a model that more accurately maps the input to the output using historical data so that the model can be used to produce the output when the desired output is unknown.

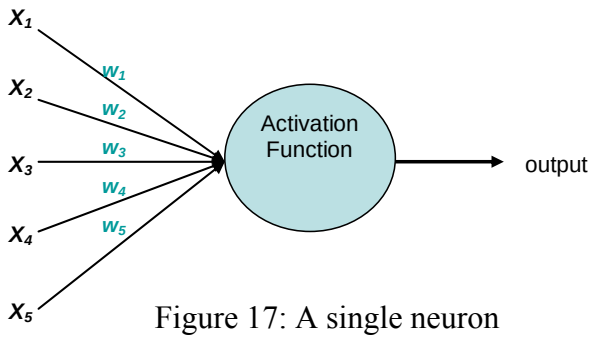


Figure 17: A single neuron

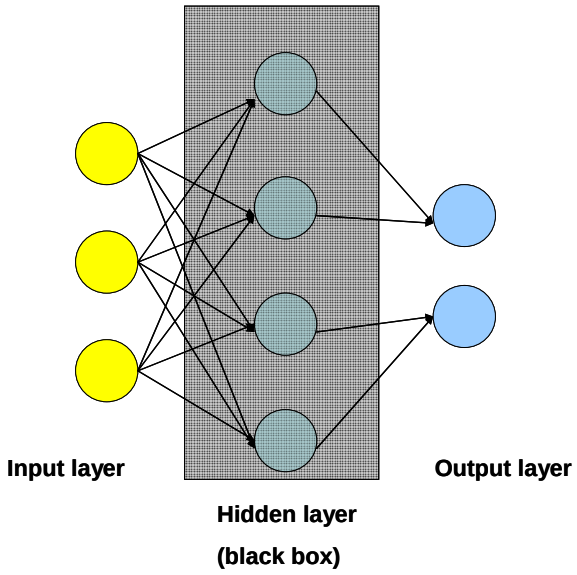


Figure 18: Artificial neural network layout.

Neural network is widely applied to business, economics, finance and engineering, where complicated models can be built to forecast market trend[77,78], to study the patterns of abnormal behaviors in business [79,80], and to solve various engineering problems[81,82,83,84]. The biggest advantage of ANNs is their ability to be used as an arbitrary function approximation mechanism which 'learns' from observed data and there is no need to assume an underlying data distribution such as what is usually done in statistical modeling. And ANNs are widely applicable to multivariate non-linear problems. However, one biggest disadvantage of ANNs is that the relations between the input variables and the output variables generated by ANNs are not developed by engineering judgment. It is solely the user's responsibility to tell if the prediction or simulation result is reasonable or not. And ANNs only cover the range where training samples

were covered. If the ANNs were not exposed to certain range of samples during training, then in no way the ANNs would give a trustworthy predictions when such input is present in testing.

4.2 Associative neural networks (ASNN)

Traditional multi-layer neural network is an approach that is "memoryless", which means after training and neural network weights are calculated from input patterns, the input data is discarded and no storage of any of those input information is stored in the system. On the other hand, k-nearest-neighbors [85] (KNN) and Parzen-window regression [86] can be classified as memory-based approaches, where the entire pool of training data are kept in memory and the prediction is based on the local approximation of the stored data. So in a sense, neural networks predict from a global point of view. The prediction result made by ANNs is based on the global characteristics extracted from all the data points in the training set. And KNN and Parzen-window regression are local methods [87]. In many applications, the global model can be insufficient since it may not describe universally well the entire state space. In some particular regions of space, a high bias may exists due to the limited size of training set. The multi-layer neural networks variance can also contribute to poor performance of ANNs, although the variance is reducible by using an ensemble of ANNs and taking the average of all networks.

ASNN is a combination of an ensemble of feed-forward neural networks and the k-nearest neighbors technique. It uses the correlation between ensemble responses as a measure of distance amid the analyzed cases for the nearest neighbors technique. This provides an improved prediction by the bias correction of the neural network ensemble. An associative neural network has a memory that can coincide with the training set. If new data becomes available, the network further improves its predictive ability and provides a reasonable approximation of the unknown function without a need to retrain the neural network ensemble. This feature of the method

dramatically improves its predictive ability over traditional neural networks and k-nearest neighbour techniques. Another important feature of ASNN is the possibility to interpret neural network results by analysis of correlations between data cases in the space of models.

To summarize, compared with conventional ANNs, ASNN explicitly corrects bias of neural network ensemble which will lead to improved prediction ability; And ASNN provides better modeling, because similarity in space of models makes it possible to interpret the ASNN results; And finally, ASNN is capable of fast and accurate extrapolation. New data are incorporated in the network without retraining its weights. However, it should be noted that the problem of bias of ANNs is not completely decreased in ASNN because even using a large number of ANNs, such networks can still fall in a local minimum and have considerable bias in its prediction.

The ASNN is implemented [88] as the following: Consider an ensemble of M neural networks, $[ANNE]_M$, with the number of neural networks in the ensemble to be M.

$$[ANNE]_M = \begin{bmatrix} ANN_1 \\ ANN_2 \\ ..... \\ ANN_M \end{bmatrix}$$

The prediction of a case x_i, i=1, …, N can be represented by a vector of output values $Z_i = \{Z_j^i\}_{j=1}^M$ where j=1, …, M is the index of the network within the ensemble.

$$x_i \times [ANNE]_M = Z_i = \begin{bmatrix} Z_1^i \\ Z_2^i \\ ... \\ Z_M^i \end{bmatrix}$$

And then, a simple average $\bar{Z} = \frac{1}{M} \sum_{j=1,M} Z_j^i$ was used. (Depending on application and implementation, other ways except taking a simple average to reduce the bias caused by the global model built from limited training data is possible.)

In some regions of data space, the ANNs ensemble predictions $\bar{Z}_i$ could have obvious bias from its true value. After obtaining $\bar{Z}_i$, in order to improve the performance of the ANNs, the ensemble prediction $\bar{Z}_i$ is corrected according to the following equation:

$\bar{Z}_i' = \bar{Z}_i + \frac{1}{k} \sum_{j \in N_k(x)} (y_j - \bar{Z}_j)$, where y_i are the true value for input x_i , $N_k(x)$ is the collection of the k nearest neighbors of x among the input vectors in the training set $\{x_i\}_{i=1}^N$ determined using Spear-man non-parametric rank correlation coefficient r_{ij}. (calculated as $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$ where d_i is the difference between each rank of corresponding values and n is the number of pairs of values. It is simply a special case of the Pearson product-moment coefficient in which the data are converted to rankings before calculating the coefficient).

Since the final prediction result of the new data is according to the prototypes of the analyzed training set which is located in the memory of the neural network, this proposed approach was named as Associative Neural Network. One big advantage of ASNN is that in case some new data become available, then these data would be loaded to the memory of the ASNN. This provides users with the ability to improve the neural network predictions without a need to retrain the weights[88].

The ASNN is solidly supported by neurophysiological background. Theories of brain coding suggest the importance of temporal coding for information processing in brain[89, 90].

Analysis of speed of processing and particular features in the visual systems indicates the importance of the rank coding for information processing[91].

ASNN has been successfully applied to the prediction of lipophilicity of chemical compounds[92] and classification of University of California, Irvine (UCI)[88] benchmark data and satellite data[93] and showed improved prediction results over traditional neural network both in regression and in classification studies.

4.3 Motivation

The last effort seen on the prediction the oxidation states of cysteines using neural networks was seen in 2002, when Martelli et al[51], implemented hybrid system that combines a neural network and a hidden Markov model (hidden neural network). An overall accuracy of 88% was obtained. This indicates to us that artificial neural network is able to successfully capture the characteristics of oxidized and reduced cysteines with very good accuracy. Since then, new progress has been made in the area of neural networks such as Associative Neural Network (ASNN)[94]. ASNN uses a combination of an ensemble of feed-forward neural networks and the k-nearest neighbor technique. It takes the correlation between ensemble responses as a measure of distance among the analyzed cases for the nearest neighbor algorithm, which has been proven to be able to provide an improved prediction by the bias correction of the neural network ensemble[95]. In our case, our previous results have shown that the amino terminus of the protein chain is highly correlated with the bonding states of the other cysteines in the same chain. If the amino terminus of the protein chains shares a set of characteristics, and if such correlation is discernable by neural networks, it is promising to explore such correlation by using a powerful machine learning technique such as ASNN. So here we propose using ASNN to capture the characteristics of oxidized/reduced cysteines on proteins.

4.4 Prediction of oxidation states of cysteines by ASNN

Oxidation states of cysteines were predicted using ASNN. The details such as software used, encoding of the sequence features as well as the prediction results are discussed in the following sections.

4.4.1 Software and dataset

ASNN software has been developed by Virtual Computational Chemistry Lab (VCCLAB)[96] and is downloadable from the VCCLAB website. Here we used the same datasets and the same evaluation on the prediction accuracy as in Chapter 2.

4.4.2 Encoding

The format of encoding to ASNN is X_1 X_2 ……. X_n **Y**, with X_1 X_2 ……. X_n the input features and Y the target value for the series of Xs. Similar to the SVM encoding, the target value is set to 1 for intra-chain oxidized cysteines and -1 for non intra chain bonded cysteines. The X variables represent the relevant features selected for the prediction. Those include the global information about the protein chain (chain length, number of cysteines on chain as well as the amino acid composition on the whole chain), the flanking sequences around interested cysteines (both in frequency and in order) and the information on the N-terminus (the oxidation states of cysteines on the N-terminus or partial amino acid sequence on N-terminus). Figure 19 shows the layout of the encoding as input to ASNN.

Features on global info	Ordinal number of cysteine in consideration	Features for flanking sequences	N-terminus information	Target value 1/-1

Figure 19: Encoding as input to ASNN

4.4.3 Results and Discussion

We evaluated the cases when the N-terminus cysteines oxidation states are known, together with other influencing factors such as flanking sequences, global protein information (amino acid composition, chain length and number of cysteines etc). Table XIV is the summary of results when using different encodings.

Table XV Prediction accuracy from ASNN

Included variables	SS sens.	SS spec.	SH sens	SH spec	accuracy	MCC	SS-Yuden	Sh-Yuden
1 st cys only	0.637	0.652	0.817	0.807	0.754	0.457	0.289	0.624
1,2 cys only	0.647	0.671	0.830	0.814	0.766	0.481	0.318	0.644
global+1,2cys	0.647	0.671	0.830	0.814	0.766	0.481	0.318	0.644
Flanking freq+1,2cys	0.646	0.671	0.830	0.813	0.765	0.480	0.317	0.643
1st and 2nd cys+flanking+global	0.647	0.662	0.823	0.812	0.761	0.459	0.309	0.635

Table XIV is the prediction performance of ASNN. Overall, the performance of ASNN is inferior to what we have observed in SVM for similar encodings. The system does not seem to be sensitive enough when adding more feature variables to the encoding.

4.4.4 Summary on prediction of oxidation states of cysteines by ASNN

A side by side ASNN was run on the similar encodings as SVM. However, unsatisfactory result was obtained. As far as the different encodings for each ASNN run, we find that adding more features does not change much of the classification of cysteines and the performance remains poor even when all the relevant features were added to the encoding.

4.5 Conclusion

In this section, models to predict the oxidation states of cysteines were built using ASNN. Features used in the ASNN are very similar to what we have used for SVM. Prediction results indicate that ASNN does not have the same satisfactory prediction performance as SVM on prediction of oxidation states of cysteines on protein.

CHAPTER 5 PREDICTING DISULFIDE CONNECTIVITY BY CALCULATING THE INTERACTIONS BETWEEN FLANKING SEQUENCES

5.1 Related works in literature on prediction of disulfide bridges

The past research efforts in prediction of disulfide connectivity can be classified into two categories [97]: i. pattern-wise prediction, which focuses on the connectivity pattern of the protein chains [97,98] and ii. Pair-wise prediction[99,100,101] which concerns more about the probability of each possible pair of cysteines. Up to date, pattern-wise prediction has reached 70% accuracy [98, 102] and pair wise prediction gave 89% accuracy [101].

The method used for prediction of disulfide bond connectivity included neural network[25,28,55], SVM[54,97,98] , fuzzy SVM [103] and use of multilevel of these machine learning techniques. The sequence features used for encoding for these machine learning techniques comprised one or more of the following factors: (1) flanking sequences such as appearing frequencies of nearest neighbor residues of bonded cysteines[99], secondary structure[9], solvent accessibility[9], (2) global spacing of bonded cysteine pairs such as position of bonded cysteines[116], sequence length[9], and separation profile of bonded cysteines[97,104]. (3) the characteristics of the protein chain itself (these include the chain length, number of disulfide bonds as well as the amino acid composition of the protein chain). So far, for pattern-wise predictions, a two level framework that incorporated both pattern-wise and pair-wise encodings to SVM has reached pattern-wise accuracy of 70%. The best result in pair-wise predictions used secondary structure information and diresidue frequencies and has the best accuracy of 89%. A summary of the efforts in literature is presented in table XV.

Table XVI: Past efforts on the prediction of disulfide bond connectivity

year	author	method	information used	accuracy
2001	Fariselli et al[99]	graph matching	residue contact potentials	0.56*
2004	Vullo et al[116]	neural networks	connectivity pattern	0.45
2005	Tsai et al[97]	two step SVM	sequence profile and distance between oxidized cysteine	0.79
2005	Ferre et al[41]	neural networks	known secondary structure and diresidue frequencies	0.89**
2006	Cheng et al[28]	2D neural networks	flanking sequence, PSSP, distance, solubility of partial sequences	0.56

*: for chains with 2 disulfide bonds only. Accuracy falls with number of disulfide bonds increase in the chain.
** With the assumption that secondary structure of protein is known

In this research, a different approach was proposed and tested. Multiple sequence alignments were first conducted on a number of proteins to confirm the existence of the conserved binding motifs on protein followed by incorporating these features into calculation of flanking sequence interaction and SVM encodings. The multiple sequence alignments were conducted first to confirm the existence of the binding motifs to PDIs on protein chain. Then two different prediction approaches in evaluating the disulfide formation probability of cysteine pairs. For the first approach, we first calculate the interaction between the flanking amino acids between cysteine pairs. Then the interaction is further adjusted according to whether the binding motif was located and by checking the distance between cysteine pairs. Based on the adjusted interactions for all possible pairs, a maximum weighted graph matching algorithm was applied to the completely connected graph constructed from all possible pairs of cysteines with the weight

of each edge to be the adjusted interaction. The prediction of disulfide bridges via calculated interaction between flanking sequences is described in detail in this chapter (chapter 4)

In the second approach, the binding motif, flanking sequences around both cysteines, global information of the protein chain and the distance between cysteine pairs were incorporated into the SVM encoding. SVMs assign the probability of disulfide formation to each possible pair of cysteines according to the models built and this probability was used as the weight for the connection between the cysteine pair. Then maximum weighted graph matching algorithm is applied to the graph obtained and top n connections in the resulted graph are selected to be disulfide bridges. Chapter 6 covers the details of the SVM prediction approach based on the binding motif to PDIs.

In this chapter and the next, we describe in greater details these two approaches and evaluate the performance. Possible factors that could potentially influence the prediction are also discussed.

5.2 Multiple sequence alignment on the protein families with disulfide bonds

Based on the experimental observations for PDIp [38], a tyrosyl side chain with a free phenolic hydroxyl group and without negative charge in vicinity is crucial for the formation of disulfide bond. Experiments indicated that if this binding motif is altered or blocked, the binding affinity of the PDIp and protein chain is greatly reduced; hence disulfide bonds would be formed in a much less efficient manner. In this research, we try to identify the binding region of protein chains to PDIs by multiple sequence alignment on proteins that share significant similarity retrieved from the Data Bank (PDB) database [105] by Blastp [106]. The identified anchor region was searched against the amino acid sequences between interested cysteine pairs and also encoded as input to SVM.

Multiple sequence alignments using clustalX [107] were performed on the retrieved sequences from PDB. It was found that indeed there are certain regions on many proteins that were highly preserved and with the characteristics of the binding motif identified for PDIp [38]. Figure 20 is a snapshot of the multiple sequence alignment on the first 250 entries returned by blastp for 4PTI. From the alignment it can be seen that the highly preserved tyrosine residues (shown in dark blue) between the cysteine pairs. Another example is shown in Figure 21, which is the multiple sequence alignment for the 1HOE family. Again we observe preserved aromatic amino acids (W and Y) in between cysteines pairs that form disulfide bonds. There are many more cases like this and from these multiple sequence alignment results, it was believed that it is highly possible to use this information as a new input feature to train SVM to recognize disulfide forming pairs in an efficient manner.

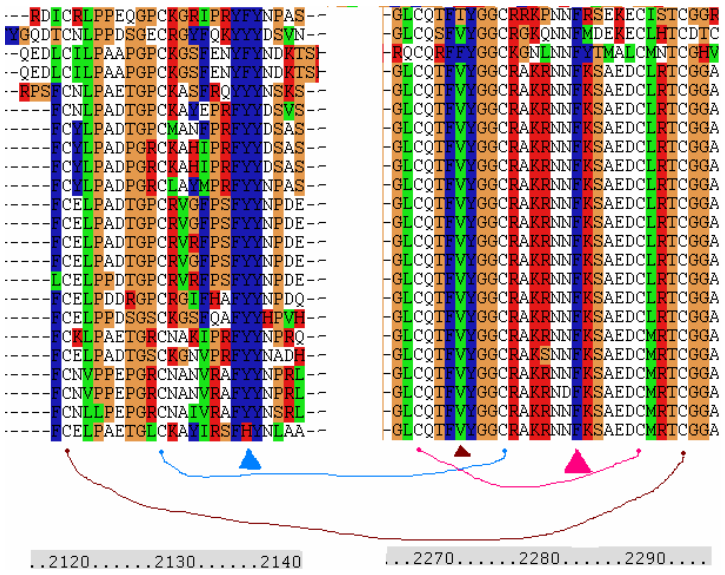


Figure 20: Binding motif found in multiple sequence alignment for 4PTI (highly reserved amino acids shown in blue)

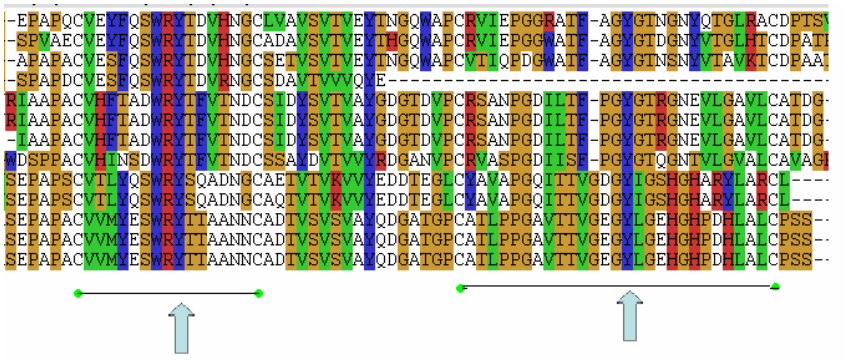


Figure 21: Another example - identified binding sites for 1HOE family.

In the original biological findings [38], the binding sites were restricted to tyrosyl side chain with a free phenolic hydroxyl group (W and Y) and without negative charge (primarily amino acid D and E) around. Since in our observation of the multiple sequence alignments, this is not a very prominent restriction and we did find negative charges around those identified binding sites between many actual disulfide bonds, we have relaxed the restrictions by including W, F, Y and C (cysteine) and allowed neighboring residues to be amino acids with negative charges.

5.3 Identifying disulfide bonds by calculating the interactions between flanking sequences

The stability of the disulfide bonds is a direct indication on the susceptibility of the protein to unfavorable changes in its surrounding environment such as radioactive damages and chemical attacks. It is crucial for the protein to maintain the correct folding and secondary structures[108]. The interaction between the flanking sequences of the paired cysteines is one of the

most direct factors that contribute to the stability of the disulfide bonds. The first research effort[99] on the prediction of bonding probability of cysteine pairs adopted this intuitive approach. The pair-wise amino acid interactions (contact potentials) used in literature were derived by Mirny and Shakhnovich[109] and these contact potentials were based on optimization of protein folding and threading. However, these contact potential values were obtained by pooling the observed number of contacts over all proteins, which is spread out among those proteins that have contacts and those that do not. And this incorrect accounting for composition can be readily corrected by only accounting for the structures that actually have such contacts. Skolnick et al[110] derived the so called composition corrected, quasi chemical pair wise potentials, which is believed to be a better description of pair wise contact potential of protein folding and threading. This is where we differ from the literature effort in terms of interaction calculation. Also we incorporated other factors such as distance between the cysteine pair in consideration and whether or not the binding motif to PDIs exists between the cysteine pairs because we believe doing so would give a better description of the disulfide bond bridges on protein.

Therefore, we computed the flanking sequence interactions using the composition corrected quasi chemical pair wise potentials. The interaction value was further adjusted by the distance factor and the anchor factor. Finally, the maximum weighted matching algorithm was applied and the top n cysteine pairs were obtained (n is the number of disulfide bonds on the protein chain). We will cover each step in a greater detail in the following sections.

5.3.1 Evaluate the interaction between cysteine pairs

The interaction between flanking sequences of cysteine i and cysteine j is computed by $\omega_{i,j} = \sum_{k \in S_i} \sum_{l \in S_j} \frac{1}{\alpha(d)} U(k,l)$, $\omega_{i,,j}$ is the interaction between the flanking sequences of cysteine i and

cysteine j. And S_i and S_j are the sets containing the nearest neighboring residues of cysteine i an cysteine j. α is the function of distance d, which is the linear distance from which the i^{th} residue is from the exact matching residue on the flanking sequence of pairing cysteine. Protein-Specific Pair Potentials based on weak Sequence Fragment similarity [110] was used as the interaction U(k,l) between pairs of amino acids. By using the protein-specific pair potentials based on sequence similarity to calculate the interaction, it enables us to obtain the interaction potential which already been adjusted by various compositions of different proteins and giving us pair potentials that are more specific than that derived from other approaches such as those assessed by gapless threading. The pair-wise interaction potentials for each pair of amino acids can be found in Appendix II. We evaluated the cases where $\alpha(d) = d$ and $\alpha(d)= d^2$ which means the interaction potential diminish in the order of $\frac{1}{d}$ and $\frac{1}{d^2}$, with d the relative distance from which the exact matching amino acid is located.

5.3.2 Adjustment with anchor factor and distance factor

Based on the experimental observations for PDIp[38] , a tyrosyl side chain with a free phenolic hydroxyl group and without negative charge in vicinity is crucial for the formation of disulfide bond. Experiments indicated that if this binding motif is altered or blocked, the binding affinity of the PDIp and protein chain is greatly reduced; hence disulfide bonds are formed in a much less efficient manner. In previous section, we have confirmed the existence of the binding motifs (anchor region) between the cysteine pairs through multiple sequence alignments.

In the following calculation and SVM encoding as well as discussion, *anchor* refers to the binding motif for the substrate protein to the enzyme (PDI). Although Ruddock et al. restricted the binding sites as Tyrosine (Y) and Tryptophane (W) without Aspartic acid (E) and

Glutamic acid (D) in the neighborhood. In our program, we have relaxed the restrictions to Y, W, F and C including those with E and D as neighboring amino acids. For the *anchor region*, based on the observation of multiple sequence alignments, we have used the region between the cysteine pairs and a window size for 9 amino acids. For the flanking sequences, a window size of 5 is used for sequences before and after the cysteine in consideration. *Distance between cysteine pairs* refers to how far the two cysteine residues are located. For example, if we have CXXXC, then the distance between the two cysteines is 4. Also in our program, we used a window size of nine amino acids in search of the correct anchors.

The interaction calculated from pair-wise interaction potentials is further adjusted by two factors: (1) whether or not an anchor is present between the considered cysteine pair (2) the distance between the cysteine pairs. The reasons for the above adjustment are that (1): based on experimental results[39], blocking the anchoring site of the PDIs will greatly reduce the efficiency of disulfide bond formation. (2): A complete statistical analysis on the distance between pairs of cysteines that form disulfide bonds indicates there is a favorable range in distance for disulfide formation. As is shown in Table XVI, for those disulfide bonds forming cysteine pairs, there is a clear preference in distance: 85% of disulfide forming cysteine pairs are within 60 amino acids apart, while this percentage is only 51% for the non-SS forming pairs.

Table XVII: Statistics on distance between cysteine pairs for disulfide bonded cysteines and non-bonded cysteine pairs

Distance between cystein pairs	2~32 as	2~50AAs	2~60AAs	2~80AAs
non-SS cys pairs	35%	46%	51%	59%
SS cys pairs	68%	80%	85%	88%

The adjustment was done according to the following equation:

$$adjusted - \omega_{i,j} = \eta \bullet \mu \bullet \omega_{i,j}$$

where η is the anchor coefficient, η=1 for those cysteine pairs with anchors and η=0.6 for those without anchors. μ is the distance coefficient, μ is 0.85 if the distance between the cysteine pairs is within 60 amino acids and μ=0.5 if the distance is above 60 amino acids.

5.3.3 Maximized graph matching for connectivity

Assuming the number of disulfide bonds is *n* and number of cysteines on the protein chain is *s*, by calculating the adjusted flanking sequence interaction between cysteine pairs in consideration; we can obtain a completely connected graph with s nodes. The weight associated with each edge in this graph is the calculated flanking sequence interaction adjusted with anchor factor and distance factor. The problem is to find the connectivity pattern within the graph with *n* edges that connect the cysteines pairs which are most probable to form disulfide bonds. Finding the maximum weight matched graph with $\lfloor s/2 \rfloor$ edges is solvable by Edmond's maximum weight matching algorithm[111] in $O(V^4)$ time. Also an $O(V^3)$ implementation using Gabow's[112] improved algorithm is available from MATHPROG's website[113].

How to select the n edges from the completely connected graph with s nodes is more complicated than simply picking n edges from maximum weight matched graph with $\lfloor s/2 \rfloor$ edges since doing so will not guarantee the picked n edges are the global optimum. Trying all possible subsets of 2n nodes from the overall s nodes in the graph and calculating which subset gave the maximized weights would guarantee optimal solution but is not practical in computation time.

Here Cheng et al's[28] approximation heuristics is adopted. This approximation heuristics is claimed to give very good result in the case where the total number of disulfide bonds are known.

The heuristics is as follows: If s is even, Gabow algorithm is applied to the s nodes and then prune down the final result by removing the s/2-n edges with lowest probabilities, from the final set of s/2 edges. If s is odd, the same strategy as above is applied s times, each time removing one of the cysteines. Then the matching with top n edges that has the maximum weights is selected. The process is illustrated in Figure 22.

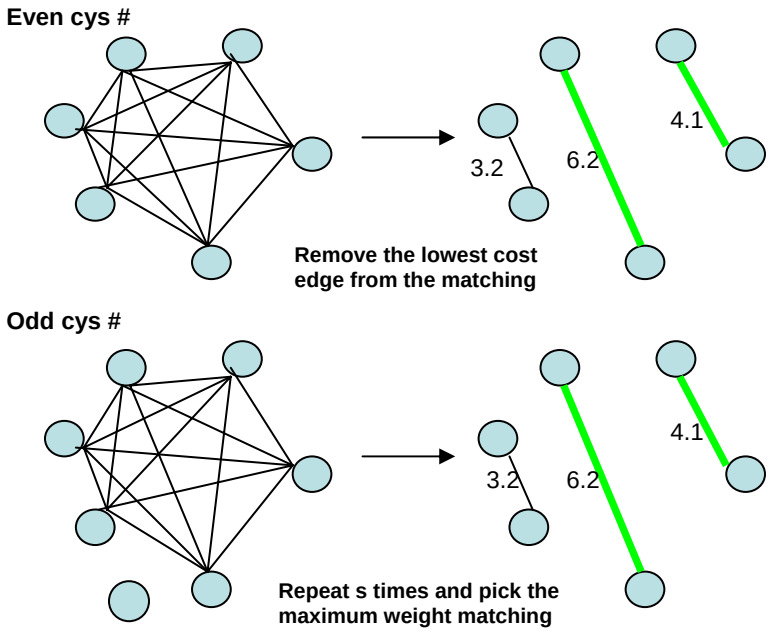


Figure 22: applying the maximum weight matching algorithm to the cysteines interactions.

5.4 Dataset

The SPX dataset described in section 2.3.4 was used for the testing with the assumption of known number of disulfide bonds in the protein chain. The dataset was divided into 10 subsets

for the purpose of cross validation. The statistics of this dataset on the number of disulfide bonds and their distribution is shown in the following table (Table XVII).

Table XVIII: Statistics on the disulfide bonds for SPX dataset

n: number of SS on chain	Number of chains with n cysteines
1	398
2	211
3	219
4	88
5	49
6	19
7	10
8	10
9	2
10	3
12	2
14	1
16	2
17	2
25	1
26	1
Total # of SS	1018

5.5 Measurement of predictive accuracy

The accuracy of the predication is calculated as the following:

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

where P denotes the number of cysteines pairs that forms disulfide bonds and N denotes the number of cysteines pairs that do not form disulfide bond. True positive (TP) denotes the number of cysteine pairs that form disulfide bonds in the protein chain and are predicted as such.

True negative (TN) is the number of cysteine pairs that are correctly predicted not to form a disulfide bond.

Sensitivity and Specificity of the predication are defined as the following:

$$Q_c = \frac{TP}{P}$$

$$Q_{nc} = \frac{TN}{N}$$

Youden index is calculated according to

Youden's index = *sensitivity* + *specificity* - 1

The Matthews correlation coefficient (MCC) is calculated as:

$$\text{MCC} = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

5.6 Result and Discussion

Table XIX shows the prediction result for SPX data set with the adjusted interaction approach. It can be seen that using this approach, without the maximum weighted graph matching, the accuracies are far below 80%, mainly because of the large number of false positives in the prediction. With the maximum weighted graph matching we are able to eliminate 89% of the non-bonded cysteine pairs from consideration, but at the same time, we are also

removing majority of the actual disulfide-bonded pairs, making the prediction essentially useless since most of the actually bonded pairs were misclassified. Overall prediction is unsatisfactory. The only information obtained from this result probably is that the interaction between flanking sequence of cysteine pairs decreases closer to linear to distance *d* between the cysteine pair than to the d^2.

.

Table XIX: Prediction result on SPX for the adjusted interaction approach (interaction cut off set as 0.8 for the result without maximum weighted graph matching)

		sensitivity	specificity	accuracy	MCC	Youden IDX
without maximum weighted graph matching	a(d)=d	0.39	0.78	0.73	0.14	0.17
	$a(d)=d^2$	0.38	0.70	0.66	0.06	0.08
with maximum weighted graph matching	a(d)=d	0.28	0.89	0.81	0.17	0.17
	$a(d)=d^2$	0.23	0.88	0.80	0.11	0.11
	literature[28,*]	0.55	0.24	N/A	N/A	N/A

*The referred literature did not give all the accuracy measures given above. Those entries were marked with N/A in Table XIX.

5.7 Summary

In this section, predicting the disulfide connectivity according to the calculated interaction between flanking sequences of cysteine pairs is attempted. The interaction is calculated from the contact potential between the flanking sequences of cysteine pairs and is further adjusted by the distance between the cysteine pair and whether anchor was observed between the cysteine pair. The prediction result is unsatisfactory. There might be some factors

other than the flanking sequence contact potential and anchor to PDIs that account for the formation of disulfide bonds. It also might be that this way of calculating interaction of flanking sequences could not describe the extent of interaction. In the next section, we try to incorporate some other factors such as global information about the protein chain into SVM encodings and see if the prediction performance can be improved.

CHAPTER 6 PREDICTING DISULFIDE CONNECTIVITY BY IDENTIFYING BINDING MOTIFS OF PROTEIN TO PDIs

6.1 Influencing factors other than flanking sequence interactions

In the previous chapter, predicting disulfide connectivity by calculating the adjusted interaction between flanking sequences of cysteine pairs was attempted. The prediction result is far from satisfactory. Important factors such as the global information of the protein chain, amino acid compositions of the flanking sequences etc were not counted for in the previous calculations. In this chapter, we try to incorporate these factors together with the binding motif into SVM encodings and aim to improve the prediction performance.

6.2 Predicting disulfide bonds with SVM models without considering the interaction

Three descriptors were used to encode each possible cysteine pairs on the protein chain as the input to SVM: (1) Flanking sequence around the cysteines considered; (2) Anchor region sequence profiles; (3) the linear distance between the two cysteine pair. (4) the global information about the protein chain. SVM-*light* [114] was used for the prediction.

For the encoding, the representation for amino acids computed by Meiler et al. [115] (see APPENDIX I) was used, which has accounted for the physical-chemical properties of the amino acids. This five-parameter representation describes amino acids in terms of steric parameter, polarizability, volume, hydrophobicity and iso-electric point. The number of occurrence of each amino acid in the window of flanking sequence of cysteines and the anchors (W, F, Y, C) in the anchor region were recorded and used in the encoding. The feature values used for amino

acid representations in the anchor region were incremented by 100 to avoid overlapping with the representations of the same amino acids in the flanking sequence region. It was tested that as long as the encoding is consistent during the training and testing, incrementing the feature value of amino acids in the anchor region by any reasonable positive integer would not change the final result.

6.3 Data set

In order to assess the performance of our model in predicting the connectivity pattern of disulfide bonds, we have used two data sets: SP39 and SPX (courtesy of Dr. J Cheng). The compilation of SP39 was carried out by Vullo and his colleagues [116] and the compilation of SPX was done by Cheng et al[28] were described in detail in section 3.2.4. Overall there are 1018 protein chains in SPX ranging from one to 26 disulfide bonds. For cross validation, the SPX data set was divided into 10 subgroups and the homology between any subgroups is set to be less than 30%. The SP39 was extracted from the Swiss-Prot database [117] release no. 39 (October 2000) and consists of 446 protein chains with the number of disulfide bridges range from 2 to 5 on each protein. For cross validation, the SP39 dataset was divided into four sub-groups and the homology between any two of them is less than 30%. The number of disulfide bonds and their distribution are listed in Table XIX and XX, respectively.

Table XX: sequence information in dataset SP39

Number of dinsulfide bond in protein chain	Number of sequences in dataset
B=2	156
B=3	146
B=4	99
B=5	45
B=2~5	446

Table XXI: statistics on disulfide bonds in SPX dataset

N: number of SS on chain	Number of chains with N disulfide bonds
1	398
2	211
3	219
4	88
5	49
6	19
7	10
8	10
9	2
10	3
12	2
14	1
16	2
17	2
25	1
26	1
total	1018

6.4 Result and discussion

In the following sections, predictions results when the local sequence characteristic, protein global characteristic as well as the characteristics from the binding motif region are discussed.

6.4.1 Prediction result with SPX dataset

Table XXII shows the prediction performance of SVM on SPX dataset. We could see that by incorporating the anchor region profile, the prediction accuracy is improved over the testing without the anchor region profile, both before and after maximum weighted matching algorithm

is applied. However, by applying the maximum weighted graph algorithm, the sensitivity of the prediction is significantly lowered; due to the fact the maximum weighted graph algorithm is removing the correctly predicted connections as it removes the non-disulfide bonded cysteine pairs.

One thing to notice from the prediction result is that, if we look at the accuracy before the maximum weighted graph matching algorithm is applied, we are able to remove 89% of the non-disulfide bonded pairs and keep a good number of bonded pairs for further weight matching. Therefore, using anchor region profile might be a nice and convenient filter for the disulfide connectivity in proteins. To have higher overall prediction accuracy after the maximized graph matching algorithm, some other adjustment on the SVM output is necessary. We also tested this anchor filtering effect on the SP39 dataset and the result is presented in the next section.

Table XXII. Predictive performance of SVMs on SPX dataset.

	encoded features	Sensitivity	specificity	accuracy	MCC	Youden Index
without maximum weighted graph matching	flanking+distance +global	0.50	0.88	0.83	0.34	0.38
	anchor+flanking+ distance+global	0.58	0.89	0.85	0.41	0.47
with maximum weighted graph matching	flanking+distance +global	0.24	0.89	0.80	0.12	0.13
	anchor+flanking+ distance+global	0.25	0.89	0.80	0.13	0.14
	Literature[28, *]	0.55	0.24	N/A	N/A	N/A

*The referred literature did not give all the accuracy measures given above. Therefore some of the entries were marked as N/A.

6.4.2 Prediction results with SP39 dataset.

We also tested our SVM encoding with the older and smaller dataset SP39. SP39 is a dataset with restricted number of disulfide bonds (up to 5). The oxidation states of the cysteines are assumed to be known. Table XXII is the prediction result.

Table XXIII: Disulfide connectivity prediction accuracy with SP39 using SVM

	Encoded factors	sensitivity	specificity	accuracy	MCC	Youden Index
without maximum weight matching	flanking+ distance +global	0.56	0.90	0.84	0.44	0.46
	anchor+ flanking+distance +global	0.63	0.91	0.86	0.51	0.54
with maximum weight matching	flanking+ distance+ global	0.35	0.86	0.77	0.21	0.21
	anchor+flanking+ distance+global	0.44	0.88	0.80	0.32	0.32
	Literature[41]	0.59	0.95	0.89	0.57	0.54

By incorporating the flanking sequence, anchors, global information about the protein and the linear distance between the cysteine pairs in the SVM encoding, we were able to obtain an overall accuracy of 80% on the SP39 dataset after the weighted matching algorithm. It is 9% lower than the record literature.

6.4.3 Screening without any assumptions

The previous testing are all based on certain assumptions such as the knowledge of the number of oxidized cysteines (for the purpose of limiting the number cysteine pairs that form disulfide bonds to the actual number of disulfide bonds), the actual oxidation states of cysteines

or the secondary structure of the protein (since literature indicates that oxidized cysteines prefer certain secondary structures). However, this assumed knowledge is not always readily available. Therefore, in this section, we tested on SP39 dataset to see how well the encodings will filter out the non-bonded cysteine pairs without the knowledge of the oxidation states and number of disulfide bonds on the protein chain.

Table XXIII shows the pair-wise prediction result of the SVM run without the application of maximum weighted graph matching. For chains with 2 to 5 disulfide bonds, an overall accuracy of 92% was obtained. From the result we can see that without any pre-knowledge of the protein chain except the sequence information, this SVM run can eliminate 92% of the non-disulfide bonded pairs from the overwhelmingly large pool of possible cysteine pairs and keeping 85% of the bounded cysteine pairs for further processing.

Table XXIV: Prediction result without pre-assumption of oxidation states of cysteines and without the weighted graph matching algorithm

Encoded features	sensitivity	specificity	accuracy	MCC	Yuden IDX
flanking+distance+global	0.81	0.80	0.80	0.15	0.61
anchor+flanking+distance+global	0.85	0.92	0.91	0.19	0.77

6.5 Conclusion

By observing the preserved amino acids between cysteine pairs in multiple sequence alignments of protein chains, we are able to confirm the anchor regions identified by biochemical experiments. Our predictions in this part has involved the profile in the anchor region, the flanking sequences of cysteine pair, together with the flanking sequences, global protein information and linear distance between the cysteine pairs.

We attempted to find disulfide connectivity by calculating the interaction between the flanking sequences of cysteine pairs, the value of which was further adjusted by the coefficients related to the existence of identified anchors in between the cysteine pairs and also by the linear distance between the cysteine pairs. Maximized weighted graph matching algorithm was applied and performance of the testing evaluated. It was found that using this approach, we can eliminate 89% of the non-bonded cysteine pairs from consideration, but at the same time, we are also incorrectly classifying majority of the disulfide bonded cysteine pairs. Overall prediction is unsatisfactory with this approach.

We also attempted to predict the disulfide connectivity by using SVM, taking anchor region profiles, distance between cysteine pairs, global information of the protein chain and the flanking sequences around the cysteine pairs. The prediction results indicated the advantage of using the anchor profile. The performance of our SVM model is 81% with knowledge of number of disulfide bonds on the protein chain for SPX dataset and after the weighted graph matching algorithms are applied, 80% for SP39 with the knowledge of oxidation states of cysteines. The accuracy and sensitivity for both dataset are lower than literature record and need further improvement.

The above discussions are based on the assumption that number of disulfide bonds in the protein chain is known beforehand or oxidation states of cysteines known. However, this information is not always available and accurate prediction is not guaranteed. Without the pre-knowledge of number of disulfide bonds, it was found that the SVM model which incorporated the anchor region information is able to act as a pre-screen for possible disulfide bond forming pairs. It eliminates 92% of the non-disulfide bonded pairs from the overwhelmingly large pool of possible cysteine pairs and keeping 85% of the bounded cysteine pairs for further processing.

CHAPTER 7 CONCLUSION AND FUTURE WORKS

7.1 Conclusion of this research effort

In this research, oxidation states of cysteines and the correct pairings of cysteines to form disulfide bonds on protein chains were studied. Models to predict the oxidation states of cysteines were developed with machine learning techniques such as Support Vector Machines (SVMs) and Associative Neural Networks (ASNN). Result indicates that the oxidation states of amino terminal cysteines infer the oxidation states of other cysteines in the same protein chain. By incorporating the oxidation states of N-terminus cysteines, flanking sequences of cysteines and global information on the protein chain (number of cysteines, length of the chain and amino acids composition of the chain etc.) into SVM encoding, accuracy of 95% is achieved. This is 5% higher than the current record. Satisfactory prediction results were also observed with the newer and more inclusive SPX dataset, especially for chains with higher number of cysteines. Compared with the literature, our approach is a one-step prediction system which is easier to implement and use. A side by side comparison of SVM and ASNN was conducted. Results indicate that SVM outperform ASNN on this particular problem.

For the prediction of correct pairings of cysteines to form disulfide bonds, we first studied disulfide connectivity by calculating the local interaction potentials between the flanking sequences of the cysteine pairs. The obtained interaction potential was further adjusted by the coefficients related to the existence of binding anchors during disulfide formation and also by the linear distance between the cysteine pairs. Finally, maximized weighted matching algorithm was

applied and performance of the interaction potentials evaluated. Overall prediction accuracy is unsatisfactory.

SVM was used to predict the disulfide connectivity with the assumption that oxidation states of cysteines on the protein are known. Information on binding region during disulfide formation, distance between cysteine pairs, global information of the protein chain and the flanking sequences around the cysteine pairs were included in the SVM encoding. Prediction results illustrated the advantage of using anchor region information.

7.2 Future directions in improving the prediction of oxidation states of cysteines

Based on our current results in prediction of oxidation states of cysteines, it is believed that the predictive accuracy may be further improved by inclusion of more sequence features such as secondary structure and solvent accessibility. Also we propose to convert the problem into a 3-class classification problem. Details can be found in the next two sub sections.

7.2.1 Improve the accuracy by introducing more sequence features

The current research used less assumption than the literature and obtained an overall accuracy of 95%. It can be further improved by either reducing the assumption made on the amino terminal cysteines or by using more comprehensive encoding schemes.

To reduce or totally eliminate the assumptions on the protein chain, we propose a two step process. The first step is to capture the information provided by amino terminal. Previous testings have already shown that the oxidation states of amino terminal shed lights on the oxidation states of other cysteines on the chain, and the flanking sequences around cysteines as well as the global amino acid composition helps define the oxidation states of cysteines too. Therefore, the partial sequence information on the amino end of chain in combination with the

flanking sequence and global sequence composition may be just enough for us to predict the oxidation states of cyesteines on the amino end. Once the oxidation states of the amino terminal cysteines are predicted with decent accuracy, it is easy to eliminate the assumption used previously and predict the oxidation states of other cysteines on the chain. This process is illustrated in the figure below.

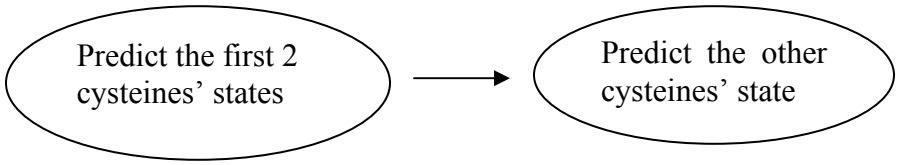


Figure 23 Proposed two-step approach to predict the oxidation states of cysteines without assumptions.

In addition, it is also possible to adopt a more comprehensive encoding scheme in our prediction. So far what we have used in our prediction is physical chemical properties of flanking sequences and sequence composition information of the whole chain. Secondary structure information and solvent accessibility information has not been incorporated into the current prediction model. Literatures[19,28,50] have shown these can be valuable if used properly.

Moreover, a valuable constraint which is that oxidized cysteines have to appear in the protein in pairs has not been fully utilized in our current research work. An effective way to use this property needs to be integrated into our current system. What we propose here is adding a step. The value returned by SVM for each cysteine can represent the probability of a cysteine

residue being in oxidized form or reduced form. When the total number of oxidized cysteines ends up an odd number for a chain, we can check the probability of that residue and determine to find an additional cysteine as oxidized or to eliminate a cysteine that was predicted to be oxidized.

7.2.2 Improving the prediction accuracy by using improved SVM

If examined closely, it can be found that the oxidation state of cysteines is actually a 3-class classification problem rather 2-class. This is because among oxidized cysteines, there are two cases. One is cysteines oxidized to form inter-chain disulfide bonds and the other is cysteines oxidized to form intra-chain disulfide bonds. Although both form disulfide bonds, the functional role and stability[118, 119] may vary from each other. And there may be sequence information around which manifests these differences and these can be captured by the SVM. Currently, the literatures on prediction of oxidation states all combine the reduced cysteines together with the cysteines that form inter-chain disulfide bonds into one big class and call them reduced in general and make it a two-class classification problem. How valid this combination is has never been studied before. To further improve the prediction accuracy, it is necessary to capture the detailed differences between the three classes without arbitrarily combination of different classes.

Support vector machine has been proven to outperform conventional classifiers especially for cases where the training dataset is not too big and the two classes does not intermingle. For multiple classes problem (n classes, for example), what a support vector can do is to transform the system into n two-class problems. Basically, it is to convert the problem into the ith class and the non-ith class. The drawback is that unclassifiable region exists and this in some cases results unsatisifiable prediction results. There has been some improved versions[120, 121] to solve the

problem. Kreßel et al[120] converts the n-class problem to $n(n-1)/2$ two-class problems which cover all pairs of classes. This method is called pairwise classification. However, it was found that the unclassifiable region remains. Decision-tree-based pairwise classification eliminates the unsatisfiable prediction results, but decision boundaries are changed as the order of tree formation is changed. Fuzzy SVM, on the other hand, claims to be able to resolve the unclassifiable region for multi class classification problem. It [122, 123] has shown superior performance on a couple of benchmark data sets.

Based on the above, here we are proposing to improve the prediction accuracy of oxidation states of cysteines by solving the 3-class classification. We would try conventional n class classifications by SVM and also see if the improved SVMs could improve the accuracy further. We believe doing so will help us better understand the sequence environment of different cysteines under different oxidation states and this also will facilitate the prediction of disulfide formation on proteins.

The natural occurrence of inter-chain disulfide bonds is about 11% in the protein level. However, one the dataset we have contains only 1% of such cysteines and the SPX dataset has specifically eliminated those chains with inter-chain disulfide bonds when compiling the dataset. Therefore the intra-chain bonded cysteines are greatly under represented. To have a multiple-class SVM, we will also need to construct our own dataset to increase the population of proteins with inter-chain disulfide bonds.

7.2.3 Improving the prediction accuracy calculating the decision tree

The oxidation states of cysteines on the same protein chain are closely related to each other. This is because it has been shown that the knowledge of statistical oxidation patter of protein chains with similar number of cysteines on the chain in the dataset is able to greatly

improve the prediction model's accuracy[54]. Also intuitively, the total number of oxidized cysteines are always even in a given chain (as far as intra-chain disulfide bonds are considered). Assigning the oxidation state of one cysteine has a significant influence on the assignment of other cysteines. Therefore we believe finding the decision tree of this oxidation state assignment process is essential to the success of the accuracy of prediction.

7.3 Future direction on improving the accuracy of cysteine pairing prediction

Preliminary results have shown that anchor identification does help with the prediction of disulfide bond prediction. Compared with the record holder in literature, we have ok accuracy level before the maximum weighted matching algorithm is applied. However, after the maximum weighted matching algorithm, many of the correctly predicted cysteine pairs were simply eliminated, leading to inferior prediction results than the literature. A mechanisms that bring more restricting factors to further differentiate bonded cysteine pairs from those non-bonded is needed.

A two level model to directly eliminate the false positives is also possible to give improved prediction. In the first level, current SVM encodings can be used to to have a initial screen. At the second level, more restrictions on those that were predicted to be disulfide bonded pairs can be applied. One of the possible restricting factors could be disulfide connectivity pattern (such as cysteine separation profile). Research has shown that by calculating the separation profile of paired cysteines, about 70%[102] of chain level prediction accuracy can be achieved. We believe this may be our future direction for protein chains with relatively small number of disulfide bonds.

APPENDIX I Amino acid representation[68] based on their physical chemical properties

name	five-layer networks trained with fiver parameter set					
	1D	2D		3D		
ALA	0.01	0.13	0.06	0.23	0.19	0.19
GLY	0	0	0	0	0	0.09
VAL	0.93	0.26	0.11	0.89	0.29	0.84
LEU	0.94	0.78	0.23	0.98	0.39	0.48
ILE	0.94	0.44	0.15	1	0.29	1
PHE	0.96	0.92	0.23	0.93	0.59	0.36
TYR	0.96	0.94	0.3	0.5	0.75	0.16
TRP	1	0.98	0.15	0.95	1	0.11
THR	0.74	0.18	0.13	0.41	0.32	0.52
SER	0.03	0.12	0.09	0.16	0.3	0.1
ARG	0.5	1	1	0.05	0.96	0
LYS	0.47	0.97	1	0.02	0.75	0
HIS	0.69	0.9	0.4	0.3	0.65	0.05
ASP	0.09	0.07	0.4	0.23	0.27	1
GLU	0.12	0.13	0.45	0.27	0.4	0.76
ASN	0.37	0.59	0.55	0.11	0.49	0.04
GLN	0.34	0.67	0.45	0.2	0.55	0.09
MET	0.94	0.84	0.3	0.73	0.52	0.21
PRO	0.77	0.18	0.09	0.7	0.17	0.71
CYS	0.92	0.21	0.02	0.93	0.3	0.44

APPENDIX II: Contact potentials used for calculation of amino acid interaction

	TRP	TYR	PHE	HIS	ARG	GLN	GLU	LYS	LEU	ASN	ASP	MET	PRO	ILE	THR	VAL	CYS	SER	ALA	GLY
GLY	0.2	0.3	0.4	0.7	0.3	0.8	1.1	0.7	0.7	0.6	0.8	0.6	1.1	0.5	0.6	0.8	0.9	0.9	1.4	1.7
ALA	-0.5	-0.2	-0.2	0.5	0.4	0.6	1.1	1	-0.1	0.8	1.1	-0.2	0.8	-0.3	0.6	-0.1	0.6	0.9	1	1.4
SER	-0.2	0.1	0.1	0.1	0.3	0.2	0.4	0.7	0.4	0.4	0.3	0.3	0.6	0.4	0.2	0.4	0.4	0.5	0.9	0.9
CYS	-0.7	-0.3	-0.8	-0.2	0.2	0.2	0.8	0.9	-0.5	0.5	0.5	-0.5	0.4	-0.5	0.1	-0.5	-1.7	0.4	0.6	0.9
VAL	-0.9	-0.7	-0.9	-0.1	0.1	0.3	0.5	0.5	-0.9	0.5	0.8	-0.8	0.2	-1	-0.1	-0.8	-0.5	0.4	-0.1	0.8
THR	-0.3	-0.2	-0.2	0	0	0.1	0.2	0.5	0	0.2	0.1	-0.1	0.5	-0.2	0.2	-0.1	0.1	0.2	0.6	0.6
ILE	-1.1	-0.8	-1.1	0	-0.2	0	0.3	0.4	-1.2	0.4	0.6	-0.8	1	-1.1	-0.2	-1	-0.5	0.4	-0.3	0.5
PRO	-0.7	-0.5	-0.2	0.3	0.1	0.4	0.5	0.8	0.1	0.6	0.9	0	0.9	1	0.5	0.2	0.4	0.6	0.8	1.1
MET	-1.2	-0.8	-1.1	-0.3	0	-0.1	0.4	0.3	-1	0.1	0.4	-1	0	-0.8	-0.1	-0.8	-0.5	0.3	-0.2	0.6
ASP	0	-0.1	0.4	-0.2	-0.6	0.2	0.7	-0.2	0.6	0	0.6	0.4	0.9	0.6	0.1	0.8	0.5	0.3	1.1	0.8
ASN	-0.2	-0.2	0	0	0	0	0.3	0.3	0.3	0.1	0	0.1	0.6	0.4	0.2	0.5	0.5	0.4	0.8	0.6
LEU	-1.1	-0.9	-1.1	0	-0.2	0	0.4	0.3	-1.1	0.3	0.6	-1	0.1	-1.2	0	-0.9	-0.5	0.4	-0.1	0.7
LYS	-0.1	-0.2	0.3	0.6	0.6	0.1	-0.4	1.6	0.3	0.3	-0.2	0.3	0.8	0.4	0.5	0.5	0.9	0.7	1	0.7
GLU	-0.2	-0.2	0.2	-0.1	-0.5	0.2	1	-0.4	0.4	0.3	0.7	0.4	0.5	0.3	0.2	0.5	0.8	0.4	1.1	1.1
GLN	-0.5	-0.3	-0.1	-0.1	0	0	0.2	0.1	0	0	0.2	-0.1	0.4	0	0.1	0.3	0.2	0.2	0.6	0.8
ARG	-0.6	-0.7	-0.4	-0.1	-0.1	0	-0.5	0.6	-0.2	0	-0.6	0	0.1	-0.2	0	0.1	0.2	0.3	0.4	0.3
HIS	-0.7	-0.7	-0.3	-0.6	-0.1	-0.1	-0.1	0.6	0	0	-0.2	-0.3	0.3	0	0	-0.1	-0.2	0.1	0.5	0.7
PHE	-1.3	-0.9	-1.2	-0.3	-0.4	-0.1	0.2	0.3	-1.1	0	0.4	-1.1	-0.2	-1.1	-0.2	-0.9	-0.8	0.1	-0.2	0.4
TYR	-1.1	-0.7	-0.9	-0.7	-0.7	-0.3	-0.2	-0.2	-0.9	-0.2	-0.1	-0.8	-0.5	-0.8	-0.2	-0.7	-0.3	0.1	-0.2	0.3
TRP	-1.1	-1.1	-1.3	-0.7	-0.6	-0.5	-0.2	-0.1	-1.1	-0.2	0	-1.2	-0.7	-1.1	-0.3	-0.9	-0.7	-0.2	-0.5	0.2

Appendix III: Publication list

1. Aiguo Du, Yi Pan, Prediction of oxidation states of cysteines (in preparation)
2. A Du, Y Pan, Amino terminus cysteines infer the oxidation states of other cysteines on protein chain, Accepted ANNIE 2007, St Louis, Missouri, Nov 11-14, 2007
3. A Du, X Hu, Y Pan, Prediction of the disulphide bridges in proteins using SVM, International Journal of Bioinformatics Research and Applications, Vol. 3, No. 2, 2007, pp. 223 – 233
4. A Du, Y Pan, Improved algorithms on ant colony optimization in routing, 2003 regional ACM conference, Savannah, GA
5. A Du, H. Ma, S. Chen, Review on discoloration of effluents from dying and printing plants, Dyeing and Printing, 1997, 22(5): 34-39

REFERENCES

1 Chamberlain, L. H. and Burgoyne, R. D., 1997. Activation of the ATPase activity of heat-shock proteins Hsc70/Hsp70 by cysteine-string protein, *Biochem. J.* **322** (3) pp853–858

2 McBride, A. A., Klausner, R. D. and Howley, P. M., 1992. Conserved Cysteine Residue in the DNA-Binding Domain of the Bovine Papillomavirus Type 1 E2 Protein Confers Redox Regulation of the DNA- Binding Activity in Vitro, *Proc. Natl. Aca. Sci,* **89** (16), 7531-7535

3 Hatch, T. P., Allan, I. and Pearce, J. H., 1984. Structural and polypeptide differences between envelopes of infective and reproductive life cycle forms of Chlamydia spp, *J Bacteriol*.; **157** (1): 13-20

4 Berlett, B. S. and Stadtman, E. R., 1997. Protein Oxidation in Aging,Disease, and Oxidative Stress, *J. Biol. Chem*., **272 (**33), 20313–20316

5 Kudo, N, Matsumori, N, Taoka, H., Fujiwara, D., Schreiner, E. P., Wolff, B., Yoshida, M., and Horinouchi, S., 1999. Leptomycin B inactivates CRM1/exportin 1 by covalent modification at a cysteine residue in the central conserved region, *Proc. Natl. Aca. Sci,* **96** (16), 9112-9117

6 Zhang, Z-Y. and Dixon, J. E., 1993. Active Site Labeling of the Yersinia Protein Tyrosine Phosphatase: The Determination of the PKa of the Active Site Cysteine and the Function of the Conserved Histidine 402, *Biochemistry*, **32** (8), 9340-9345

7 Elias, S. J. ArneÂ, R. and Arne, H., 2000. Physiological functions of thioredoxin and thioredoxin reductase, *Eur. J. Biochem*. **267** (2), 6102-6109

8 Vallee, B. L. and Auld, D. S., 1990. Zinc Coordination, Function, and Structure of Zinc Enzymes and Other Proteins, *Biochemistry*, **29** (24):5647–5659.

9 Baldi, P. ,Cheng, J. and Vullo, A., 2005. Large_scale Prediction of Disulphide bond

connectivity, *Advances in Neural Information Processing systems*, 17, MIT Press, Cambridge, MA, 97-104

10 Lehrer, SS, 1978. Effects of an interchain disulfide bond on tropomyosin structure: intrinsic fluorescence and circular dichroism studies, *J Mol Biol.* **118**(2):209-26.

11 Wagner DD, Lawrence SO, Ohlsson-Wilhelm BM, Fay PJ, Marder VJ., 1987. Topology and order of formation of interchain disulfide bonds in von Willebrand factor, *Blood.* **69**(1):27-32

12 Reiter, Y. ; Brinkmann, U. ; Webber, K. O. , Jung, S-H; Lee, B. ; Pastan, I., 1994. Engineering interchain disulfide bonds into conserved framework regions of Fv fragments : improved biochemical characteristics of recombinant immunotoxins containing disulfide-stabilized Fv, *Protein eng*., **7** (5), 697-704

13 Reiter, Y., Brinkmann, U., Jung, SH, Lee, B., Kasprzyk, PG., King, CR. and Pastan, I., 1994. Improved binding and antitumor activity of a recombinant anti-erbB2 immunotoxin by disulfide stabilization of the Fv fragment, *J. Biol. Chem*., **269** (28), 18327-18331

14 Nagahara, N., Yoshii, T., Abe, Y., Matsumura, T., 2006. Thioredoxin-dependent enzymatic activation of mercaptopyruvate sulfurtransferase: An intersubunit disulfide bond serves as a redox switch for activation. *J Biol Chem*. 2006 Nov 27; [Epub ahead of print]

15 Mannick, JB., 2006. Regulation of apoptosis by protein S-nitrosylation. *Amino Acids.* 2006 Nov 30; [Epub ahead of print].

16 Vielle, C., Zeikus, G., 2001. Hyperthermophilic enzymes: sources, uses and molecular mechanisms for thermostability, *Microbiol Mol Biol Rew*; **65** (1), 1-43.

17 Hogg, P. J., 2003. Disulfide bonds as switches for protein function, *Trends Biochem Sci*, **28** (4), 210-214.

18 Mitsunori, F., Eiko, K., and Katsuhiko, M**.,** 1999. Conserved N-terminal Cysteine Motif Is

Essential for Homo- and Heterodimer Formation of Synaptotagmins III, V, VI, and X, *J Biol Chem*, **274** (44), 31421-31427

19 Chuang, CC., Chen, CY., Yang, J-M., Lyu, PC., Hwang, JK., 2003. Relationship between protein structures and disulfide-bonding patterns, *Proteins: Structure, Function, and Bioinformatics,* **53** (1), 1-5

20 Huck C.W., Bakry R., Bonn G. K., 2005. Progress in capillary electrophoresis of biomarkers and metabolites between 2002 and 2005, *Electrophoresis*,**27** (1) , 111 - 125

21 Toyo'oka T. et al, 1986. Amino acid composition analysis of minute amounts of cysteine-containing proteins using 4-(aminosulfonyl)-7-fluoro-2,1,3- benzoxadiazole and 4-fluoro-7-nitro-2,1,3-benzoxadiazole in combination with HPLC, *Biomed. Chromatogr*. 1(1), 15-20.

22 Kim S. O., Merchant K. et al., 2002. OxyR A Molecular Code for Redox-Related Signaling, *cell*, **109** (3), 383-396

23 Horton, H.R., Moran, L.A., Ochs, R. S., Ravn, J. D. and Scrimgeour K.G., 1996. *Principles of Biochemistry*, second edition, Upper Saddle River, NJ: Prentice-Hall, Inc, 1996, pp102.

24 Stenesh J., 1998. Chapter 3: Proteins, *Biochemistry,* New York*:* Springer,1998, pp.54.

25 Baldi P., Cheng J. and Vullo A., 2004. Large scale Prediction of disulphide bond connectivity, *Advances in Neural Information Processing systems*, vol. 17, Cambridge, MA: MIT Press, pp97-104.

26 Hinck A. P., Truckses Dm and Markley J. L., 1996. Engineered disulfide bonds in staphylococcal nuclease: effect on the stability and conformation of the folded protein. *Biochemistry*, **35** (32): 10328-38.

27 Shimalka M, Lu C, Salas A., Xiao T., Takagi J. and Springer T. A., 2002. Stabilizing the integrin alpha M inserted domain in alternative conformations with a range of engineered disulfide bonds, *Proc. Natl. Aca. Sci USA*, **99** (26), 16737-41.

28 Cheng J., Saigo H. and Baldi P., 2006. Large-scale prediction of disulphide bridges using kernel methods, two dimensional recursive neural networks and weighted graph matching, *Proteins: structure, function and bioinformatics,* **62** (3), pp.617-29.

29 Bardwell J., 2004. The dance of disulfide formation, *Nat. Struct. Mol. Biol*., **11** (7), pp582-3

30 Melissa F. Schwaller, Bonney Wilkinson, and Hiram Gilbert, 2002, Reduction/reoxidation cycles contribute to catalysis of disulfide isomerization by protein disulfide isomerase, *J. Biol. Chem*, pp 1074,

31 Darby, N. J., Penka E. and Vincentelli R., 1998. The multi-domain structure of protein disulfide isomerase is essential for high catalytic efficiency, *Journal of Molecular biology*, **276** (1), pp.239-47.

32 Bommarius A. S. and Riebel B. R., 2004. *Biocatalysis*, Weihneim, Germany: Wiley-VCH, 2004, pp.20.

33 Bisswanger H, 2004. *Practical Enzymology,* Germany: Wiley-VCH, 2004

34 Fischer E., 1894. Einfluss der configuration auf die wirkung derenzyme, *Ber. Dt.Chem. Ges*. **27**, pp.2985-93

35 Cooper, G. M., 2000. *The Cell: A Molecular Approach*, Second Edition, Washington DC: Sinauer Associates, Inc

36 Berg, J. M., 2002. Tymoczko J. L., Stryer L. and Clarke N., *Biochemistry*, fifth edition, England: W. H. Freeman and Company, 2002.

37 Andrew G. McArthur, Leigh A. Knodler, Jeffrey D. Silberman, Barbara J. Davids, Frances D. Gillin and Mitchell L. Sogin, 2001, The Evolutionary Origins of Eukaryotic Protein Disulfide Isomerase Domains: New Evidence from the Amitochondriate Protist Giardia lamblia, *Molecular Biology and Evolution* **18**:1455-1463

38 Ruddock L. W., Freedman R. B. and Klappa P., 2000. Specificity in the substrate binding by protein folding catalysts: Tyrosine and tryptophan residues are the recognition motifs for the binding of peptides to the pancreas-specific protein disulfide isomerase PDIp, *Protein Science*, **9** (4), pp.758-64.

39 Pirneskoski A., Klappa P., Lobell M., Williamson R. A., Byrne L. and Alanen H., Molecular characterization of the principal substrate binding site of the ubiquitous folding catalyst protein disulfide isomerase, *J. Biol. Chem*., 279 (11), pp.10374-81.

40 Goldberg M. E. and Guillou Y.,1994, Native disulfide bonds greatly accelerate secondary structure formation in the folding of lysozyme, *Protein Science*, **3**(6), pp 883-887

41 Ferre F. and Clote P., 2005. Disulfide connectivity prediction using secondary structure information and diresidue frequencies, *Bioinformatics* 21 (10), pp.2336-46.

42 James D. Jamieson 1 and George E. Palade, 1967, Intracellular transport of Secretory Proteins in the Pancreatic Exocrine Cell: II. Transport to Condensing Vacuoles and Zymogen Granules, *The J. of Cell Bio.*, **34**, 597-615

43 James D. Jamieson 1 and George E. Palade, 1967, Intracellular transport of Secretory Proteins in the Pancreatic Exocrine Cell: I. Role of the Peripheral Elements of the Golgi Complex, *The J. of Cell Bio*, **34**, 577-596

44 James D. Jamieson 1 and George E. Palade, 1971, Synthesis, intracellular transport and discharge of secretory proteins in stimulated pancreatic exocrine cells, *The J. of Cell Bio*, **50**, 135-158

45 Palade, G (1975). Intracellular aspects of the process of protein synthesis. *Science,* **189**: 347-358

46 Gunter Blobel, Bernhard Dobberstein, 1975, Transfer of Proteins across Membranes. I. Presence of Proteolytically Processed and Unprocessed Nascent Immunoglobulin Light Chains on Membrane-Bound Ribosomes of Murine Myeloma**,** *The Journal of Cell Biology*, **67**(3), pp. 835-851

47 Theodoridis S and Koutroumbas K., 2001, Pattern recognition and neural networks, in Machine learning and its applications: advanced lectures, Paliouras G, Karkaletsis V. and Spyropoulos C (eds), Springer-Verlag, Berlin Heideberg, Germany, pp169

48 Muskal, S., Holbrook, S. and Kim, S., 1990. Prediction of the disulfide-bonding state of cysteine in proteins, *Prot. Eng.,* **3** (8), 667-672

49 Fiser, A., Cserzo, M., Tudos, E., Simon, I., 1992. Different sequence environments of cysteines and half cystines in proteins. Application to predict disulfide forming residues., *FEBS Lett.* **302** (2), 117-20**.**

50 Muccielli-Giorgi, MH., Hazout, S., Tuffery, P., 2002. Predicting the disulfide bonding state of cysteines using protein descriptors. *Proteins*; 46: 243-249

51 Martelli, P.L., Fariselli, P., Malaguti, L. and Casadio, R., 2002. Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy, *Protein Science*, **11**:2735-2739.

52 Martell, P. L., Fariselli, P., Malaguti, L. and Casadio, R., 2002. Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks, *Protein Engineering,* **15** (

12), 951-953

53 Frasconi, P. Passerini, A., Vullo, A., 2002. A two-stage SVM architecture for predicting the disulfide bonding state of cysteines**,** *Neural Networks for Signal Processing,. Proceedings of the 2002 12th IEEE Workshop on*, 25- 34

54 Chen, YC., Lin, YS., Lin, CJ. and Hwang, JK., 2004. Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences, *Proteins: structure, function and bioinformatics,* **55** (4),1036-1042

55 Ceroni, A., Passerini, A., Vullo, A., and Frasconi, P., 2006. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server, *nucleic acids research*, **34**:W177-W181

56 Piero Fariselli, Paola Riccobelli, Rita Casadio, 1999, Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins, *Proteins: Structure, Function, and Genetics*, 36 (3) , pp340 – 346

57 Cristianini, N., and Shawe-Taylor J., 2000, *An introduction to Support Vector Machines*. Cambridge University Press

58 Vapnik, V. and Cortes, C., 1995. Support vector networks. *Machine Learning,* **20**, 273-293

59 Joachims, T., 2004, SVMlight Support Vector Machine, http://svmlight.joachims.org/

60 http://en.wikipedia.org/wiki/Support_vector_machine

61 Howard, E. A., Zupan, J. R., Citovsky V. and Zambryski P. C**.**, 1992**,** The VirD2 protein of A. tumefaciens contains a C-terminal bipartite nuclear localization signal: Implications for nuclear uptake of DNA in plant cells, *Cell*, **68**(1):109-118

62 Dingwall C., R A Laskey R. A., 1986, Protein Import into the Cell Nucleus, *Annual Review of Cell Biology*, **2,** pp367-390

63 Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1983) *Molecular Biology of The Cell,* Garland Publishing, New York pp. 340-349

64 Darnell, J., Lodish, H. and Baltimore, D. (1986) Molecular Cell Biology, W.H. Freeman & Co., New York, pp. 940-957

65 Stryer, L., (1981) *Biochemistry*, W.H. Freeman & Co., New York pp. pp712-714

66 Voet, D. and Voet, J.G. (1990) *Biochemistry*, John Wiley & Sons, New York , pp. 298-304

67 Walter, P., Gilmore, R. and Blobel, G. (1984) Protein translocation across the endoplasmic-reticulum, *Cell* **38**, pp5-8

68 Meiler, J., Muller, M., Zeidler, A., 2001. Schmaschke, F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks, *J Mol Model*; **7**,360-369

69 Berman, HM., Westbrook, J., Feng, Z., Gilliland, G., Bhat, TN., Weissig, H., Shindyalov, IN., Bourne PE. 2000. The Protein Data Bank, *Nucleic Acids Res*, **28**, pp235-242.

70 Sander C. Schneider R. 1991, Database of homology-derived protein structures and the structural meaning of the sequence alignment, proteins; **9**: pp56-68

71 Vapnik, V., 1995. *The nature of statistical learning theory*. New York: Springer

72 Matthews, B. W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim, *Biophys Acta*; **405**, pp442-451

73 Youden, W.J. 1950. Index for rating diagnostic tests. *Cancer,* **3**, 32-35.

74 Scuse D., 1999 , "Neural networks", Slides for course 74.436 Machine Learning: University of Manitoba, MB, Canada

[75] http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

76 Udo S., 2001, Multiple Layer Perceptron Training Using Genetic Algorithms, *ESANN'2001 proceedings - European Symposium on Artificial Neural Networks,* Bruges (Belgium), D-Facto public, pp159-164

77 Utans, J., Moody, J. Rehfuss, S., Siegelmann, H.; 1995. Input variable selection for neural networks: application to predicting the U.S. business cycle, *Proc. Of IEEE/IAFE 1995 Comput. Intellig. for Financial Eng.*, pp 118-122

78 Ahn B. S., Cho S. S. and Kim C. Y., The integrated methodology of rough set theory and artificial neural network for business failure prediction , *Expert Systems with Applications* , **18**(2), pp65-74

79 Rodger J.A., 2003, Utilization of data mining techniques to detect and predict accounting fraud: a comparison of neural networks and discriminant analysis, *Managing data mining technologies in organizations: techniques and applications*, Idea Group Publishing, Hershey, PA, pp174 – 187

80 Phua C., Alahakoon D., Lee V., 2004, Minority report in fraud detection: classification of skewed data, *ACM SIGKDD Explorations Newsletter*, **6**(1), pp50 - 59

81 Misra H., Ikbal S., Yegnanarayana B., 2003, Speaker-specific mapping for text-independent speaker recognition, *Speech Communication*, **39**(3-4), pp301-310

82 Sarishvili A., Andersson C., Franke J., Kroisandt G., 2006, On the Consistency of the Blocked Neural Network Estimator in Time Series Analysis, *Neural Computation*, **18** (10), p.2568-2581

83 Ambrožič T., Turk G., 2003, Prediction of subsidence due to underground mining by artificial neural networks, *Computers & Geosciences*, **29** (5), pp.627-637

84 Cancelliere R., Gai M., 2003, A comparative analysis of neural network performances in astronomical imaging, *Applied Numerical Mathematics*, **45** (1), pp.87-98

85 Dasarthy, B., 1991, *Nearest neighbor(NN) norms*, IEEE computer society press, Washington, DC

86 Hardle, W., 1990, *Smoothing techniques with implementation in S*, Springer-Verlag, New York

87 Lawrence, S., Tsoi, A.C. and Back, A.D, 1996, Function approximation with neural networks and local methods: bias, variance and smoothness, in: P. Bartlett, A. Burkitt and R. Williamson (eds), *Australian conference on neural networks,* Australian National University, pp16-21

88 Tetko I. V., 2002, Associative Neural Network, *Neural Processing Letters,* **16**: 187-199.

89 Abeles, M., 1991, *Corticotronics: Neural circuits of the cerebral cortex*, Cambridge University Press, New York

90 Villa, A. E. P., Tetko, I. V., Hyland, B. and Najem, A, 1999, Spatiotemporal activity patterns of rat cortical neurons predict responses in a conditioined task, *Proceedings of the Natioanl Academy of Sciences of the USA,* **96**, pp1106-1111

91 Thorpe, S., Fize, D. and Marlot, C., 1996, Speed of processing in the human visual system, *Nature*, **381**, pp520-522.

92 Tetko, I. V., Tanchuk, V. Y., 2002, Application of associative neural networks for predictions of lipophilicity in ALGOGPS 2.1 program. *Journal of Chemical Information and Computer Sciences*, **42**(5), pp1136-45.

93 Tetko I. V., 2002, Neural Network studies. 4. Introduction to Associative Neural Networks, *Journal of Chemical Information and Computer Sciences*, **42** (3), pp717 -728

94 Tetko, I. V. 2002, Associative neural network, *Neural Processing Letters*, **16**, pp187-199

95Tetko, I. V.; Tanchuk, V. Y. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program, *J. Chem. Inf. Comput. Sci*., **42**, pp.1136-45

96 VCCLAB, 2005, Virtual Computational Chemistry Laboratory, http://www.vcclab.org

97 Tsai C.H., Chen B.J., Chan C.H., Liu H.L. and Kao C.Y., 2005. Improving disulfide connectivity prediction with sequential distance between oxidized cysteines, *bioinformatics*, **21** (24), pp.4416-9.

98 Chen B-J., Tsai C-H., Chan C-H., and Kao C-Y., 2006. Disulfide connectivity prediction with 70% accuracy using two level models, Proteins: Structure, Function and Bioinformatics **64**(1), pp246-252.

99 Fariselli P. and Casadio R., 2001. Prediction of disulfide connectivity in proteins, *bioinformatics*, **17** (10), pp957-64.

100 Fariselli P., Riccobelli P. and Casadio R., 2002. A neural network based method for predicting the disulfide connectivity in proteins, in E. Damiani et al., ed., *Knowledge based intelligent information engineering systems and allied technologies(KES 2002)*, Amsterdam: IOS Press, pp. 464-8.

101 Ferre F. and Clote P., 2005. Disulfide connectivity prediction using secondary structure information and diresidue frequencies, *Bioinformatics* **21** (10), pp.2336-46.

102 Deng H., 2007, Ph.D thesis, Identifying Calcium-binding Sites and Predicting Disulfide Connectivity Georgia State University,

103 Rama G. L. J., Shjilton A., Parker M. M., Palaniswami M., 2005, Disulfide bridge prediction using fuzzy support vector machines, Proc. of the International Conference on Intelligent Sensing and Information Processing, Bangalore, India. pp49-54

104 Chen B.J. Tsai C-H, Chan C-H and Cheng YK, 2006, Disulfide connectivity Prediction with 70% accuracy using two-level models, Proteins: structure, function and bioinformatics, **64**, pp246-252.

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

106 Altschul, S. F., Warren G., Webb M., Eugene W. M. and David J. L., 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**, pp.403-10.

107 Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. 1997, The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, 24:4876-4882.

108 Bryan Philip N, 1995, Chapter 12 Site-Directed Mutagenesis to study protein folding and stability, in Protein Stability and Folding: theory and practice, Bret A. Shirley (eds), Humana Press Inc, New Jersey.

109 Mirny, L. A., and Shakhnovich, I. I. 1996. How to derive a protein folding potential? A new approach to an old problem. *Journal of moleculear biology,* **264**, 1164-1179.

110 Skolnick J., Kolinski A., and Ortiz A., 2000, Derivation of Protein-Specific Pair Potentials Based onWeak Sequence Fragment Similarity, *Proteins: Structure, Function, and Genetics* **38**, pp.3–16

111 Edmonds J. Paths, trees and flowers, 1965, *Canadian J. Matehmatics*,; **17**, pp. 449-467

112 Gabow H. N. 1976, An efficient implementation of Edmond's algorithm for maximum weight matching on graphs, *J. of ACM* , **23**, pp. 221-234.

113 http://elib.zib.de/pub/Packages/mathprog/matching/weighted/

114 Joachims T., 1999. Making large-Scale SVM Learning Practical. In Schölkopf B., Burges C.

and Smola A. ed., *Advances in Kernel Methods - Support Vector Learning*, MIT-Press

115 Meiler J., Muller M., Zeidler A., Schmaschke F., 2001. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks, *J Mol Model;* **7**(9). pp.360-9.

116 Vullo A., Frasconi P., 2004. Disulfide connectivity prediction using recursive neural networks and evolutionary information. *Bioinformatics*, **20** (5), pp653-659.

117 Bairoch A. and Apweiler R., 2000. The SWISS-PROT protein sequence database and its supplemental TrEMBL. *Nucleic Acids Res,* **28** (1), pp.45-8.

118 Ishibashi J, Kataoka H, Isogai A, Kawakami A, Saegusa H, Yagi Y, Mizoguchi A, Ishizaki H, Suzuki A. 1994, Assignment of disulfide bond location in prothoracicotropic hormone of the silkworm, Bombyx mori: a homodimeric peptide. *Biochemistry*. **33**(19), pp5912-9

119 Kownatzki E. 1973, Disulfide Bonds of Human IgM: Differential Sensitivity to Reductive Cleavage, Scandinavian. *Journal of Immunology,* **2** (4), pp 433–438

120 Kreßel U. H.-G., 1999, Pairwise classification and support vector machines. In B. Sch¨olkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pp 255–268. The MIT Press, Cambridge, MA

121 Platt J. C., Cristianini N., and Shawe-Taylor J.. Large margin DAGs for multiclass classification. In S. A. Solla, T. K. Leen, and K.-R. M¨uller, editors, *Advances in Neural Information Processing Systems*, **12**, pp.547– 553. The MIT Press

122 Inoue T. and Abe S. 2001, Fuzzy support vector machines for pattern classification. *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, **2**, pp 1449–1454

123 Abe S. and Inoue T., Nada R., 2002, Fuzzy Support Vector Machines for Multiclass Problems, *ESANN'2002 proceedings - European Symposium on Artificial Neural Networks*, Bruges (Belgium), 24-26 April 2002, d-side publi., pp. 113-118