IN SILICO DETECTION AND PREDICTION OF GLYCOSYLATION SITES IN THE EPIDERMAL GROWTH FACTOR-LIKE PROTEINS USING FEED-FORWARD NEURAL NETWORKS

Alireza Darissi Shaneh

A thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Computer Science
Concordia University
Montréal, Québec, Canada

September 2006

Your file Votre référence
ISBN: 978-0-494-20774-1
Our file Notre référence
ISBN: 978-0-494-20774-1

Abstract

In silico Detection and Prediction of Glycosylation Sites in the Epidermal Growth Factor-Like Proteins using Feed-Forward Neural Networks

Alireza Darissi Shaneh

Biological databases are sparse, huge and redundant. Therefore, knowledge inference from those databases needs a consistent approach. Widely accepted as a most complex process of protein modification, *glycosylation* has been the main focus in this study. In this process a simple chain of carbohydrates attaches to a target protein at a specific amino acid, so-called *glycosylation site. Epidermal Growth Factor-Like* (EGFL) repeats have been the target proteins of this study because of having a particular glycosylation process. Moreover, they may associate with many type of cancer as wel las other diseases. The objective of this study was to detect and predict the number of glycosylation sites in EGFL protein sequences using feed-forward neural networks. Bayesian automated regularization was exploited to prune the unnecessary weights and biases of the feed-forward neural network. The result of applying eight learning algorithms showed that One Step Secant (OSS) learning algorithm is more reliable than the others in terms of the accuracy and performance as measured in this study. The Bayesian regularized neural network outperformed OSS method according to the employed assessment measures. Compared to the existing neural detectors, Bayesian automated learning could improve the consistency of the model by 39.48%. The concept of *Reduction Factor* was also introduced to determine the efficiency of Bayesian automated learning quantitatively.Glycobiologists can use and validate such connectionist models to choose and study on the selected EGF-like proteins which are associated with cell malignancy.

Acknowledgments

I had a unique opportunity to work under the supervision of Prof. Gregory Butler. His bright mathematical mind kept me in the right track of the systematic and scientific method necessary for carrying out this project. I would also like to appreciate the effective opinions of my friends at Dr. Butler's group.

My sincere gratitude goes to Prof. Annette Herscovics, FRSC, (McGill Cancer Centre, McGill University) for hours of discussion on protein glycosylation, especially mannosylation- and α-mannosidase-related protein sequences. In addition, I would like to thank Prof. Robert Haltiwanger's Lab (Department of Biochemistry and Cell Biology, State University of New York at Stony Brook) for kind provision of O-fucosylation and O-glucosylation data of EGF-like oncoprotein sequences.

My special thank and love goes to Patricia Gaitan, a molecular biologist whose exceptional ability in problem analysis proved that she should also have studied mathematics.

I dedicate this work to my family for their forever supports.

Finally, I was inspired from two representatives of harmony and perfection in this world: Johann Sebastian Bach for his highly calculated music and Erwin Schrödinger for his outstanding classical work on molecular biology and its physical perspective.

Dedication

Now, for there is an infinity of possible universes among God's ideas, and only one of them can exist, there must be a sufficient reason for God's choice, which determines him to the one rather than to the other. This reason can be found only in harmony, or the degrees of perfection which these worlds contain, since each possible world has the right to claim existence in proportion to the perfection it includes. Thus, nothing is entirely arbitrary. This is the cause of the existence of the best, which his wisdom makes him know, which his goodness makes him choose, and which his power makes him produce. This is the means for obtaining as much variety as possible, but with the greatest order as possible. In other words, it is the means for obtaining as much perfection as possible.

— Baron Gottfried Wilhelm von Leibniz (1646-1716)
La Monadologie, theses 53–56 & 58

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The remarkable advances in multidisciplinary areas of the science are logical consequences of an organized cooperation among the scientists working in different branches. Bioinformatics is the result of such cooperation between biologists and computer scientists. As a fast growing field, bioinformatics has been revolutionized by significant progress in database technology. Biological datasets are stored in databases for further manipulation and annotation. Widely accepted as one the most vital molecules in the nature, proteins participate in many critical biological pathways in mammalian species. Consequently, biologists utilize those databases to store protein data. Nevertheless, protein datasets are huge and sparse in terms of their structural variation and functionality [78]. In addition, such varieties in function and shape produce an intrinsically redundant database. As a result, inferring desired and specific knowledge from protein datasets is a challenging process. PROSITE [66], UniProt [4, 8] and SWISS-PROT [19] are examples of protein sequence databases with their own format and annotation styles.

PTMs are the necessary chemical modifications applied on proteins to regulate their functions [63]. About 325 PTMs have been discovered [42, 68, 35]. One of the most common modifications is *glycosylation* in which a simple chain of saccharides attaches to a specific amino acid of the target protein [113]. The attached amino acid, *glycosylation site*, is specific and sometimes unique in proteins. Moreover, the complex of carbohydrate-protein, *glycoprotein*, is responsible for many important biological interactions inside and outside of the cell [21].

The distribution of glycosylation sites along the sequence of a glycoprotein is an interesting subject for biochemists [113, 116, 1, 77, 47], since aberration in the number or order of glycosylation sites causes irreversible and serious diseases such as brain or lung cancer [77, 33, 54, 55, 31, 32, 46]. Therefore, studying protein glycosylation with respect to the distribution of the number of glycosylation sites provides biologists with the information necessary for selecting particular proteins for their research.

Incorporated with large datasets, soft computing techniques are powerful tools to infer the necessary knowledge out of those data. They find a solution or set of solutions in a non-linear search space; in fact, soft computing methods optimize a search space by intelligently limiting it around extreme solution or solutions. Among the various kinds of computational intelligence methods, neural networks are shown to quickly adapt with large sets of data whose non-linear functionalities are under study in conjunction with some certain selected features [123, 74]. Hence, It is possible to employ a class of neural networks, so called *feed-forward networks* or *multilayer perceptron*, to mine the required glycosylation sites information in a particular subfamily of a protein. While extracting that information, it is important to choose the characteristics special to the protein which is under study.

Epidermal Growth Factor-Like superfamily (EGFLs or EGF-like) [23, 3] belongs to *growth factor* proteins. EGF-like proteins have a distinct glycosylation process [59, 118, 79] which makes them good candidates for this study, showing in Section 1.2. The abnormal process of glycosylation in EGFLs leads to unexpected results in cell growth as well as serious diseases such as Congenital Disorders of Glycosylation (CDGs) and cancer [124, 40, 98, 37, 112, 33, 7, 107]. EGFLs are well-annotated, and biochemists have determined different types of glycosylation in those proteins [127, 41]. Consequently, EGFLs provide a reasonably valid information for a feed-forward neural network as prior knowledge.

The non-linear mapping of the distribution of the number of glycosylation sites in EGFLs can be inferred by a feed-forward neural network which, in turn, can detect and predict the glycosylation sites, the subject building the foundation of this thesis.

Figure 1 shows the organization of the thesis. First, the mechanism of glycosylation as well as a review on Epidermal Growth Factor-like proteins will be presented. After that, the application of computational intelligence methods in the field of protein glycosylation will be discussed. Feed-forward neural networks is the topic covered in the third section. The final section of Chapter 1 explains the criteria of choosing this project. The nature, compilation, and encoding of the EGFLs data will be presented in Chapter 2. *Statistical Learning Theory* elucidates *ab initio* laws governing machine learning methods including neural networks. Extensively used in many scientific applications, feed-forward networks are the subject emphasized in Chapter 2 according to the statistical learning theory. In phase I of this study, eight learning algorithms were applied and used to choose the most reliable algorithm to train the underlying network. Bayesian regularization is a consistent technique which has effectively reduced the size of the neural network to improve the generalization of the detection and prediction [111]. In phase II, Bayesian learning was compared to the chosen algorithm in phase I. The workflows of those phases are introduced in Chapter 2. Subsequently, the last chapter discusses the results obtained from both phases. Moreover, this chapter reviews the assessment measures utilized. The results of Bayesian neural network were also compared to the existing systems of detection and prediction of glycosylation sites in protein sequences, which is covered by Chapter 3. In conclusion, the lessons learned during this project as well as the possible future works will be outlined in Chapter 4.

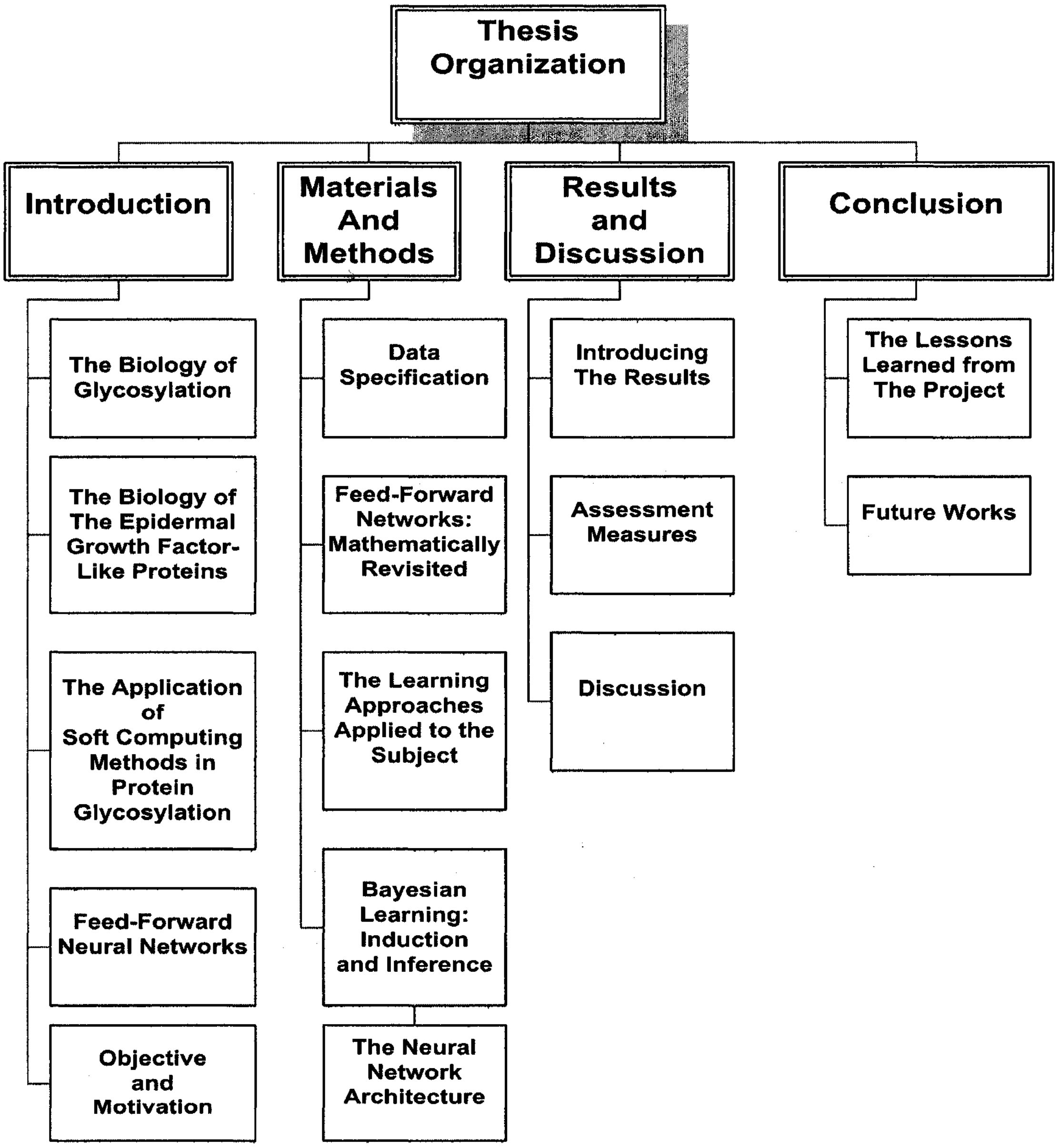

Thesis Organization
Introduction
Materials And Methods
Results and Discussion
Conclusion
The Biology of Glycosylation
The Biology of The Epidermal Growth Factor-Like Proteins
The Application of Soft Computing Methods in Protein Glycosylation
Feed-Forward Neural Networks
Objective and Motivation
Data Specification
Feed-Forward Networks: Mathematically Revisited
The Learning Approaches Applied to the Subject
Bayesian Learning: Induction and Inference
The Neural Network Architecture
Introducing The Results
Assessment Measures
Discussion
The Lessons Learned from The Project
Future Works

Figure 1: The Organization of the Thesis

1.1 The Biology of Protein Glycosylation

Having been synthesized and translocated, proteins usually need to undergo structural modifications to achieve their full functionality. These processes are called *post translational modifications* (PTMs) cite [63]. Biologists have detected over one hundred of such modifications [5]. Among those structural changes, *glycosylation* is the most diverse and common one by which all membrane proteins and most of secretory ones are modified [81]. In the process of glycosylation, a chain of saccharides attaches to a specific amino acid of the target protein. The attachment site is called *glycosylation site.* The product of glycosylation process is a complex of saccharides-protein so-called *glycoprotein.* Depending on the type of the process, glycosylation site varies from one type to another, and glycobiologists have discovered 13 different monosaccharides along with 8 amino acids, which make participate in 41 carbohydrate-peptide linkage [113]; however, in this thesis Asparagine (*Asn* or *N*), Serine (*Ser* or *S*) and Threonine (*Thr* or *T*) have been emphasized because they are the most common glycosylation sites observed in proteins [5]. The diversity in glycosylation sites leads to the polymorphism of glycoproteins, referred to as *site heterogeneity.* Site heterogeneity yields various forms of a glycoprotein, *glycoforms*, which may have different biological properties [81, 113]. Glycosylation process occurs in two major subcellular compartments: *Endoplasmic Reticulum* (ER) and *Golgi apparatus.* In ER, the simple sugars are added to the protein. Subsequently, the folded glycoprotein is transported to Golgi apparatus to obtain more chains of carbohydrates, and some of already incorporated saccharides are removed in Golgi apparatus. The matured glycoprotein is then translocated to other organelles to do its functions. Alternatively, it may be sent out of the cell to perform particular tasks, a phenomenon called *secretion.*

In terms of biochemical characteristics, the chain of saccharides, which are usually referred to as *glycans*, are responsible for controlling solubility, electrical charge, mass, size and viscosity of a glycoprotein [81]. Glycans regulate biological function of a glycoprotein such as intracellular traffic, localization, activity and cell-cell interaction [81, 113]. Furthermore, It has been shown that glycosylation has a remarkable effect on reproducing hormones [122].

The major identified types of glycosylation are *N-linked glycosylation*, *O-linked glycosylation*, *glypiation and GPI anchoring*, and *C-linked mannosylation* and *P-glycosylation* [113]. The first two types, N- and O-linked glycosylation, form the structure of the subjects discussed in this study.

1.1.1 N-Linked Glycosylation

N-linked glycosylation (Figure 2) is a prominent and stable mechanism in which the amide side of Asparagine ($-NH_2$) of the target polypeptide attaches to the chain of polysaccharides or glycans. This process is necessary for a proper protein folding. Site-specific and enzyme-directed, N-linked glycosylation is a co-translational process, and it occurs in endoplasmic reticulum while translating m-RNA to the protein [61, 1, 125, 77]. Almost all membrane proteins in eukaryotes and archaea undergo this process. In prokaryotes, N-linked glycosylation rarely happens; however, there are few cases reported [116].

Glycosylation process heavily depends on the structure of the underlying glycans. The glycan is a 14-carbohydrate precursor consisting of 3 Glucose, 9 Mannose and 2 N-Acetylglucosamine ($GlcNAc$). In the first step, the glycan appends to a carrier molecule called *dolichol* through a complicated enzymatic reactions. In this step, the attachment is subject to Mannose and GlcNAc monosaccharides. In the second step, the complex of glycan-dolichol now translocated into the lumen of ER using the enzyme *flippase.* Catalyzed by an enzyme, *oligosaccharyl transferase*, the glycosylation goes further with addition of more Mannoses and finally Glucose inside the ER. The final step is to release dolichol from the glycan and attach it to the amide aide of Asparagine in the target protein.

Afterward, the produced glycoprotein (Figure 3(a)), is forwarded to Golgi apparatus. In that compartment, the glycoprotein goes through the trimming and adding process. Some Mannoses added to the glycoprotein in the ER are removed. Consequently, the removal of simple carbohydrates in Golgi apparatus results in a *core* N-glycan, which may be elongated by adding other types of sugars. Figure 3(b) shows N-glycan core structure. The structure consists of 3 Mannoses and 2 GlcNAc. There is a strong evidence that the participating Asparagine in the process of N-linked glycosylation should match with the consensus pattern $Asn - X - Ser|Thr$ where X is any amino acid but Proline [87]. This consensus sequence is literally known as *N-linked sequon.* There are two types of N-linked glycans: High Mannose and complex polysaccharides.

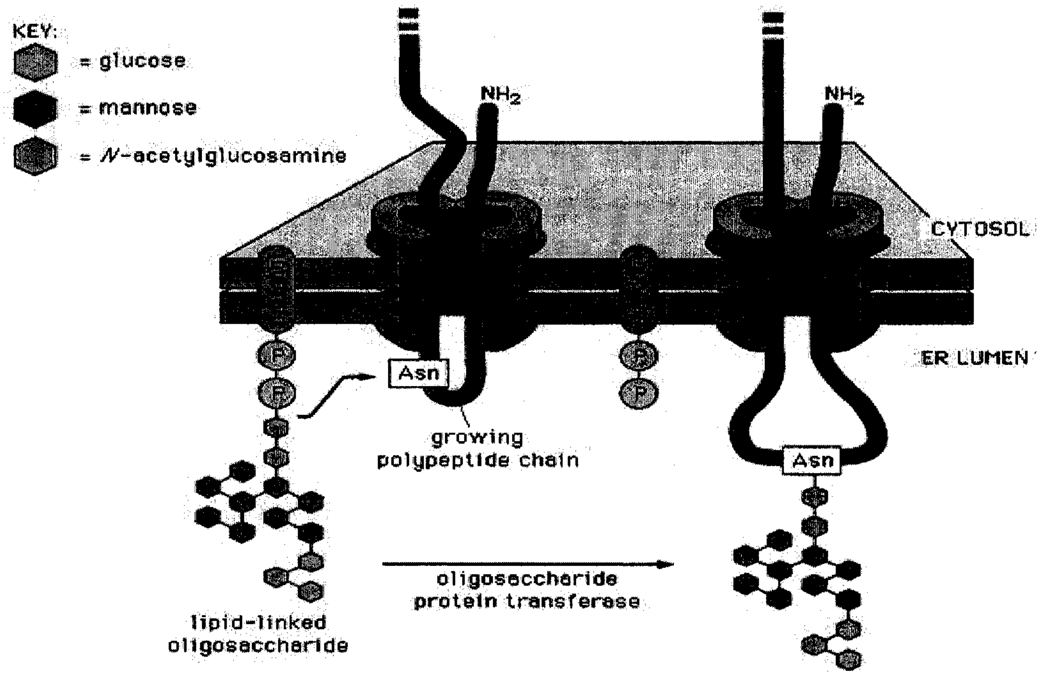
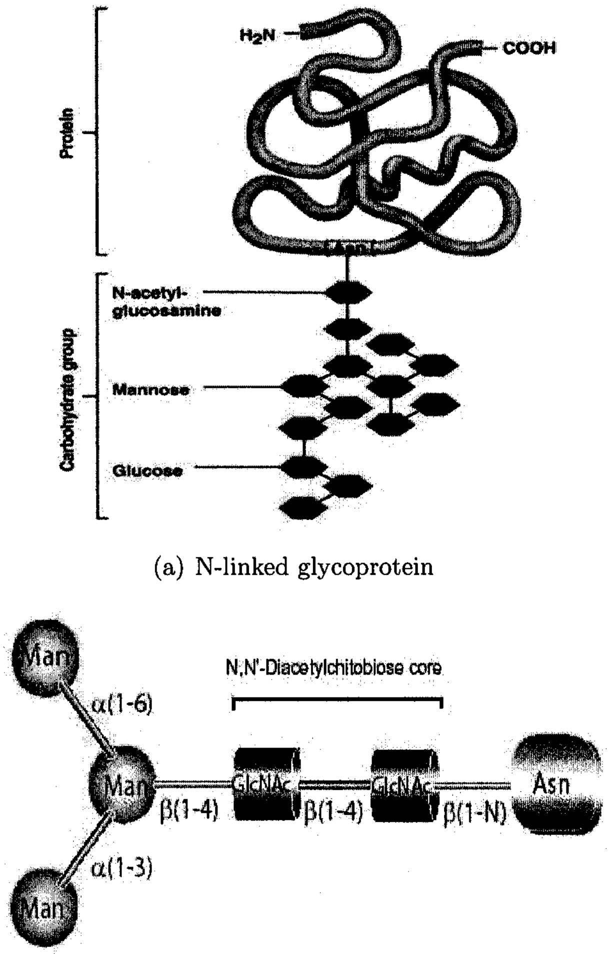

KEY:
= glucose
= mannose
= *N*-acetylglucosamine
NH2
NH2
CYTOSOL
ER LUMEN
Asn
growing polypeptide chain
Asn
P
P
P
P
oligosaccharide protein transferase
lipid-linked oligosaccharide

Figure 2: N-linked glycosylation Mechanism [31]



(a) N-linked glycoprotein

(b) N-glycan core structure

Figure 3: Hallmarks of N-linked glycosylation

High Mannose glycans have simply two GlcNAc as well as large numbers of Mannose in their structure. Those glycans are accessible by Golgi apparatus for Mannose trimming purposes. On the other hand, complex type glycans contains more than two basic GlcNAc molecules; furthermore, they may have diverse saccharides in their structure. Similar to High Mannose ones, complex polysaccharides are also accessible to Golgi apparatus for further trimming [113].

N-linked glycosylation plays an important role in the stability and proper folding of a protein [125]; nevertheless, it is not the only important modification. In the next section, another necessary type of glycosylation will be reviewed.

1.1.2 O-Linked Glycosylation

O-linked glycosylation is another conserved and well-studied process [116]. In contrast to N-linked glycosylation which is a co-translational process, this modification is a real post-translational procedure. O-glycosylation leads to a fully folded glycoprotein maintaining the stability and structure of the produced glycoprotein. In this way, O-linked glycosylation provides the glycoprotein with resistance to an unusual situation such as heat-shock by the proper conforming of the secondary, tertiary and quaternary structures of that protein. Moreover, this mechanism allows the protein to avoid from *aggregation*, a phenomenon in which, because of misfolding, the protein deposits in the cell. A study shows that O-linked glycosylation, in contrast to N-linked glycosylation, modulates enzymatic activity [10]. In addition, it has been demonstrated that O-glycosylation regulates critical glycoprotein hormones [116].

Undergoing the process of O-linked modification, the target protein attaches to the hydroxyl group ($-OH$) of Serine or Threonine amino acids in Golgi complex. There is no clear sequon identified for O-glycosylation, and each protein should be studied individually regarding to O-linked glycosylation sites [116]. Notably, this type of glycosylation is well-conserved in EGF-like proteins , which will be discussed in the next section.

The known types of O-linked glycosylation are *mucin-type*, *O-fucose*, *O-glucose*, *O-GlcNAc*, and *O-arabinose* glycosylation. Mucin-type glycosylation (Figure 4) is the attachment of GalNAc monosaccharide to Serine or Threonine amino acids of a protein. This type is the most common one observed in membrane and secretory proteins of mammals [58]. O-fucose, which has been reported in EGF-like proteins, is the attachment of a glycan to Serine or Threonine of EGF-like domain through *fucose*, a simple monosaccharide [57]. O-glucose is a similar type of attachment, but the core molecule attaching to EGF-like domain is glucose. Recognized as a critical type of O-glycosylation, O-GlcNAc is the attachment of GlcNAc to Serine or Threonine of a protein. Most of the nuclear and cytoplasmic proteins are subject to that type of O-linked glycosylation [116]. In addition, it is revealed that O-GlcNAc affects the presence of *phosphorylation*, another PTM in which the phosphate radicals attach to a specific amino acid of the target protein. Whenever O-GlcNAc is available, there is no phosphorylation and *vice versa* [120]. O-arabinose only happens in plants.

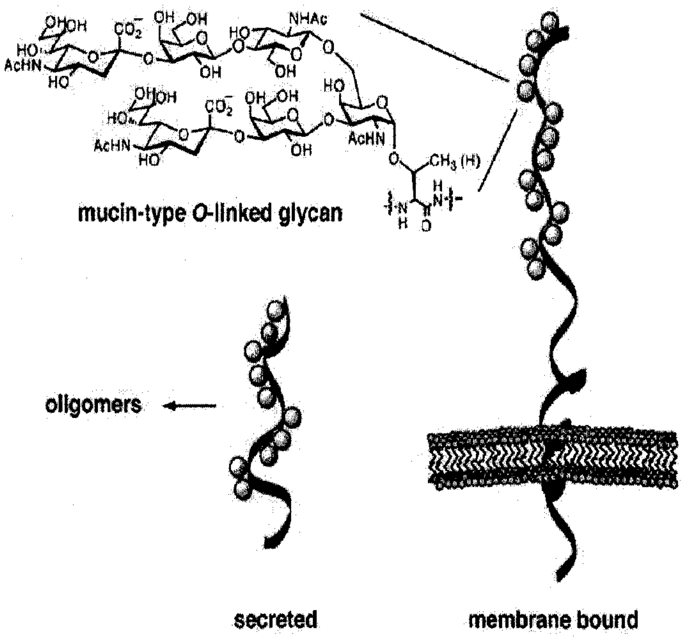


Figure 4: Mucin-type O-linked glycosylation [58]

There are 8 core structures discovered for mucin-type O-glycosylation although it is not proved whether the number of core structures is fixed or not [116].

Aberration in O-linked glycosylation cause severe diseases such as carbohydrate deficiencies and cancer. For example, *MUC EGFL oncoproteins* potentially participate in breast cancer [100]. Furthermore, O-glycosylation plays a major role in neuronal adhesion in brain, so the anomaly in glycosylation sites or glycan structures leads to irreversible brain damages [31, 32].

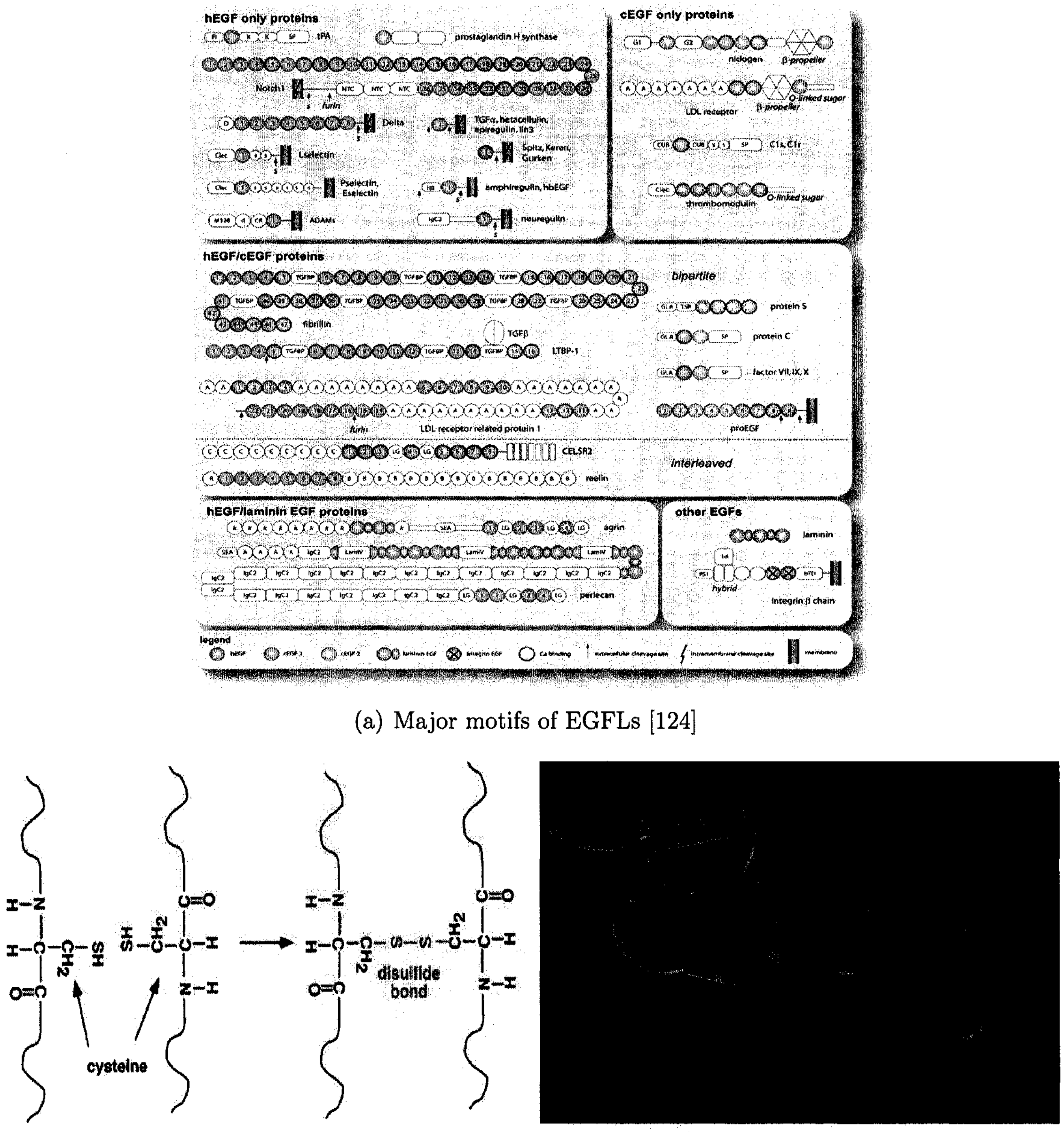
1.2 The Biology of Epidermal Growth Factor-Like Proteins (EGFLs)

Epidermal Growth Factors-Like repeats (Figure 5) are 30 to 40 amino acids domain containing of six conserved *Cysteines* amino acids. Cysteines make a three covalently attached sulfur-to-sulfur bond, known as *disulfide bridges* (Figure 5(b)), which are necessary for the proper conformation of EGF-like proteins [71]. Conserved Cysteines in EGF-like peptide are typically named as C_1 to C_6; therefore, the disulfide bridges cross-link $C_1 - C_3$, $C_2 - C_4$ and $C_5 - C_6$, structurally making three loops in EGF-like domains. EGFLs are essential for cell growth, proliferation, cancer formation and wound healing [3, 2]. Moreover, they interact with special membrane-bound proteins called *receptors* [127].

Figure 5(a) shows various kinds of EGF-like domains. About 10 subfamilies form EGF-like superfamily. Evolutionary speaking, EGF-like proteins are divided into human EGFLs, *hEGF*, and complement Clr-like or *cEGF* proteins. Clr-like EGF proteins exist in virus POX glycoproteins [124].

EGF-like proteins increase the affinity between receptors (*dimerization*) and makes the receptors to release Tyrosine; consequently, Tyrosine signaling initiates the process of proliferation in the cell. Genetic aberration of EGFL signaling causes critical diseases such as carcinomas. The wrong signals may also up- or down-regulate the growth factors, which in turn participate in tumor formation [60].

EGFLs are membrane proteins, and they are naturally glycosylated. The glycosylation of EGF-like proteins is carried out in the luminal side of the rough endoplasmic reticulum (RER) membrane. It has been studied that the glycosylation process in EGFLs alters in various cancers such as lung, brain and melanoma. Moreover, it has been suggested that the alteration of glycosylation sites in the sequences of the proteins of cell membrane responsible for intercellular



(a) Major motifs of EGFLs [124]

(b) Disulfide bonds in EGFLs

(c) EGFL Secondary Structure (gi|16975128)

Figure 5: Hallmarks of Epidermal Growth Factor-Like (EGFL) repeats

interaction, such as *Notch* and *Delta*,may cause abnormality in cell fate decision [60, 7, 107].

Glycosylation in EGF-like proteins is different from that of other subfamilies. In contrast to mucin-type O-linked glycosylation which has no well-defined consensus pattern, EGF-like domains have two O-glycan conserved sequons. Having recently been discovered, the following sequons are the known putative consensus patterns for EGF-like proteins O-glycosylation:

$$Cys_2 - X_{4-5} - (Ser|Thr) - Cys_3 \qquad \text{O-fucose}$$

$$Cys_1 - X_{1-2} - Ser - X - Pro - Cys_2 \qquad \text{O-glucose}$$

where Cys_1 to Cys_6 are conserved Cysteine amino acids in EGFLs and X is any amino acid [79, 57]. X_{4-5} refers to a chain of 4- or 5-residue of any arbitrary amino acid and so does X_{1-2} to 1- or 2-residue one . Dissimilar to N-linked glycosylation, Proline (Pro) also contribute to the consensus pattern of O-glucose. Considering the necessary consensus patterns, it is possible to detect and predict the N- and O-linked glycosylation in these proteins when appropriate non-deterministic models are applied to this subject. Improper conformation of EGF-like repeats is one of the main reason behind prostate cancer in men. Furthermore, glycosylation sites of EGF-like domains heavily altered in other carcinomas such as liver, bladder, renal, colon and gastric cancer [98]. Embodying the obvious flag for tumor growth, EGF-like repeats have been found to be essential for the formation, expansion and fate of brain tumors [54, 55].

EGF-like domains and their glycosylation process have broadly been studied in biomedical laboratories, as reviewed on the latest breakthroughs in the previously discussed sections [116]. Quantitative analysis of glycoproteins have recently been noticed as an effective tool in studying the concepts and principles of protein glycsoylations [109]. Machine learning tools encompass the context of glycosylation as well, and there are several applications developed or being developed to help biologists to better understand that complicated context.

1.3 The Application of Soft Computing Methods in Protein Glycosylation

Broadly utilized as effective means to solve many biological problems, *soft computing techniques* or *computational intelligence methods* can unravel the highly complex mechanism of post-translational modification whose non-linear nature limit them to be solved analytically [68]. The soft computing approaches used in optimization problem can be divided into *computational*, *statistical* and *metaheuristic* frameworks. Computational methods optimize the learning algorithm by predicting the future state of a solution based on the past evaluation of data in both supervised and unsupervised ways. Artificial neural networks are obvious examples of such schemes. Statistical learning theory emphasizes the statistical methods used for automated learning; for example, kernel-based methods such as support vector machines find an optimal separating hyperplane by mapping data to a higher dimensional search space [117]. Metaheuristic algorithms are usually used to find a best solution through other combinatorial optimizers; in fact, metaheuristic approaches suggest a framework for other heuristic frameworks. Fuzzy systems and genetic algorithms are examples of metaheuristic methods [126, 44].

The application background of computational intelligence methods in the context of protein glycosylation can be divided into three main research areas:

Glycan Structure Analysis and Prediction. Providing tools for predicting and analysis of different types of carbohydrate chains attaching to a target protein.

Glycan and Glycoprotein Knowledgebase and Ontology. Modeling a comprehensive ontology for describing of the biological properties of complex carbohydrates

Glycosylation Sites Detection and Prediction. Detecting and predicting the distribution of the number of glycosylation sites in biologically interested proteins

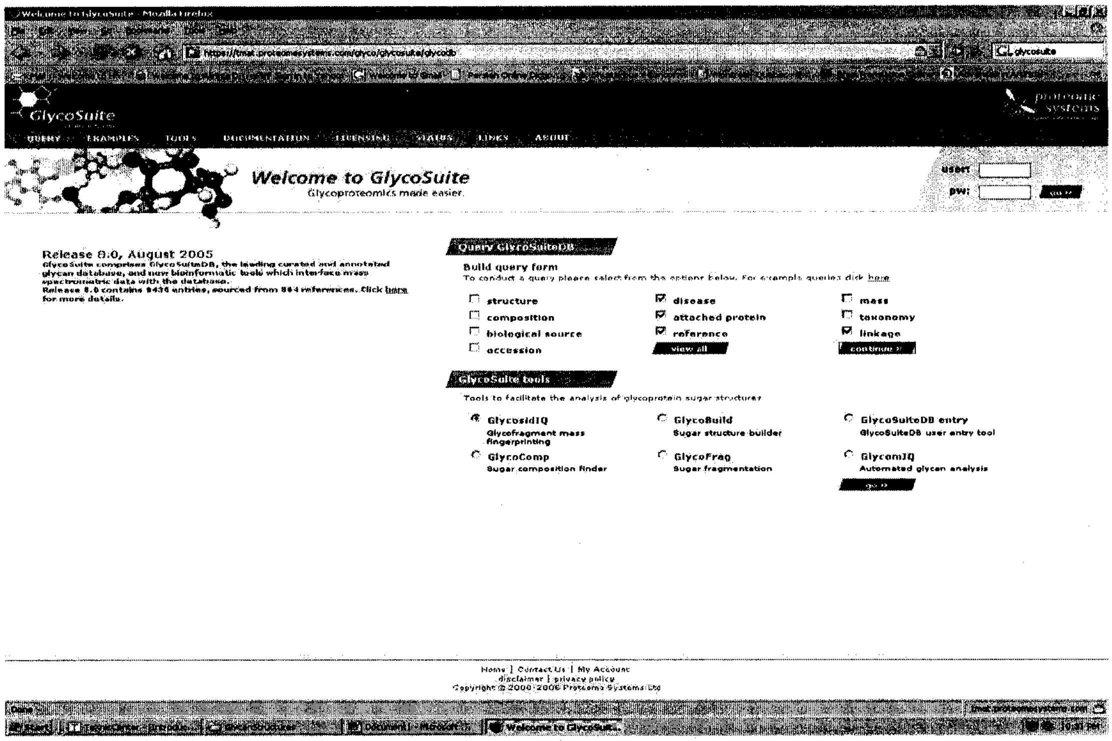
1.3.1 Glycan Structure Analysis and Prediction

Glycans have various structures, and this variety leads to different forms of glycoproteins. To recognize the pattern of those structure, biologists benefit from the structural databases available to deposit glycan structures [27]. For example, GlycoSuiteDB [26] is the database containing reported glycan structures. This database is updated from time-to-time to make sure of encompassing all available glycan structures. Moreover, there are useful tools developed by GlycoSuiteDB curators to manipulated the stored glycan structures.

CAZy (Carbohydrate Active enZYmes) [28] is a data bank of structurally related carbohydrate of enzymes. The conformational information of each carbohydrate-active site has been curated and annotated in this database.

Glycosciences.de [82] is a comprehensive and well-integrated collection of suites as well as databases necessary to explore in glycan structures. Developed by German Cancer Research Center at Heidelberg, Germany, Glycosciences.de has several relational databases including CSS (Carbohydrate Structure Suite) [83], which automatically mine the 3D structure of polysaccharides through Protein Data Bank (PDB) [62], and GlyProt [20] which accepts the fundamental parameters for geometrical position of glycosylations sites (torsion angle, anomeric data, etc.) and returns the *in-silico* built glycoprotein. This integrated database contains 3920 N- and O-linked glycan structures.

One major challenge for glycan structures is to have a unified standard to exchange and share glycan structures files. Cooper *et al* [25] have reviewed the data standardization for GlycoSuiteDB.



Welcome to GlycoSuite
Glycoproteomics made easier.

Release 8.0, August 2005
GlycoSuite comprises GlycoSuiteDB, the leading curated and annotated glycan database, and new bioinformatic tools which interface mass spectrometric data with the database.
Release 8.0 contains 9436 entries, sourced from 864 references. Click here for more details.

Query GlycoSuiteDB

Build query form
To conduct a query please select from the options below. For example queries click here

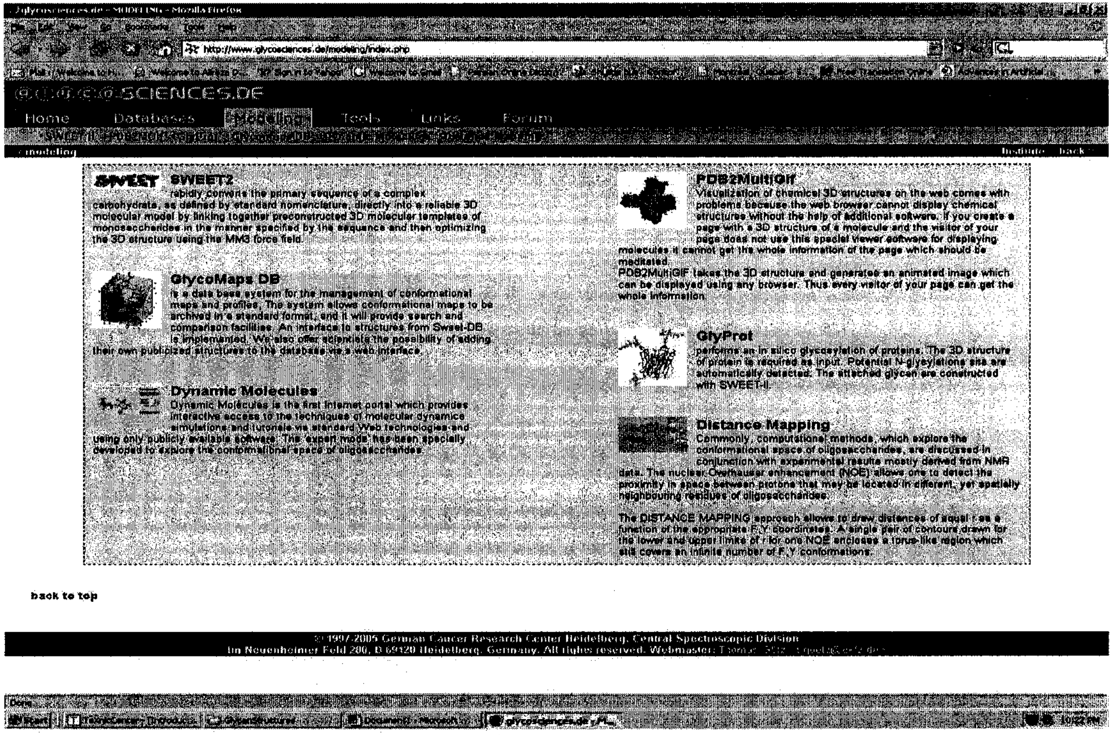
□ structure	☑ disease	□ mass
□ composition	☑ attached protein	□ taxonomy
□ biological source	☑ reference	☑ linkage
□ accession	view all	continue >>

GlycoSuite tools
Tools to facilitate the analysis of glycoprotein sugar structures

● GlycosidIQ Glycofragment mass fingerprinting	○ GlycoBuild Sugar structure builder	○ GlycoSuiteDB entry GlycoSuiteDB user entry tool
○ GlycoComp Sugar composition finder	○ GlycoFrag Sugar fragmentation	○ GlycomIQ Automated glycan analysis
		go >>

Home | Contact Us | My Account
disclaimer | privacy policy
Copyright © 2000-2006 Proteome Systems Ltd

(a)



SWEET2
rapidly converts the primary sequence of a complex carbohydrate, as defined by standard nomenclature, directly into a reliable 3D molecular model by linking together preconstructed 3D molecular templates of monosaccharides in the manner specified by the sequence and then optimizing the 3D structure using the MM3 force field.

GlycoMaps DB
is a data base system for the management of conformational maps and profiles. The system allows conformational maps to be archived in a standard format, and it will provide search and comparison facilities. An interface to structures from Sweet-DB is implemented. We also offer scientists the possibility of adding their own publicized structures to the database via a web interface.

Dynamic Molecules
Dynamic Molecules is the first Internet portal which provides interactive access to the techniques of molecular dynamics simulations and tutorials via standard Web technologies and using only publicly available software. The expert mode has been specially developed to explore the conformational space of oligosaccharides.

PDB2MultiGif
Visualization of chemical 3D structures on the web comes with problems because the web browser cannot display chemical structures without the help of additional software. If you create a page with a 3D structure of a molecule and the visitor of your page does not use this special viewer software for displaying molecules it cannot get the whole information of the page which should be medikated.
PDB2MultiGIF takes the 3D structure and generates an animated image which can be displayed using any browser. Thus every visitor of your page can get the whole information.

GlyProt
performs an in silico glycosylation of proteins. The 3D structure of protein is required as input. Potential N-glycylations site are automatically detected. The attached glycan are constructed with SWEET-II.

Distance Mapping
Commonly, computational methods, which explore the conformational space of oligosaccharides, are discussed in conjunction with experimental results mostly derived from NMR data. The nuclear Overhauser enhancement (NOE) allows one to detect the proximity in space between protons that may be located in different, yet spatially neighbouring residues of oligosaccharides.

The DISTANCE MAPPING approach allows to draw distances of equal r as a function of the appropriate F,Y coordinates. A single pair of contours drawn for the lower and upper limits of r for one NOE encloses a torus-like region which still covers an infinite number of F,Y conformations.

back to top

© 1997-2005 German Cancer Research Center Heidelberg, Central Spectroscopic Division
Im Neuenheimer Feld 280, D 69120 Heidelberg, Germany. All rights reserved. Webmaster: Thomas Lütteke

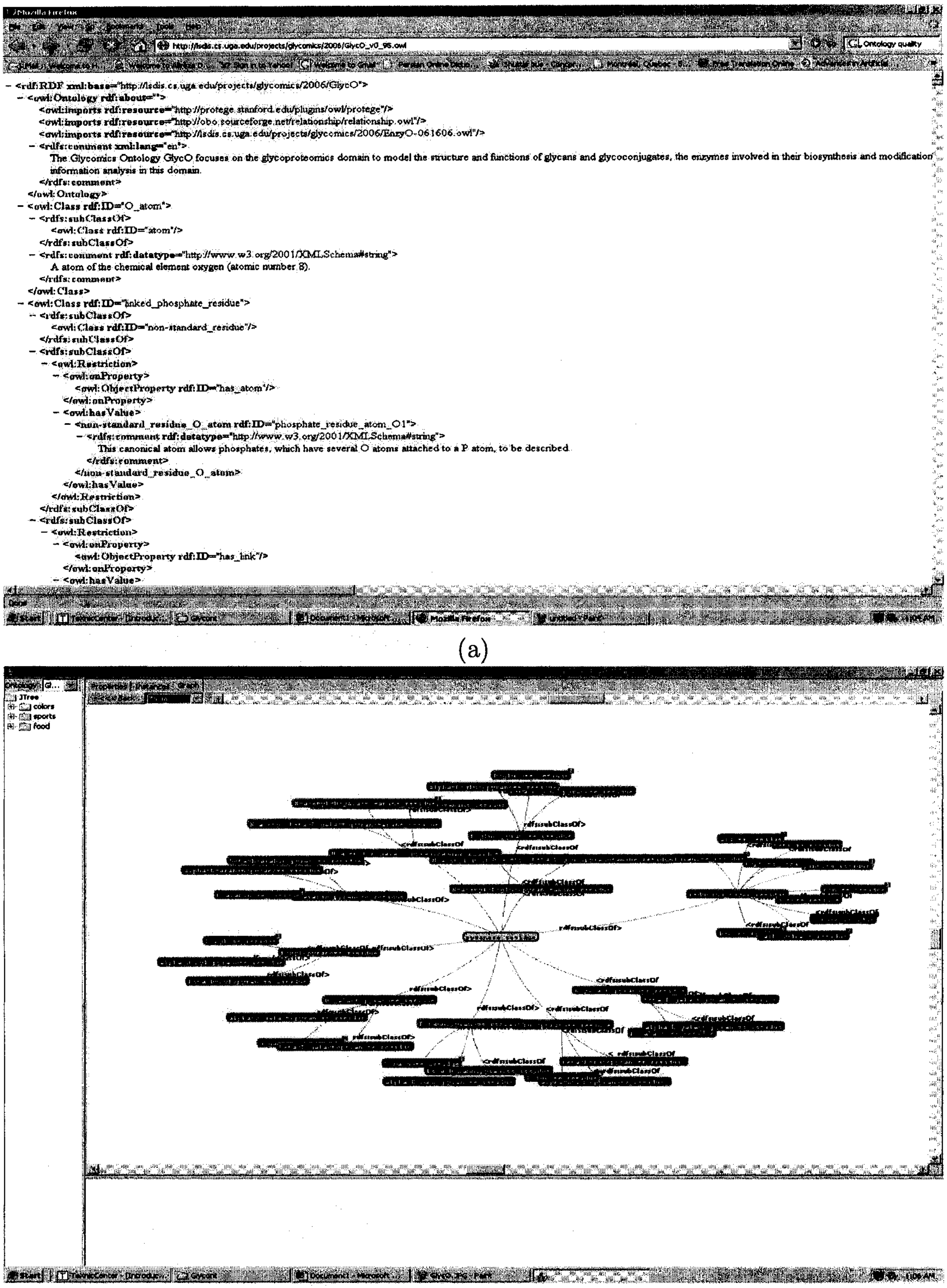
(b)

Figure 6: (a) GlycoSuiteDB (b) Glycosciences.de

1.3.2 Knowledge Representation and Ontology for Glycans and Glycoproteins

Glycosylation is a complex enzymatic process; therefore, biologists need an ontology tool to facilitate access, share and represent the proper description of carbohydrates annotation. Besides, it is necessary to develop an ontology as large as possible to retrieve the different data from the Web. GlycOViz [106] is an ontology browser as an integrated environment to edit and build ontologies for glycans and glycoproteins. GlycO (Glycomics Ontology) [106] is domain ontology with over 770 classes and is used to classify the web services describing the characteristics of glycans. Figure 7 shows snapshots of each ontology tool.

Since some of web contents are redundant, the conflict problem may happen to the knowledge representing of ontology at semantic level. The conflict occurs when there is different interpretation of problem domain, thereby leading to inconsistency in the extracted data. Arpinar *et al* have reviewed conflict data in ontology and have proposed a rule-based approach to overcome conflicts [6].



Mozilla Firefox
http://lsdis.cs.uga.edu/projects/glycomics/2006/GlycO_v0_95.owl
Ontology quality

```
- <rdf:RDF xml:base="http://lsdis.cs.uga.edu/projects/glycomics/2006/GlycO">
  - <owl:Ontology rdf:about="">
      <owl:imports rdf:resource="http://protege.stanford.edu/plugins/owl/protege"/>
      <owl:imports rdf:resource="http://obo.sourceforge.net/relationship/relationship.owl"/>
      <owl:imports rdf:resource="http://lsdis.cs.uga.edu/projects/glycomics/2006/EnzyO-061606.owl"/>
    - <rdfs:comment xml:lang="en">
        The Glycomics Ontology GlycO focuses on the glycoproteomics domain to model the structure and functions of glycans and glycoconjugates, the enzymes involved in their biosynthesis and modification
        information analysis in this domain.
      </rdfs:comment>
    </owl:Ontology>
  - <owl:Class rdf:ID="O_atom">
    - <rdfs:subClassOf>
        <owl:Class rdf:ID="atom"/>
      </rdfs:subClassOf>
    - <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        A atom of the chemical element oxygen (atomic number 8).
      </rdfs:comment>
    </owl:Class>
  - <owl:Class rdf:ID="linked_phosphate_residue">
    - <rdfs:subClassOf>
        <owl:Class rdf:ID="non-standard_residue"/>
      </rdfs:subClassOf>
    - <rdfs:subClassOf>
      - <owl:Restriction>
        - <owl:onProperty>
            <owl:ObjectProperty rdf:ID="has_atom"/>
          </owl:onProperty>
        - <owl:hasValue>
          - <non-standard_residue_O_atom rdf:ID="phosphate_residue_atom_O1">
            - <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
                This canonical atom allows phosphates, which have several O atoms attached to a P atom, to be described.
              </rdfs:comment>
            </non-standard_residue_O_atom>
          </owl:hasValue>
        </owl:Restriction>
      </rdfs:subClassOf>
    - <rdfs:subClassOf>
      - <owl:Restriction>
        - <owl:onProperty>
            <owl:ObjectProperty rdf:ID="has_link"/>
          </owl:onProperty>
        - <owl:hasValue>
```

(a)

(b)

Figure 7: (a) GlycO (b) GlycOViz

1.3.3 Glycosylation Sites Detection and Prediction

Blom *et al* [18] have well reviewed the topic of glycosylation sites detection and prediction from amino acid sequences. The major detection and prediction resources have been shown in Table 1.

Table 1: Existing Glycosylation Site Predictors

Predictor	**Description**	**Reference(s)**
NetOGlyc	Predicts mucin-type O-glycosylation in mammals	[102]
NetNGlyc	Predicts N-glycosylation in human proteins	[70]
YinOYang	Predicts O-GlcNAc O-glycosylation eukaryotes	[50]
Big-Π	Predict GPI-anchor in a protein	[114, 34]
DictyOGlyc	Predicts O-GlcNAc O-glycosylation in *Dictyostelium discoideum* proteins	[51]

NetNGlyc [102] is a web-based predictor for N-linked glycosylation sites using neural networks. Although the system reported 86% of glycosylated sites of O-GlycBase [48], the authors have not explained the learning method and the type of the neural network they have used. One site for mucin-type O-linked glycosylation is NetOGlyc [70], which could correctly identify 76% of O-linked glycosylation sites. The system uses two-layer network with error backpropagation learning algorithm. YinOYang [50] is a detector of *yin-yang* sites, the amino acids which can be attached through either phosphorylation or O-GlcNAc glycosylation using neural network. The server also uses NetPhos [17] server to detect the possible phosphorylation sites if those sites are also O-linked glycosylated. Big-Π [114, 34] predicts *glycosylphosphatidylinisotol* (GPI) anchored glycosylation sites. Using a jury of neural network with backpropagation algorithm for each network, Gupta *et al* have developed DictyOGlyc [51] to predict O-GlcNAc O-linked glycosylation sites in *D. discoideum*, which is an appropriate model paradigm for glycobiologists.

Other approaches such as mathematical modeling as well as statistical analysis of glycosylation sites have also been employed [75, 115, 5, 101, 94, 22].

1.4 Feed Forward Neural Networks

Artificial neural network (ANN) is a mathematical model of biological neurons of the human brain. An ANN simulates the characteristics of its biological counterpart for correlating or associating meaningful concepts. For example, it may correspond negative and positive numbers to black and white colors respectively (*pattern classification*). On the other hand, an ANN can keep track of a certain pattern in a protein sequence by memorizing the previously introduced examples (*pattern association*). Furthermore, they process the information coming as a *vector* or *batch* of input and produce a vector or batch of output which is an estimation of the *target* value(s). Neural networks consist of interconnected units, known as *neurons* or *nodes*. Each connection between two neurons is tagged by a number called *weight*. One can biologically interpret the weight between two neurons as helping to trigger a neuron to response or *fire* an output. The neurons and the connections among them forms the network's *architecture*. If all neurons of a network are connected to each other, the network is called an *ergodic* or *fully connected* network otherwise a *semi-connected* one. The algorithm being capable of adjusting the weights to minimize the difference between the target and observed values (*error*) is called *learning* or *training* algorithm [36, 15, 30].

The structure of a biological neuron provides a general model of an artificial neuron. A biological neuron (Figure 8) includes *dendrites*, cell body or *soma*, and *axon.* Soma is a hard body comprising the nucleus and other nuclear materials. Around soma, there are short extensions known as dendrites. Dendrites connect a neuron to another one. The interconnection area is called *synapse.* Axon is the longer extension ending from one side to soma and from other side to *terminal branches.* Terminal branches are micro-sensors, the cell body sends the electrochemical pulse generated by an external stimulator to the terminal branches through an axon [38].

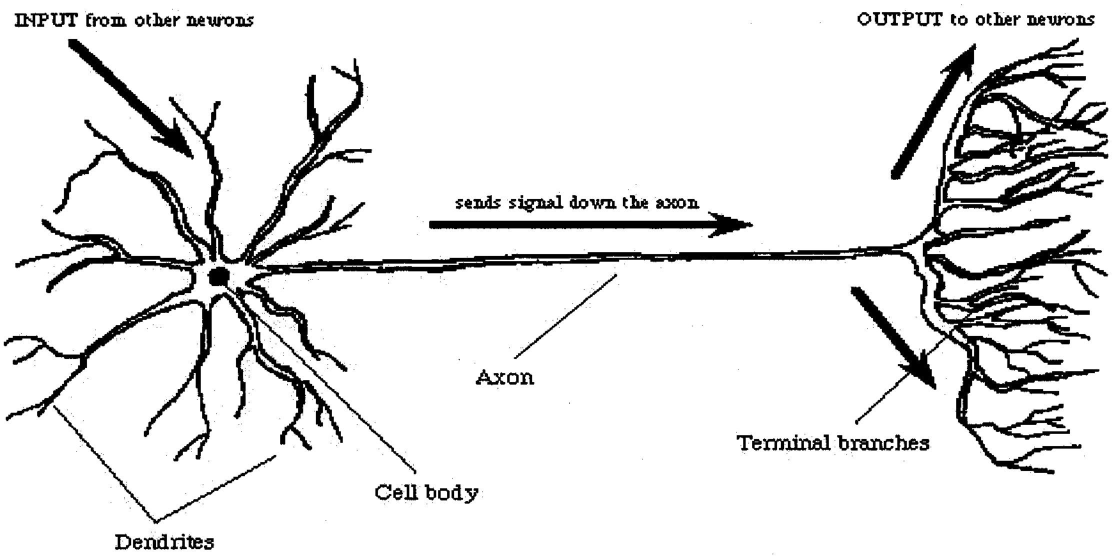


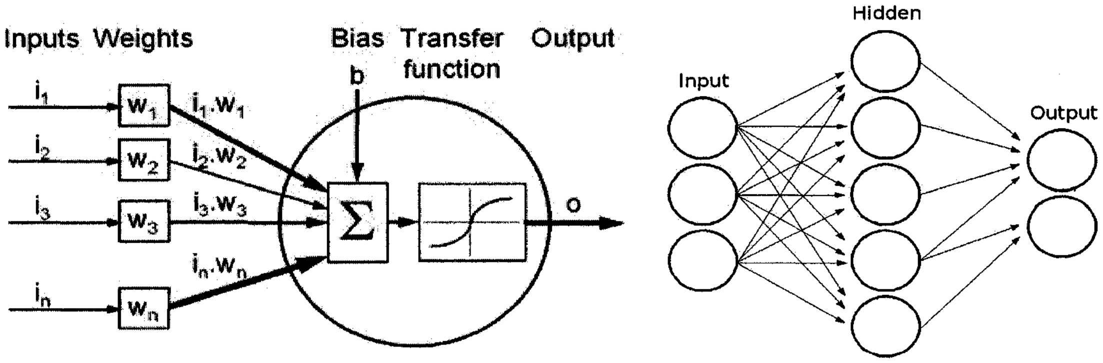
Figure 8: Biological neuron

A *McCulloch-Pitts neuron* [91] is an artificial model of biological neuron which is simply a linear function summing up the multiplication of all inputs by their corresponding weights (Figure 9(a)). After that, an *activation* or *transfer* function receives the summation and evaluates whether it is larger than a specific *threshold* value or not. If the summation is larger than threshold, it returns either the value of summation or any other pre-defined output; otherwise, it returns zero. Depending on the form and nature of target values, artificial neurons usually incorporate *logistic* functions:

$$f(x) = \frac{1}{1+e^{-\sigma x}} \quad ,\sigma > 0 \qquad \text{Sigmoid function} \tag{1}$$

$$g(x) = \frac{1-e^{-\sigma x}}{1+e^{-\sigma x}} \quad ,\sigma > 0 \qquad \text{Tangent sigmoid function} \tag{2}$$

The transfer function should be continuous and differentiable and logistic functions satisfy in those conditions. The artificial neuron is the building block of the feed-forward neural network.



(a) Artificial Neuron

(b) A multi-layer feed-forward network

Figure 9: Multi-layer perceptron and the structure of its neuron

Minsky and Papert have shown that the feed-forward networks with only input and output layer, so-called *perceptrons*, are incapable of solving a class of problems known as *linear separable* ones [92]. For example, a single-layer network cannot solve an XOR problem because XOR problem is not linearly separable. As shown in Figure 9(b), a feed-forward neural network may have a layer consisting of artificial neurons between input and output layer, so called a *hidden layer*. The neurons of the hidden layer are sometimes called *hidden units*. It has been demonstrated that a multi-layer neural network with an arbitrary number of hidden units can estimate any non-linear approximation [65, 64].

In the process of training, the weights of the network are adjusted so that they produce a vector of output having the minimum difference with target values. The most popular learning algorithm for feed-forward weight training is called *error back-propagation* algorithm [105, 119]. In standard backpropagation, the network first computes the weight updates for neurons at the output layer, and then the error between the output and that target value is *propagated backward* to the hidden and input layer [13].

1.5 Objective and Motivation

There are considerable studies on the mechanism of post-translational modifications, especially protein glycosylation. Glycoproteins are important by-product of PTMs. Accordingly, glycosylation is a research topic of interest.

Buskas *et al* have counted some of the recent advances of glycoprotein engineering and its application in drug discovery process [21]. The available applications to predict glycosylation sites run an artificial neural network over the amino acid sequences to obtain biologically meaningful response. Those techniques use standard backpropagation or jury of networks and consider various families of proteins. However, few studies have reported on the distribution of glycosylation sites in a specific superfamily [22, 69]. An examination of the previous literature encouraged for an extension of the topic to one specific protein family. Epidermal Growth Factor-like protein sequences were the case of this study. As reviewed, EGF-like peptides have a particular glycosylation process, and they contain distinct sequons for O-linked glycosylation. Furthermore, their O-fucose and O-glucose patterns have been well studied [57]. As a result, EGF-like domains are suitable candidates for evaluation of their glycosylation sites pattern. While most of the predictor servers for glycosylation sites utilize standard backpropagation with their neural networks, this study examines other learning paradigms and also introduces Bayesian learning to glycosylation sites detection and prediction problems.

The purpose of this research was to determine the non-linear correlation between EGF-like domain sequences and the distribution of glycosylation sites using feed-forward neural networks. The study emphasized different learning technique fitting best to the protein datasets. In addition, it investigates the dependency between the prior knowledge and prediction based on Bayesian inference.

The results of this study may help glycobiologists to classify and choose the EGF-like protein sequences of interest. They may also suggest further research into the importance of *in silico* detection and prediction of glycosylation sites within the procedures of glycoprotein engineering and drug discovery.

Chapter 2

Materials and Methods

This chapter describes the procedural steps applied in this study and the materials employed at each step. The nature and compilation of the data will be illustrated in the first section. Following that section, the learning algorithms applied to the subject will be explained. Finally, the concerned Bayesian framework set up for this study will be demonstrated.

2.1 Data Specification

The EGF-like protein sequences were collected from PROSITE [66], UniProt [4, 8], SWISS-PROT [19], InterPro [95], Mouse Tumor Biology Database (MTBD) [96] and a lung cancer database [99]. The EGF-like oncoproteins, the proteins potentially associated with malignancy and glycosylated in the form of O-fucose and O-glucose, were also gathered [56, 57], *vide* Appendices E and F. The sequences with 'Potentially' and 'by Similarity' annotation tags were also included to enrich the datasets. Those sequences indicate the putative glycosylation sites in EGF-like domains. Pfam [39] is a repository containing related families of gathered proteins into clans. Pfam was run through the EGF-like protein sequences to find the most similar sequences. To avoid redundancy in the the final dataset as well as keep the dataset unbiased, the sequences with more than 80% of similarity were ignored. Concluded from the series of experiments between 1971-75, Günter Blobel postulated that secretory proteins have intrinsic short peptides recognizable by cell membranes in the mechanism of secretion (*signal hypothesis*). Consequently, the glycosylated proteins need to have those peptides, so-called *signal peptides*, to pass through the secretory pathway [16, 39]. Biologically speaking, the EGF-like sequences are membrane proteins, thereby containing signal peptides. Accordingly, the EGF-like sequences without reported signal peptide were also removed from the dataset.

To study the eventual effects of prior knowledge on model response, the protein sequences were partitioned in various lengths or *window frames.* Window frame was a single window centered on the glycosylation site. Table 2 introduces the window frames of this study. $X_{GlycoSite}$ represents the glycosylation site. X is any arbitrary amino acid. The multipliers (2, 5, 7, 9 and 14) indicate the number of amino acid around the glycosylation site. To avoid over- or under-feeding of the network, the window frames in this study were restricted to be 5-, 11-, 15-, 19- and 29-residue frames.

Table 2: Window Frame Specification

Window Frame	**Sequence Pattern**
5	$2(X) - X_{GlycoSite} - 2(X)$
11	$5(X) - X_{GlycoSite} - 5(X)$
15	$7(X) - X_{GlycoSite} - 7(X)$
19	$9(X) - X_{GlycoSite} - 9(X)$
29	$14(X) - X_{GlycoSite} - 14(X)$

The target set was $\{1|0\}^n$ where n denotes the number of underlying window frames, 1 if the central amino acid in the window frame is glycosylated and 0 if it is not. The intended target set was taken as 0.9 and 0.1 for detected glycosylated and non-glycosylated sites respectively. Using sigmoid functions as a transfer function with binary target values, the learning algorithm of the feed-forward network forces the weights and biases of the network to grow quickly, a phenomenon known as *shifting effect*. The pre-determined values for the target set prevents such phenomenon by restricting the output of the model within an appropriately small range. Table 3 shows the datasets after preprocessing. The total number of the data originally used was 8037 window frames, out of which 7157 were used for training and the remaining 880 window frames for testing the neural network model.

Table 3: Datasets Specification

Dataset	**No. of Protein Sequences**	**No. of Window Frames**	**Target Value**
A	3400	3700 3045	0.9 0.1
B	412	412	0.1
C	880	880	0.9|0.1

The data were divided into the following datasets:

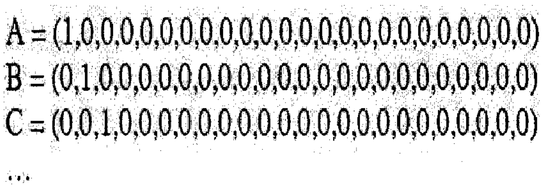
- **Dataset A.** This set consisted of EGFL proteins used for training the feed-forward neural network. At least one and at most three window frames were selected from each sequence. There was no preference in terms of the number of window frames required to be selected from each sequence. The window frames were arbitrarily taken from the middle region of each sequence. This set consisted of 3700 window frames containing glycosylation site and 3045 window frames not containing glycosylation sites (Appendix B).

- **Dataset B.** This set covered the EGFL sequences glycosylated but not shown to be associated with cell malignancy in order to avoid the knowledge coming from abnormal glycosylation sites. One window frame which did not contain glycosylation site was arbitrarily selected from the middle region of each sequence and included with the training set. The number of window frames selected for this set was 412. Similar to Dataset A, there was no preference in choosing of a specific window frame from a sequence (Appendix C).

- **Dataset C.** This set was part of the full data. Once the network trained with Datasets A and B, this set was introduced to the model to determine the performance of the network (Appendix D). The 880 window frames of this set were partitioned into the following subsets:

  - The first 220 window frames contained glycosylation site
  - The next 220 window frames had no glycosylation site
  - The third subset covered 220 window frames, which also contained glycosylation sites
  - The last subset included *noises.* The first 175 window frames were taken from non-EGFL sequences having glycosylation sites, and the last 45 window frames were any arbitrary window frame from any arbitrary sequence.

Orthogonal scheme [80] was used to encode the dataset before feeding to the neural network. In this scheme, each amino acid is represented with binary '1' while others remain '0'. Although this scheme produces sparse input units, it prevents the network to learn a false correlation between amino acids [104]. Figure 10 represents the concept of orthogonal encoding.



A = (1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
B = (0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
C = (0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
...

Figure 10: Orthogonal Encoding

The input to the neural network was the concatenation of the orthogonal encoding for ℓ amino acids in the window frame. Consequently, the input sequence to the network consisted of a 20ℓ-dimensional vector with a sparse binary string of 0s and 1s where ℓ is the length of the window frame.

According to orthogonal encoding scheme and the introduced topology for the feed-forward neural network, the number of parameters for the network was computed:

$$\Gamma = h\,(20\ell + 2) + 1 \tag{3}$$

where Γ is the total number of weights and biases, ℓ is the length of the window frame and h is the number of the hidden units.

The 10-fold cross validation was the chosen approach to improve the generalization of the model. In each fold, from the total number of the window frames in Datasets A and B, 10% were selected as test set and 90% as training set. The 10% of each of Datasets A and B were picked up sequentially. Consequently, there were 715 window frames in the test set, which 370 window frames contained glycosylation site (Dataset A), 305 window frames did not contain glycosylation site (Dataset A), and 40 window frames were taken from Dataset B. After that, the window frames in the test set were uniformly distributed. The first 92 window frames containing glycosylation sites (Dataset A), the second 76 window frames not containing glycosylation sites (Dataset A) and the third 10 window frames taken from Dataset B formed the first partition of the test set. The second and third partitions were also formed the same way as explained. The last partition consisted of 94 window frames having glycosylation sites, 77 window frames not containing glycosylation sites and 10 window frames taken from Dataset B. The choice of such partitioning was arbitrary.

2.2 Feed-Forward Networks: Mathematically Revisited

The intention of this section is to formulate feed-forward neural networks as well as the learning rule governing it.

2.2.1 An Overview

Let ℓ be the number of labeled observations, and (X_i, y_i) where $X_i \in \Re^n$ is the vector of labeled data in which $i = 1, \ldots, \ell$. This vector is called *training set.* $y_i \in \{-1, 1\}$ is the set of target values, or *target set*, which classifies the problem or phenomenon.

Definition 2.2.1. *Let $P(X, y)$ is a distribution with abstract parameters θ from which training set is generated. $f_\theta : X_i \rightarrow y_i$ is called a hypothesis over the training set.*

Definition 2.2.2. *Let $\mathcal{H} = \{f(X_i, \theta), \Theta\}$ be a hypothesis space for training data. Θ is the set of parameters of the distribution function.*

Definition 2.2.3. *If $\exists \theta^* \in \Theta$ such that:*

$$\forall \theta \in \Theta : |f(X_i, \theta) - f(X_i, \theta^*)| \leq \varepsilon \tag{4}$$

then $f(X_i, \theta^*)$ is called a trained machine over the hypothesis space $\mathcal{H}$. θ^* is the set of final trained parameters using in prediction of unseen data.

Definition 2.2.4. *A trained machine has an expected test error such as $R(\theta)$, where*

$$R(\theta) = E \left\| Y_i - f(X_i, \theta) \right\| \tag{5}$$

Definition 2.2.5. *The mean error rate of a trained machine output, training error, such as $R_{emp}(\theta)$ is the emperical error of the trained machine and is defined as:*

$$R_{emp}(\theta) = \frac{1}{\ell} \sum_{i=1}^{\ell} [Y_i - f(X_i, \theta)]^2 \tag{6}$$

$R_{emp}(\theta)$ approaches to Equiation 5. There is an upper bound for $R_{emp}(\theta)$ so that it theoretically equals to $R(\theta)$. According to Vapnik-Chervonenkis (VC) theorem [117], there is an upper bound with certain probability for the expected test error. Glivenko-Cantelli theorem [43] implies the error of a model running infinitely approaches to a probability distribution; as a result, it is simply the estimation of the error distribution with computed parameters. For more details on function superimposition and learning, see [73] and [29].

2.2.2 The Complexity of the Feed-forward Neural Network

Karpinski and MacIntyre [72] have shown that VC dimension for a single-output feed-forward network has an upper bound with an order of magnitude $O(H^2W^2)$, where H is the number of hidden neurons and W is the number of parameters of the network. The hidden-network selected in this study was a single-layer network, so $W \sim O(NH)$. Consequently, the upper bound had an order of magnitude $O(N^2H^4)$ as the complexity value. According to Equation (3), the complexity of the neural network depended on the length of the protein sequences (ℓ). It was the major reason to restrict the length of the sequences to a maximum pre-determined value. The detailed discussion on feed-forward complexity can be found in [11, 74, 121, 84].

2.3 The Learning Approaches Applied

Figure 11 illustrates the workflow taken for this study. Data normalization was a step in which encoding and scaling of the dataset took place. To improve the generalization of the network, *early stopping* and *regularization* were used. According to early stopping, the dataset was divided into three subsets of training, validation or test, and verification. Subsequently, in each training step, or *epoch*, the network was trained with training set and validated with test set. Verfication sets were used to determine the performance of the network after generalization process. Regularization will be discussed in Section 2.4.

Table 4 shows the learning algorithms incorporated with the feed-forward network. In addition to using early stopping and regularization terms, the values for the intrinsic parameters of the algorithm, *hyperparameters*, were chosen in such a way to prevent the *overfitting* phenomenon during the training. For example, the values of backtracking minimization parameters for [BPROP-OSS] were set to $\alpha = 0.001$, $\beta = 0.1$ and $\gamma = 0.1$. The *threshold goal*, a criterion for ending the training epochs, varied between 10^{-2} and 10^{-5}.

When the most consistent algorithm was found for the data, it was used against Bayesian automated learning for comparison between two model paradigms. Appendix A shows the results of training as well as the model response for each algorithm. The interpretation of the results will be discussed in Chapter 3.

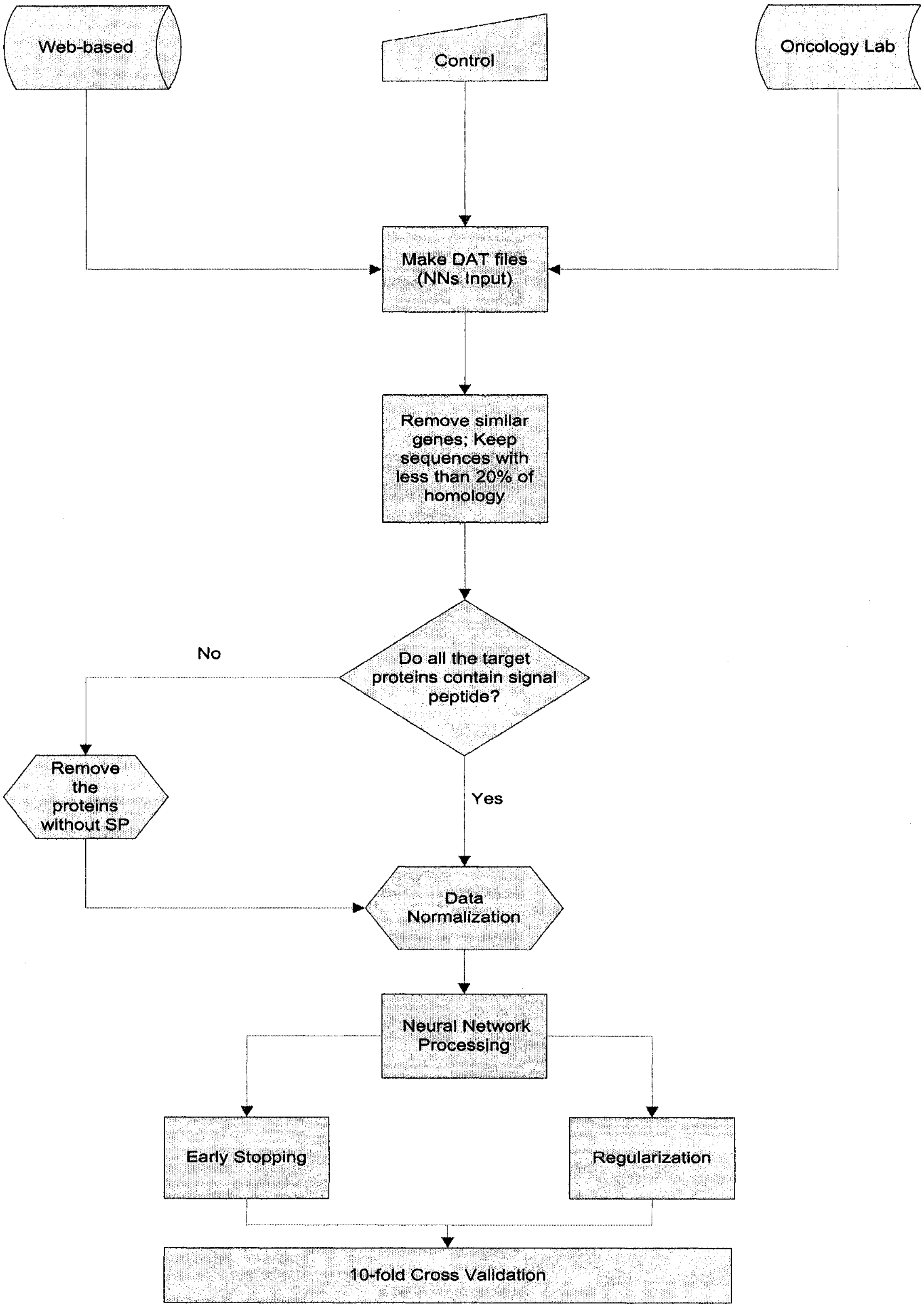

Web-based
Control
Oncology Lab
Make DAT files (NNs Input)
Remove similar genes; Keep sequences with less than 20% of homology
Do all the target proteins contain signal peptide?
No
Remove the proteins without SP
Yes
Data Normalization
Neural Network Processing
Early Stopping
Regularization
10-fold Cross Validation

Figure 11: The workflow for phase I of this study

Table 4: The learning algorithms applied in this study

Label	Learning Algorithm	Reference(s)
[BPROP-GD]	Backpropagation Gradient Descent	[105]
[BPROP-GD-M]	Backpropagation Gradient Descent with Momentum	[36, 15]
[BPROP-GD-AL]	Backpropagation Gradient Descent with Adaptive Learning rate	[36, 15]
[BPROP-GD-X]	Backpropagation Gradient Descent with Adaptive Learning rate and Momentum	[36, 15]
[BPROP-RPROP]	Resilient Backpropagation	[103]
[BPROP-CGF]	Conjugate Gradient Descent with Fletcher-Reeves Updates	[108]
[BPROP-SCG]	Scaled Conjugate Gradient Descent Backpropagation	[93]
[BPROP-OSS]	One Step Secant Backpropagation	[12]
[BPROP-BRNN]	Backpropagation with Bayesian Automated Regularization	[86, 85]

2.4 Bayesian Learning: Induction and Inference

There are usually three steps in an inductive learning:

- Observing a phenomenon
- Making a model from observation
- Predict outcomes based on the learned model

Therefore, machine learning *automates* the process of learning, and the theory of learning establishes a solid *formalization* on top of that automated process. Hence, defining a learning machine from the scratch is a useful approach to implement a practical tool for induction. Bayesian inference is a consistent method which can be applied to extract the required knowledge using a systematic induction approach. A neural network can be traditionally trained by either steepest descent algorithms or other optimizers such as conjugate gradients [53]. It is more accurate if the structural parameters of a network are also taken into account while computing training errors to penalize large weights, a mechanism called *regularization* [24]. This can be done by incorporating a regularization term to Equation (6):

$$R_{emp}(\theta) = \frac{1}{\ell} \sum_{\ell}^{i=1} [y_i - f(X_i, \theta)]^2 + \lambda \sum_{j=1}^{(n+d)H} w_j^2 \tag{7}$$

where the second term on the right shows the sum of squared weights (SSW) of a neural network. $n+d$ is the sum of all the weights including biases. λ is *regularization coefficient* and should be set up such a way that the total performance of network increases. SSW is the *regularization term* to penalize the large values of the weights in the network. One approach is to use a simple Genetic Algorithm (sGA) to determine the regularization coefficient [90]. One benefit of this method is to find the local optima of the ratio simultaneously. On the other hand, this method is too slow to fit into implementation, and as more data arrive, it is computationally complex.

Another approach is to set up neural network as a probabilistic model. The non-linear inner-product space of a neural network's parameters is assumed to be related to a probability distribution function with unknown variances, and the goal is to estimate the parameters of the network based on that function [86, 85]. Since these models adjust a network without the need of additional test set, they are able to automatically optimize the regularization ratio; consequently, they provide an *automated regularization* to a neural network.

Let both terms in Equation (7) receive different coefficients representing the importance of each term. Equation (7) can be re-written as follows:

$$R_{emp}(\theta) = \beta E_D + \alpha E_R = \beta \frac{1}{\ell} \sum_{\ell}^{i=1} [y_i - f(X_i, \theta)]^2 + \alpha\lambda \sum_{j=1}^{(n+d)H} w_j^2 \tag{8}$$

According to Bayes' theorem, general knowledge of every model can be formulated as probability distribution (density). In terms of neural networks, *posterior probability* of weights given a model, D, is integrated out using *prior distribution* of weights as well as measuring noise process model on target set or *likelihood function* given certain set of weights. To make sure posterior distribution remains a probability function, those measurements are normalized by distribution of model itself.

$$\mathcal{P}(W|D) = \frac{\mathcal{P}(D|W)\mathcal{P}(W)}{\mathcal{P}(D)} \tag{9}$$

The goal is to find best weights, W^*, that maximize posterior probability. To find a practical algorithm which could calculate Equation (9), prior, likelihood and posterior probabilities were computed individually.

2.4.1 The Prior Probability

Let the network choose a zero-mean standard Gaussian distribution, $N(0, \sigma^2)$, where $\sigma^2 = \frac{1}{\alpha^2}$.

$$\begin{aligned} \mathcal{P}(W|\alpha) &= \prod_{i=1}^{W} \mathcal{P}(w_i|\alpha) \\ &= \frac{1}{Z_W(\alpha)} e^{-\alpha E_W} \\ Z_W(\alpha) &= \left(\frac{2\pi}{\alpha}\right)^{\frac{\|W\|}{2}} \\ E_W &= \frac{1}{2} \sum_{i=1}^{W} w_i^2 \end{aligned} \tag{10}$$

Choosing the variance such as mentioned simplied the calculation of the network and prevented the algorithm to be more complex. The prior distribution becomes as follows, where W, E, and $Z_W(\alpha)$ are the number of weights, weight decay and normalization factor respectively. In case of standard normal distribution, which is used in this study, the SSW is called *weight decay.*

2.4.2 The Likelihood Estimation

From the same method to obtain the prior, it is possible to calculate the likelihood function. The likelihood expresses how data energy is *likely* to decay through the learning process. Hence, the following approach was used to reach at the model likelihood. If data, D, consist of training-target set pairs (X_i, t_i), for 1 $\leq i \leq N$, the likelihood will be such as following:

$$\begin{aligned} \mathcal{P}(D|\vec{W}) &= \prod_{i=1}^{N} \mathcal{P}(t_i|x_i, \vec{W}, \beta) \\ &= \frac{1}{Z_D(\beta)} e^{-\beta E_D} \\ Z_D(\beta) &= \left(\frac{2\pi}{\beta}\right)^{\frac{N}{2}} \\ E_D &= -\frac{\beta}{2} \sum_{i=1}^{N} [M(x_i) - t_i]^2 \end{aligned} \tag{11}$$

where $M(x_i)$, $Z_D(\beta)$ and E_D are model output given inputs x_i, normalization factor and model ouput error respectively.

2.4.3 The Posterior Probability

The posterior distribution was simply calculated from Equations (9), (10) and (11):

$$\begin{aligned}\mathcal{P}(W|D,\alpha,\beta) &= \frac{1}{Z_S(\alpha,\beta)}\,e^{-(\beta E_D+\alpha E_W)} \\ Z_S(\alpha,\beta) &= \frac{1}{Z_W(\alpha)Z_D(\beta)\mathcal{P}(D)}\end{aligned} \tag{12}$$

Maximizing the posterior probability is easier by minimizing the total error of the network according to Equation (12). The reason of why the regularization term is sometimes called *weight decay* is obvious from the above mentioned equation.

2.4.4 Updating Hyperparameters of the Network

It has been shown that the priors are reliable for every re-parameterization when they are proportional reciprocally to the parameters *per se* [86]. It means to choose $\mathcal{P}(\alpha) = \dfrac{1}{\alpha}$ and $\mathcal{P}(\beta) = \dfrac{1}{\beta}$. By plugging these values in (9) and then expanding a Taylor-approximation of (12), one can get the following inference:

$$\begin{aligned}\ln\mathcal{P}(\alpha,\beta|D) \propto \ln\mathcal{P}(D|\alpha,\beta) + \ln\mathcal{P}(\alpha,\beta) &= -\alpha E_W^* - \beta E_D^* - \frac{1}{2}\ln\|H\| \\ &+ \frac{W}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) \\ &- \ln\alpha - \ln\beta \qquad (13) \\ H &= -\nabla\nabla\ln\mathcal{P}(\vec{W}|D) \qquad (14)\end{aligned}$$

where the starred parameters referred to optimized values of energies obtained from the last procedure. $\mathcal{P}(\alpha,\beta)$ is the non-informative prior to minimize the emperical risk. Equation (14) is Hessian of (7) at optimized weights W^* [15].

The goal was to find the optimum energy parameters done by calculating the partial derivatives of Equation (13) with respect to α and β and set the result to zero. Therefore, the optimum values for the parameters were found:

$$\begin{aligned} \alpha^* &= \frac{\gamma}{2E_W^*} \\ \beta^* &= \frac{N - \gamma}{2E_D^*} \\ \gamma &= N - 2\alpha_{old}^* tr(H^*)^{-1} \end{aligned} \tag{15}$$

N, H^{*-1} and γ are the number of parameters, inverse of the Hessian matrix and the remaining parameters after training the network respectively. Having obtained the updated parameters of energy function, one can simply implement an iterative algorithm embedded with any learning technique. Figures 12 and 13 shows the flowchart of Bayesian automated inference and the methodology utilized in this study.

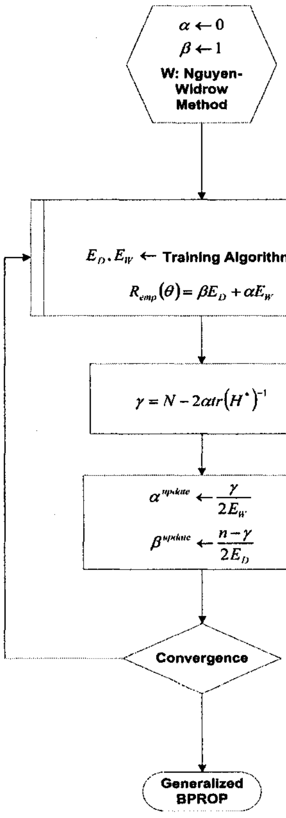

$\alpha \leftarrow 0$
$\beta \leftarrow 1$
W: Nguyen-Widrow Method
$E_D, E_W \leftarrow$ Training Algorithm
$R_{emp}(\theta) = \beta E_D + \alpha E_W$
$\gamma = N - 2\alpha tr(H^*)^{-1}$
$\alpha^{update} \leftarrow \frac{\gamma}{2E_W}$
$\beta^{update} \leftarrow \frac{n-\gamma}{2E_D}$
Convergence
Generalized BPROP

Figure 12: A Bayesian framework for pruning a feed-forward neural network

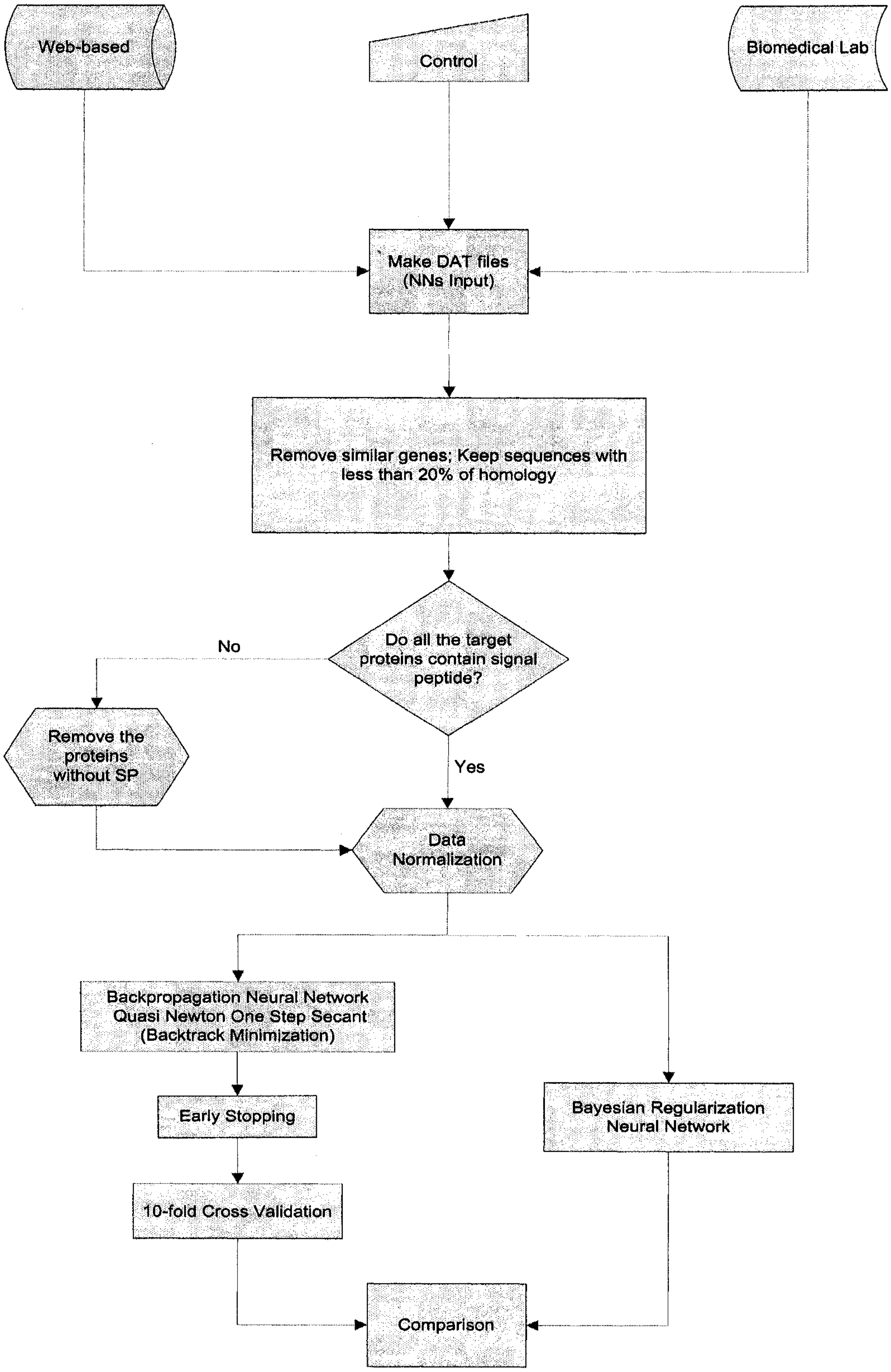

Web-based
Control
Biomedical Lab
Make DAT files (NNs Input)
Remove similar genes; Keep sequences with less than 20% of homology
Do all the target proteins contain signal peptide?
No
Yes
Remove the proteins without SP
Data Normalization
Backpropagation Neural Network Quasi Newton One Step Secant (Backtrack Minimization)
Early Stopping
10-fold Cross Validation
Bayesian Regularization Neural Network
Comparison

Figure 13: The phase II methedology for embedded Bayesian automated learning

2.5 The Neural Network Architecture

The nerual network had a single hidden layer to reduce the complexity of the search space. For model selection, different hidden units between 1 to 10 was examined along with the implementation of the network. The sigmoid function was the choice of the activation function. The learning algorithm exploited with Bayesian learning was Levenberg-Marquardt [52].

The programming environment was MATLAB ® 7.2 [88]. Table 5 shows the necessary codes implemented to provide an integrated environment for developing the entire project.

Table 5: The MATLAB and C++ Programs used for This study

Code Name	Description
seqencoder	Encode the input sequences of proteins using orthogonal encoding
kfoldcv	General purpose K-fold Cross Validation for evaluation of Backpropagation Neural Networks with any learning algorithm
CVencoder	Encode the input sequences of proteins using orthogonal encoding
f	Sigmoid transfer function for backpropagation network
window_frame	Reduce the size of window frame to the desired length
assess	Assessment measures calculation routine
display_discussion	A utility to nicely present the results
ASSESS	Modified C++ class version of 'assess'

The regularizaton ratio (RR) was set to 0.5 to make balance between regularization and mean square error (MSE) terms. The Nguyen-Widrow adaptive weight initialization [97] was the approach for initializing the network parameters.

For each window frame and learning algorithm, one Pentium 4 (3.99GHz) with 1GB of RAM computer was exploited as *psuedo-parallelization* of the project. The total training time was 9.0 hours/CPU. The average run-time for each window frame using Bayesian framework was 30 minutes on a Dual Processor Pentium 4 (3.00-2.99GHz) with 1GB of RAM. Finally, the results of training [BPROP-BRNN] were compared to [BPROP-OSS]. One reason was shown that the most consistent result among the applied learning algorithms in this study for EGF-like protein data belonged to [BPROP-OSS] [110]. The more detail will be explained in Chapter 3.

Chapter 3

Results and Discussion

3.1 Introducing the Results

The model response was superimposed on the test set distribution graph (Figure 14). The horizontal axis shows the distribution of the test set consisting of 880 sequences. The vertical axis represents the model response which was between 0.1 and 0.9. The first and third sets were those window frames which are EGFL and glycosylated. The second set was the window frames of non-glycosylated EGFLs. Finally, the last set consisted of glycosylated non-EGFL as well as arbitrary sequences or noises (Figure 14).

In the first phase of the study, the model response was found for the feed-forward network trained with different learning algorithms. The findings have been illustrated through Figures 27 to 56 in Appendix A. For each response, the corresponding training chart has also been shown. In the training chart, the x-axis indicates for the number of training iterations (epochs). y-axis is either mean squared errors (MSE) or log-MSE. The MSEs were all calculated after 10-fold cross validation to generalize the related feed-forward network. Blue, red and black colors specify training response, test response and the goal index respectively. Table 10 defines the nomenclature used

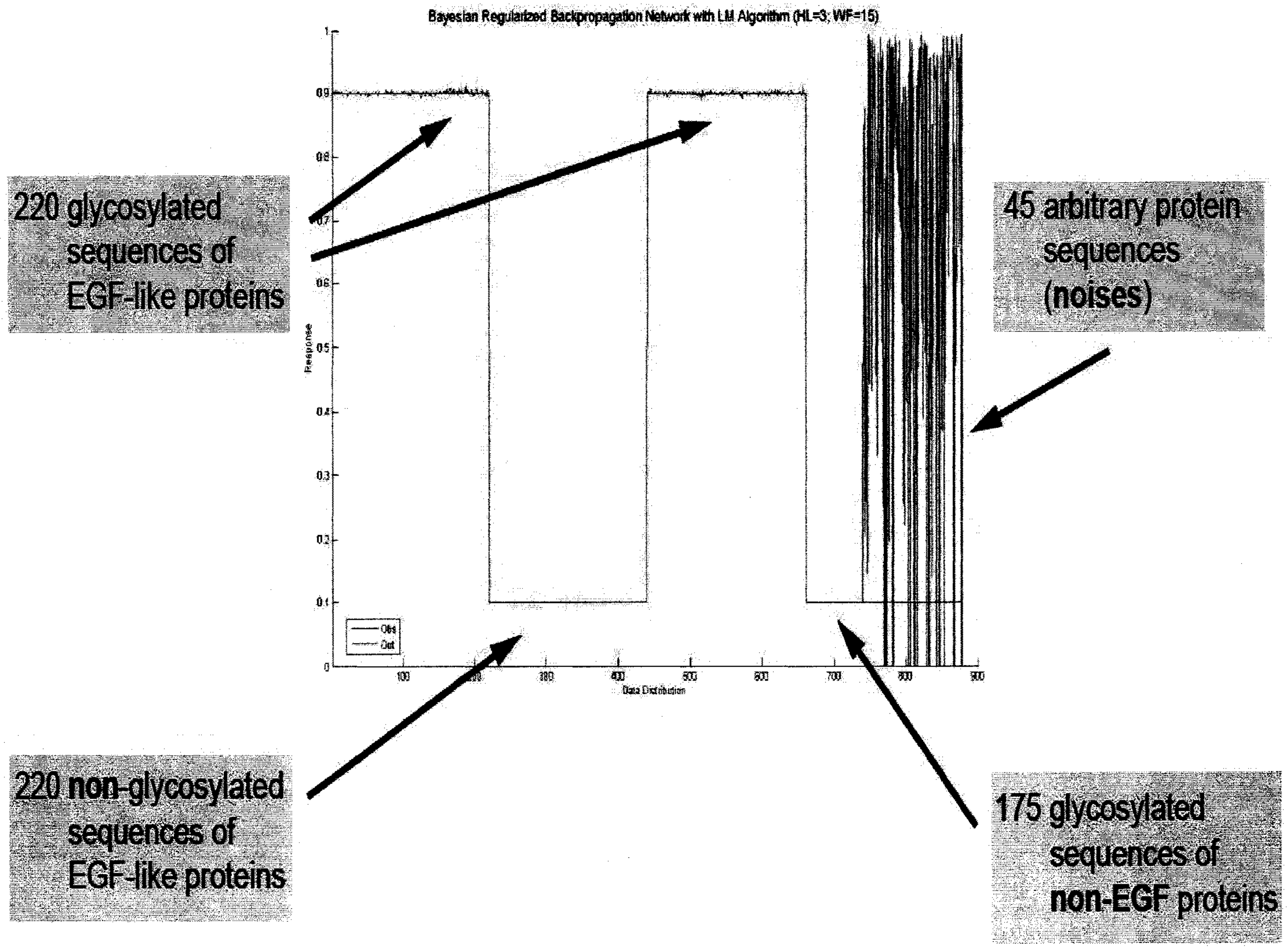
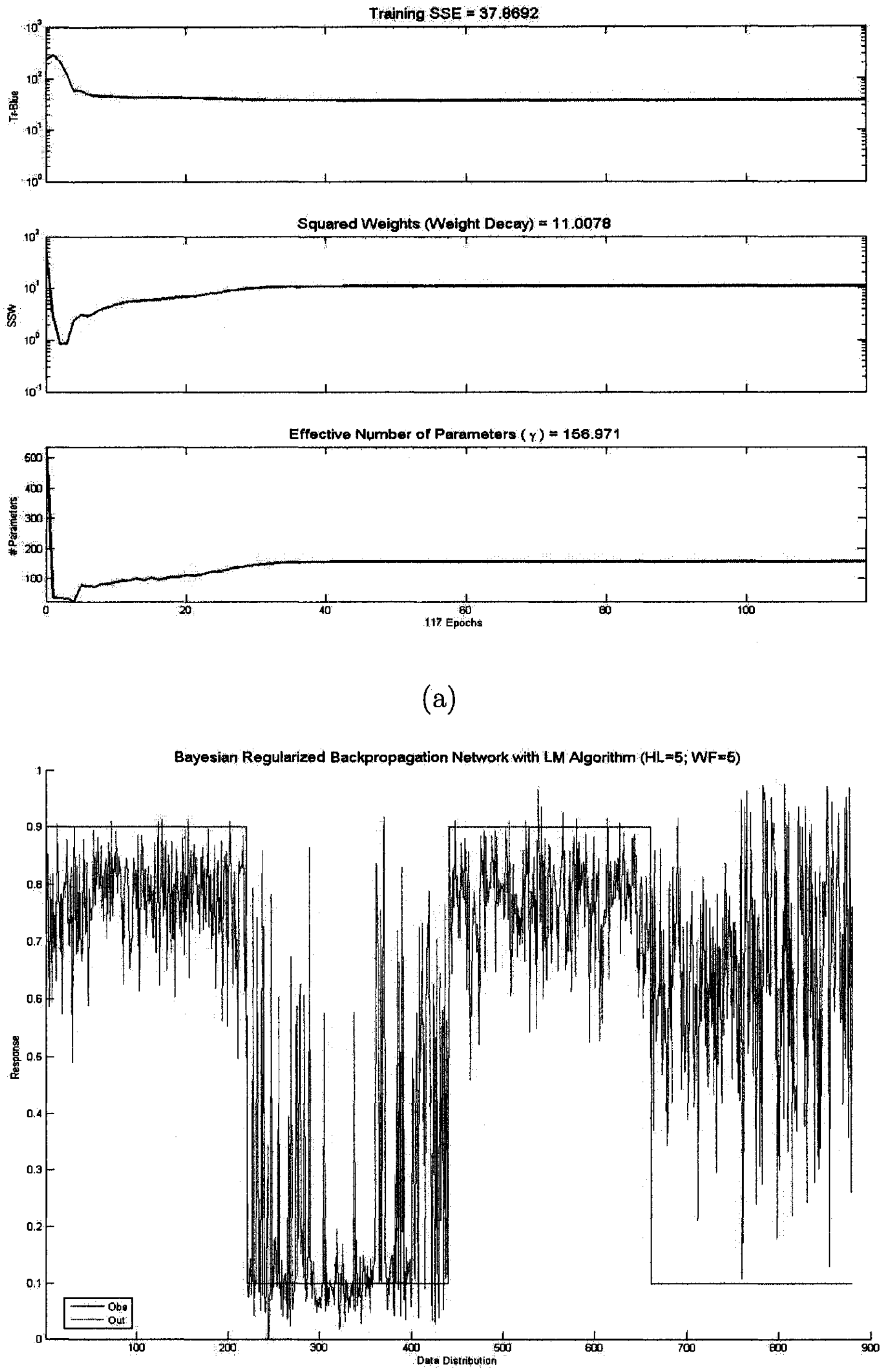


Figure 14: Introducing the test set regions as a graph

to label the graph designations as well as the parameters of the networks. In phase II, Figures 15 to 24 present the model response for [BPROP-BRNN] and [BPROP-OSS], which was more consistent than others in terms of response. The respective training charts for each type of neural network has also been demonstrated. In [BPROP-BRNN] training chart, SSE stands for the Sum of Sqaured errors. Sum of squared weights or weight decay is the term of regularization to penalize the network for large values of weights. The effective number of network parameters after training (γ) has also been show in the training chart.

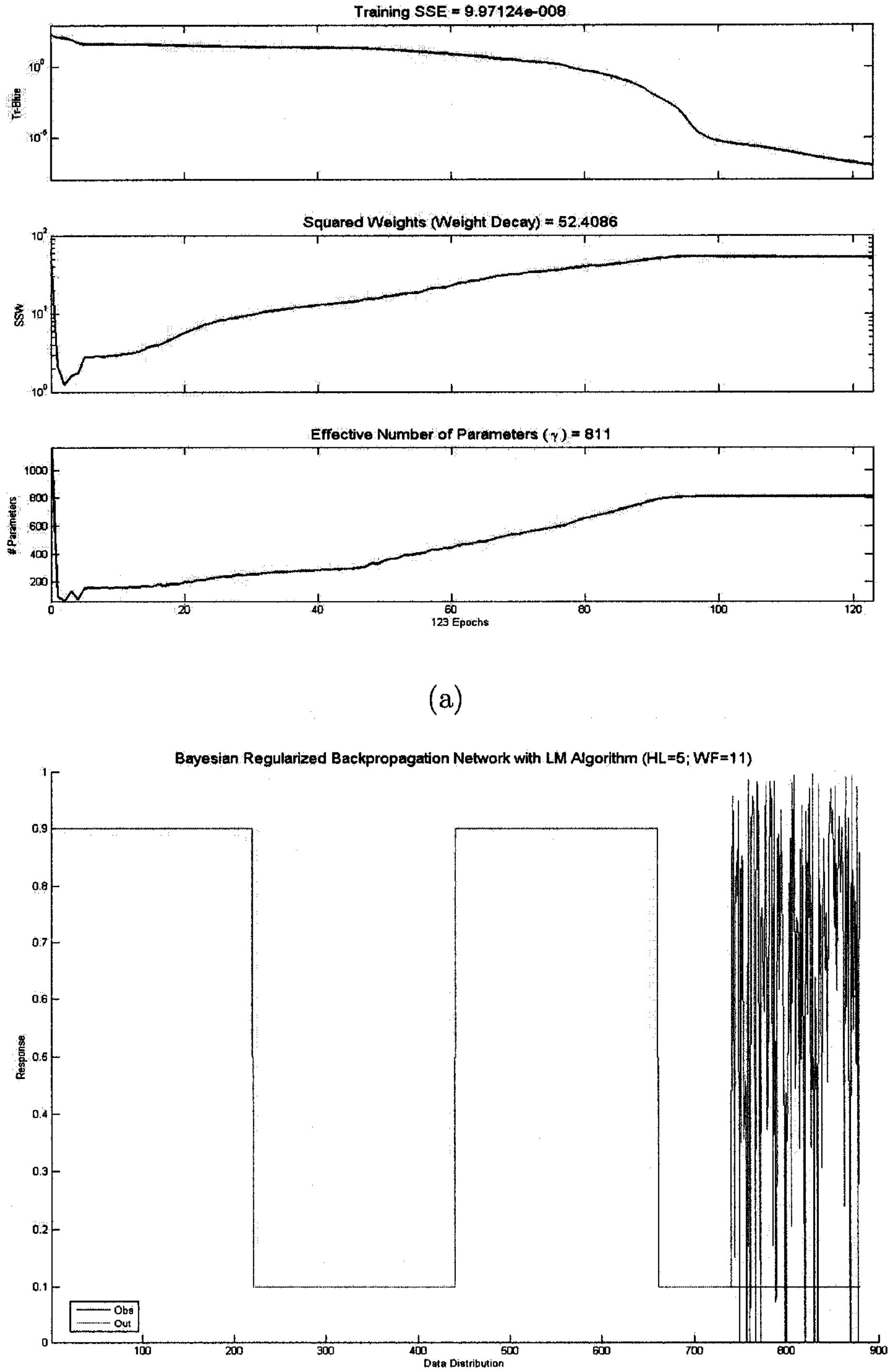
Bayesian learning prunes the unnecessary parameters of a neural network. In other words, the parameters with large variances with respect to others are set aside, so it was not needed to apply cross validation. Consequently, the training charts for [BPROP-BRNN] conatin only sum of squared errors.



(a)

(b)

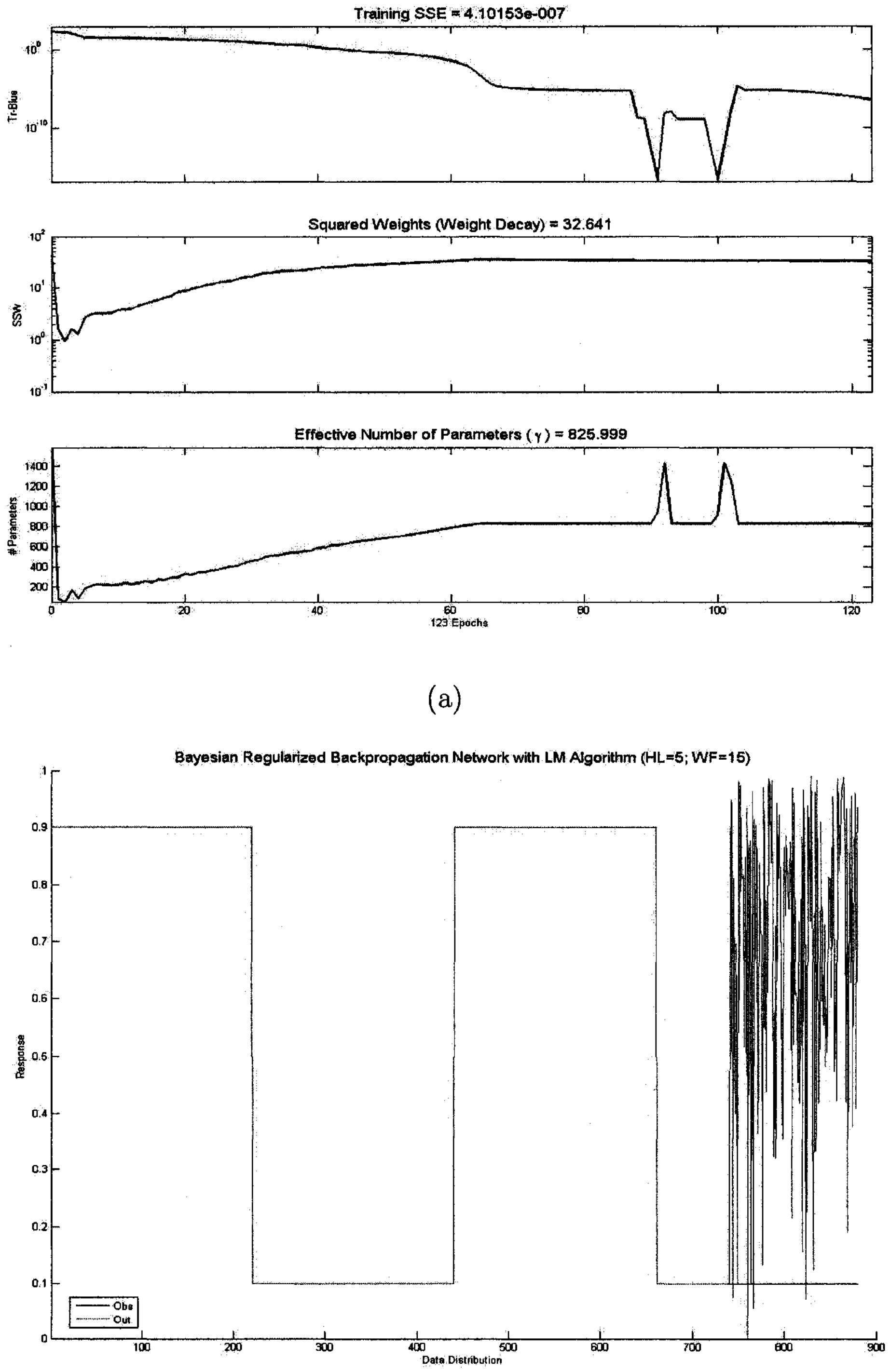
Figure 15: (a) [BPROP-BRNN], WF=5, HU=5 (b) Model Response



(a)

(b)

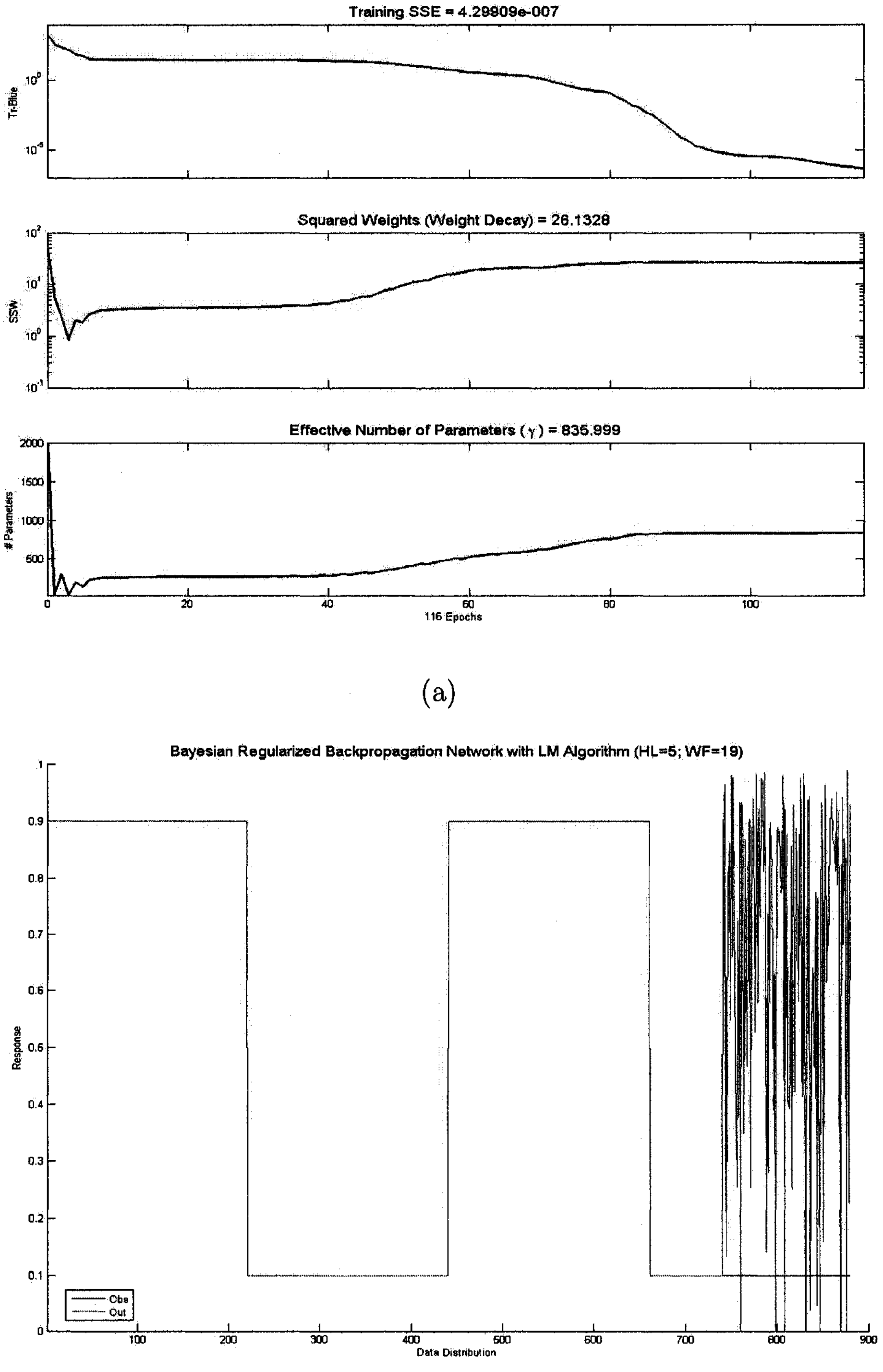
Figure 16: (a) [BPROP-BRNN] , WF=11, HU=5 (b) Model Response



(a)

(b)

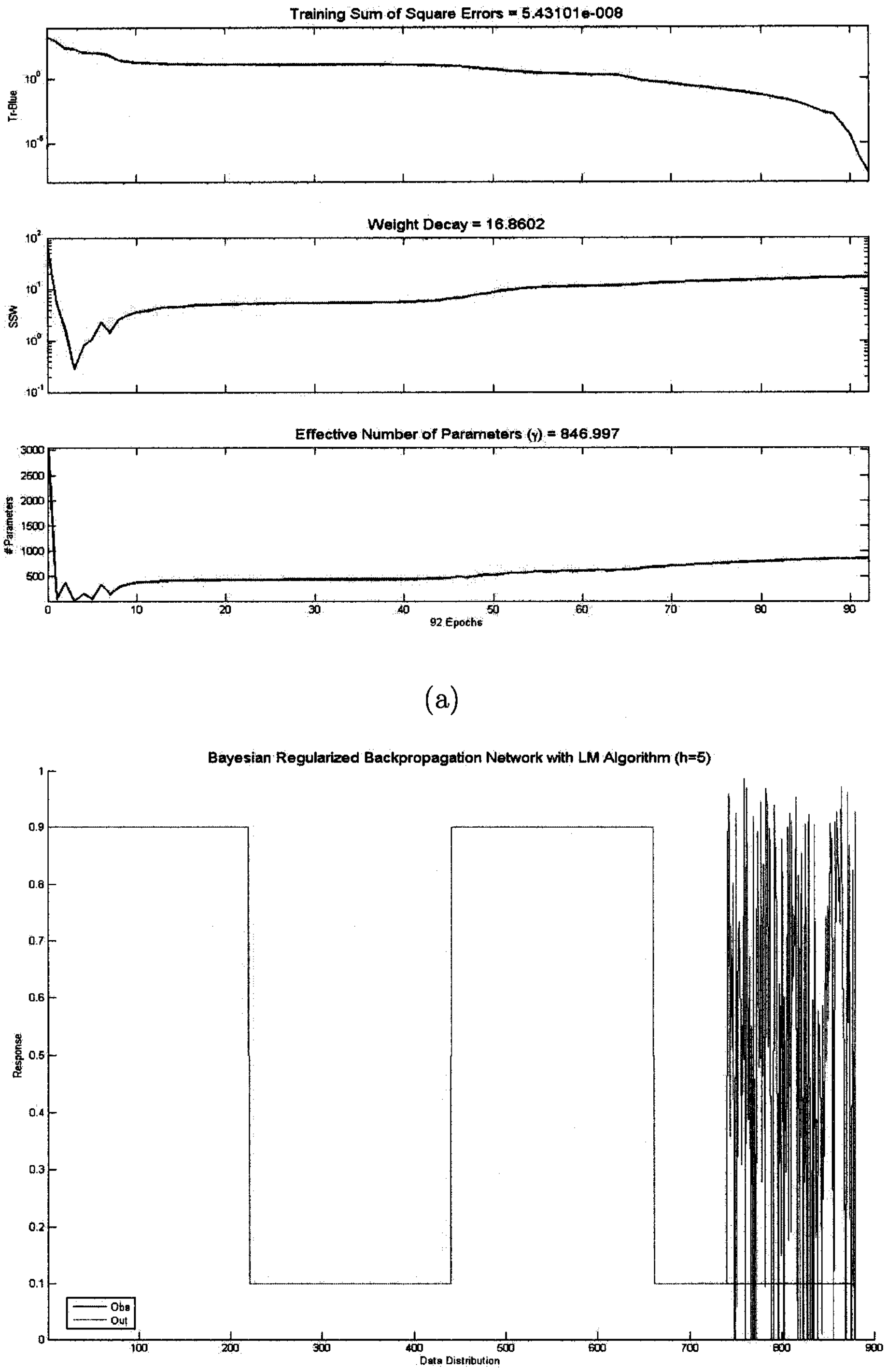
Figure 17: (a) [BPROP-BRNN] , WF=15, HU=5 (b) Model Response



(a)

(b)

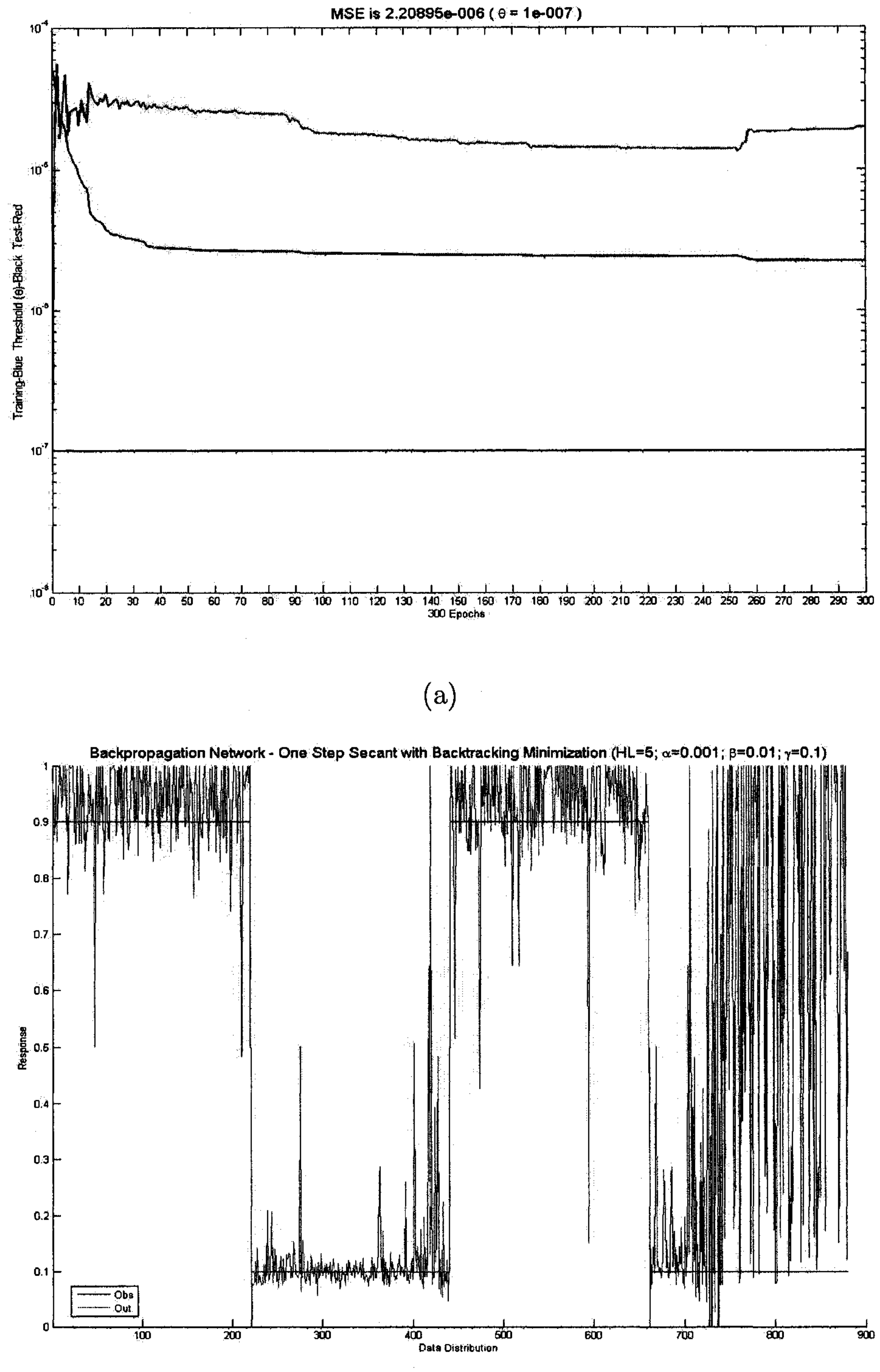
Figure 18: (a) [BPROP-BRNN] , WF=19, HU=5 (b) Model Response



(a)

(b)

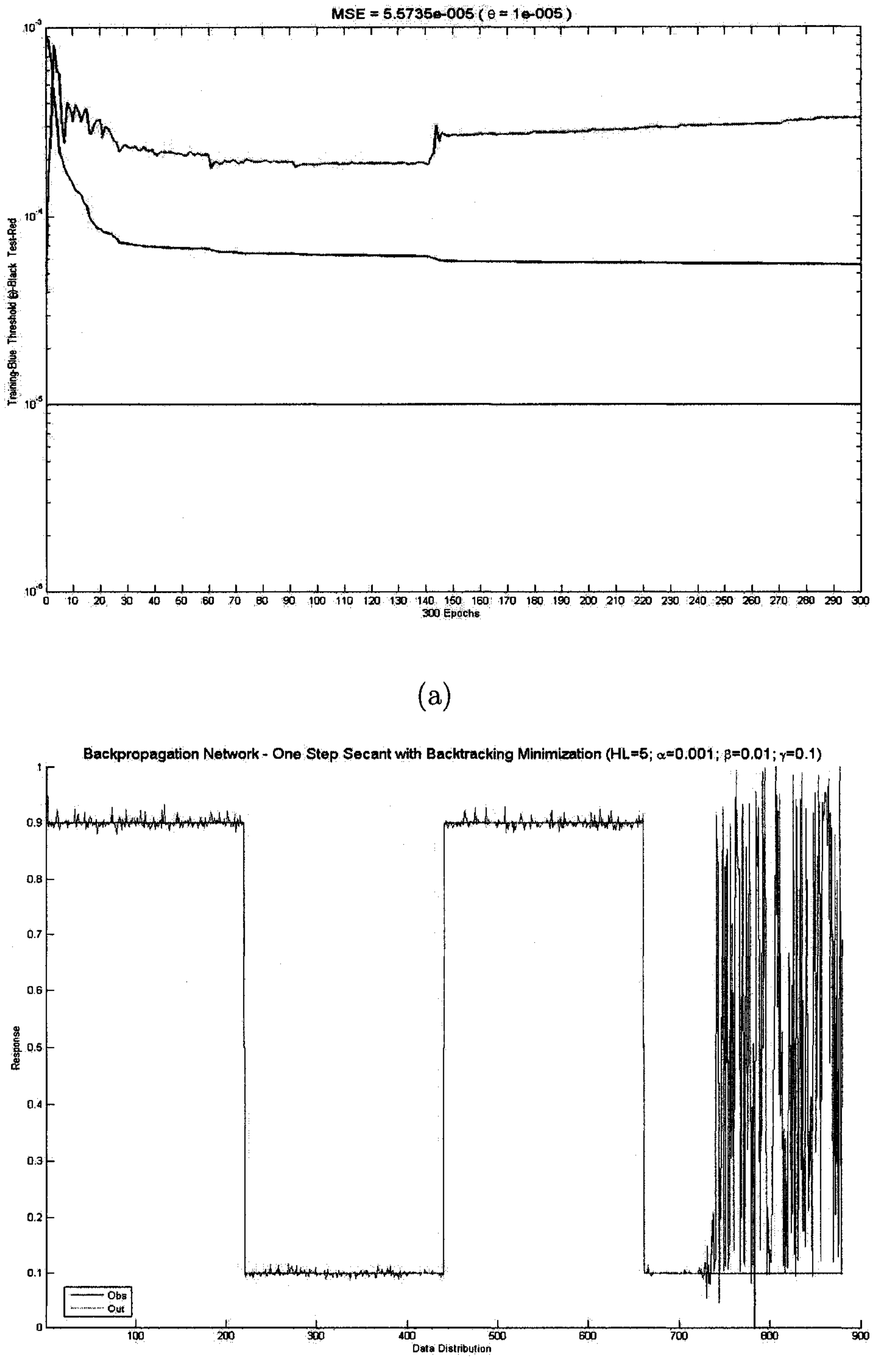
Figure 19: (a) [BPROP-BRNN] , WF=29, HU=5 (b) Model Response



(a)

(b)

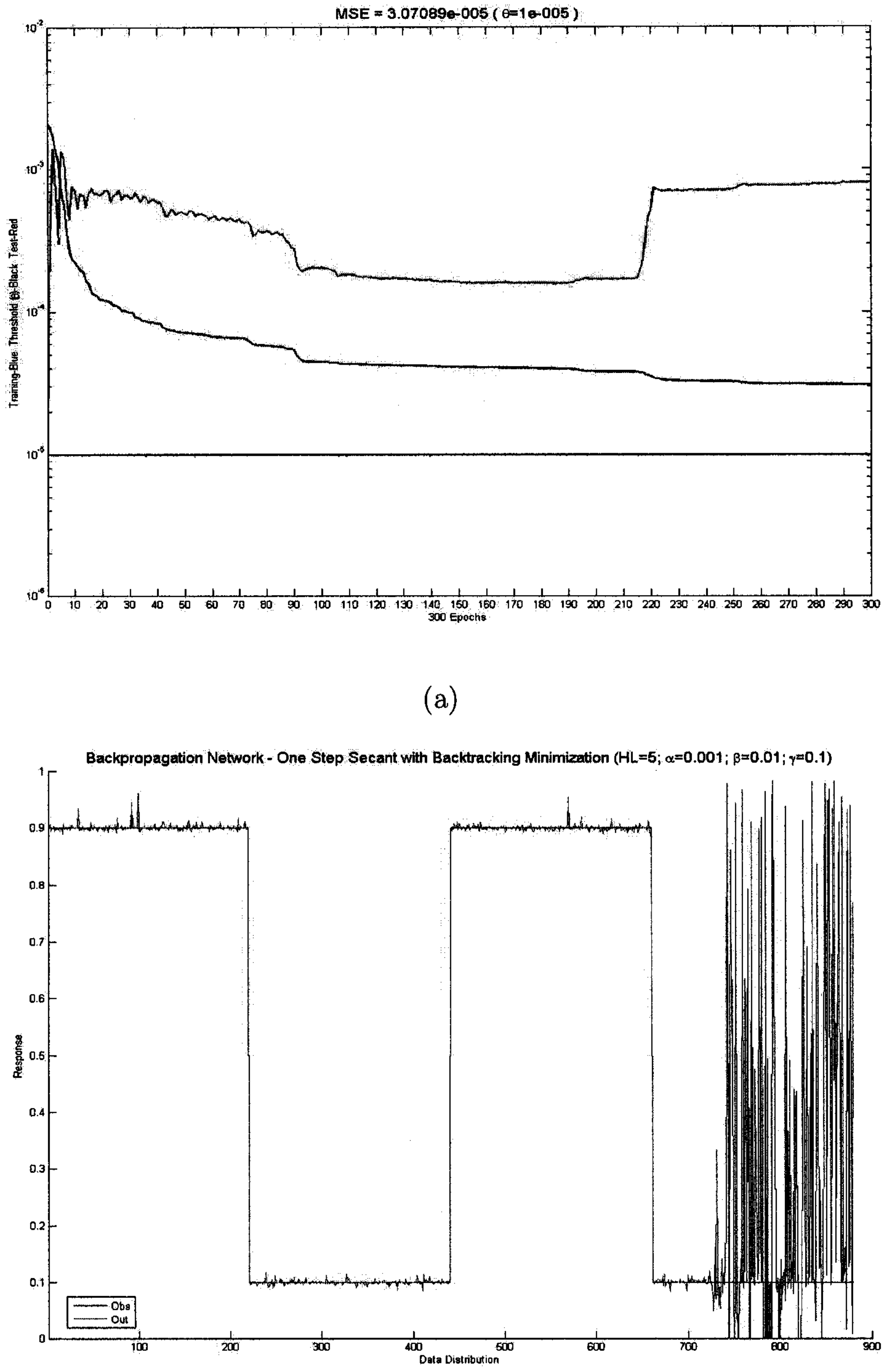
Figure 20: (a) [BPROP-OSS] , WF=5, HU=5 (b) Model Response



(a)

(b)

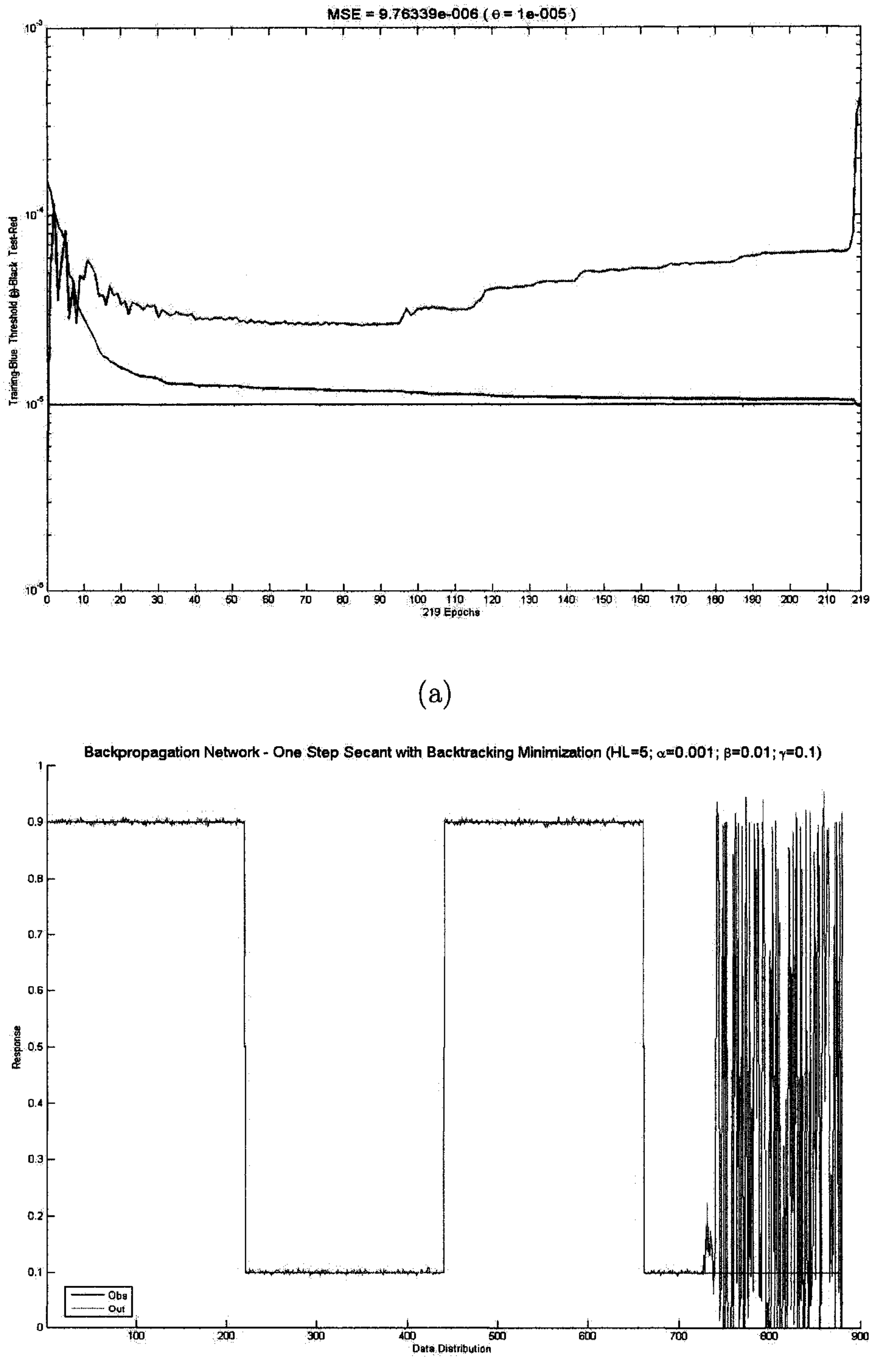
Figure 21: (a) [BPROP-OSS] , WF=11, HU=5 (b) Model Response



(a)

(b)

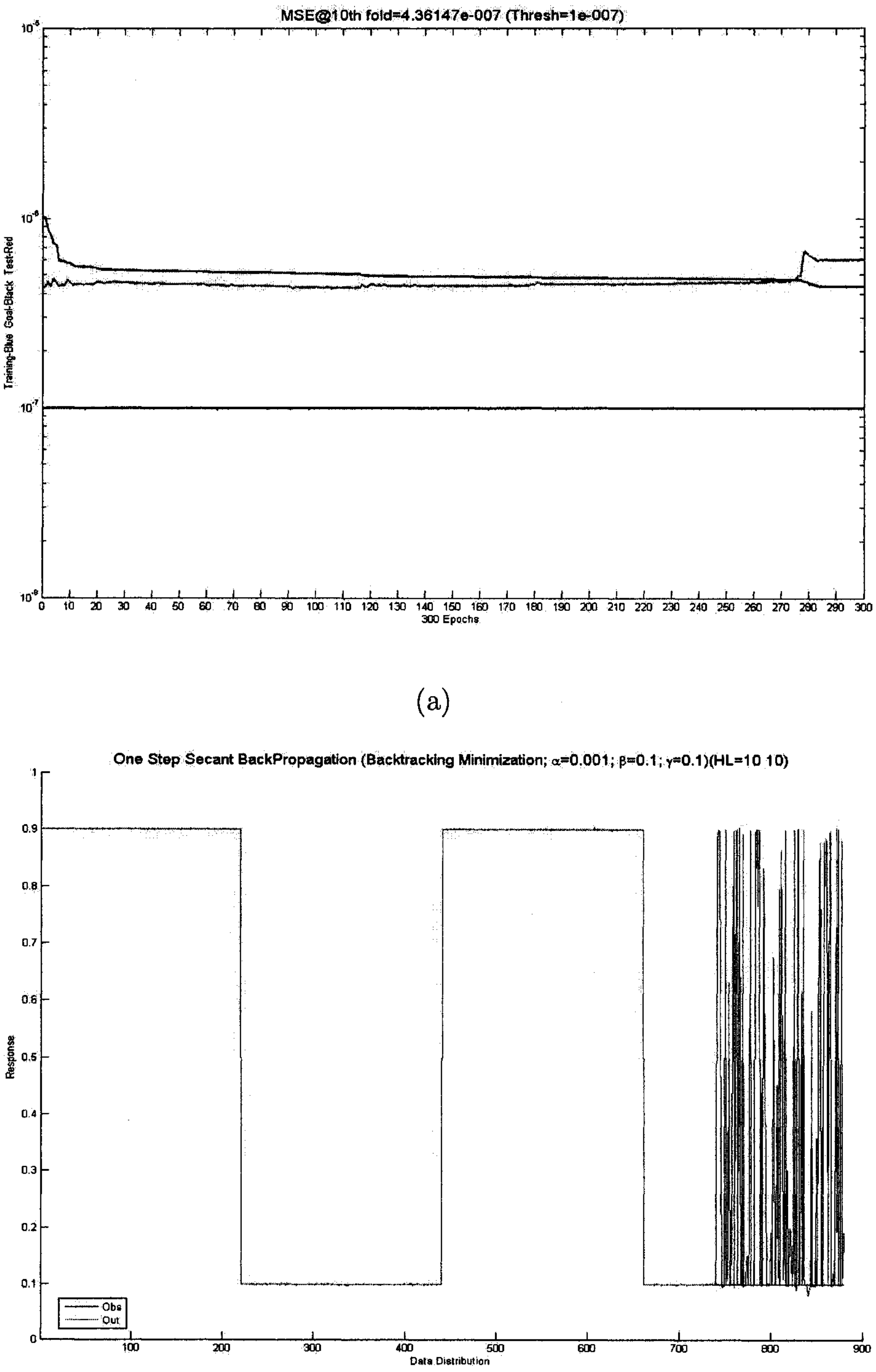
Figure 22: (a) [BPROP-OSS] , WF=15, HU=5 (b) Model Response



(a)

(b)

Figure 23: (a) [BPROP-OSS] , WF=19, HU=5 (b) Model Response



(a)

(b)

Figure 24: (a) [BPROP-OSS] , WF=29, HU=5 (b) Model Response

3.2 Assessment Measures

For the purpose of discussion, two types of assessment measures were selected:

- **Accuracy measures:** To evaluate the accuracy of the model response
- **Consistency measures:** To measure the reliability of the network

3.2.1 Accuracy Measures

one way to investigate the stability of the network is to measure the performance of the model using binary comparison [9]. In this study, M and $\widetilde{M}$ represented the glycosylated and not-glycosylated sites reported by the model respectively. D and $\widetilde{D}$ were also assumed to be the similar sites, but observed in the real dataset. The values of Figure 25 display:

True Positive Hits (TP). The number of times the amino acid glycosylated, and the network could detect them correctly.

True Negative Hits (TN). The number of times the amino acid not glycosylated, but the network had reported them as glycosylated.

False Positive Hits (FP). The frequency of amino acids glycosylated according to the model while they did not.

False Negative Hits (FN). The frequency of amino acids glycosylated, but the model detected them as non-glycosylated.

Figure 25 also shows the specificity and sensitivity, two major paramters to measure the accuracy of the network. *Sensitivity* is the probability of properly reported the true positive hits. *Specificity* is a criterion representing to what extent the detection of true positive hits is correct.

	M	$\widetilde{M}$	
D	TP	FN	$\text{SENS} = \frac{TP}{TP+FN}$
$\widetilde{D}$	FP	TN	
	$\text{SPEC} = \frac{TP}{TP+FP}$		

Figure 25: Dichotomy between glycosylated and non-glycosylated sites

The percentile of those assessment parameters were taken into account.

3.2.2 Consistency Measure

Similar to Pearson correlation, *Matthews Corelation Coefficient* (MCC) [89, 9] is a modified standard correlation used in the context of bioinformatics. The value of the correlation is always between -1 and 1. The zero value for MCC indicates a complete random estimation. On the other hand, the extreme value 1 of -1 for MCC represents complete correlation or uncorrelation respectively. MCC is calculated from the following Equation:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \tag{16}$$

$$MCC \in [-1, 1]$$

MCC is the measure of dichotomy features problems. If the feature selection process requires more than two features, other performance meaures may take place. MCC was found to be an appropriate for this study because the nature of the subject was to differentiate between glycosylated and non-glycosylated sites. For example, R_k is a generalized version of MCC and can be used for multi-features problems [45].

3.3 Discussion

This research has attempted to assess two field studies. The first field was the bioinformatics significance of the study in which the performance and accuracy of the model response was of interest. The main goal was to achieve to a feed-forward neural network producing statistically significant outputs. In addition, the topology of the neural network was important as a second field study, since a consistent framework could generally improve the output of the feed forward networks.

In the first section, the bioinformatics significance of this study will be discussed, and the related tables be introduced. In the next section, the concept and analysis of *Reduction Factor* (ρ) as an index of the feed-forward network will be introduced and discussed.

3.3.1 Bioinformatics Interpretations

The first part of the study was considered with 29-residue sequences and a single-layer feed-forward network with 10 hidden units. It was originally assumed that giving a pre-determined complexity to the network helps to obtain consistent results along with a fast training. However, the results of *HU* 10 10, a network with two layers with 10 hidden units for each did not confirm that assumption ((b) subfigures of Figures 29 to 56). The figures suggest that a single-layer neural networks is enough for study the large datasets of EGF-like protein sequences. Compared to other algorithms, standard backpropagation was less consistent. Even adding another extra hidden layer to give a complexity to the network could not help to improve the response. Other classes of standard backpropagation such as [BPROP-GD-X] and [BPROP-GD-M] showed the similar results. Nevertheless, provided by an adaptive learning term, the results for [BPROP-GD-X] could be better that those of other standard backpropagation algorithms, as displayed in Figures 32 to 34.

Table 6 summarizes the performance and accuracy meaurements for the learning algorithms applied in this study. While [BPROP-GD] reported positive hits more than others, BPROP-OSS] could detect glycosylation sites more consistent by 89.97% of specificity. [BPROP-OSS] has the minimum average error for detecting positive glycosylation sites among other algorithms. Moreover, it had the minimum sensitivity. On the other hand, [BPROP-GD] has the maximum average error.

Table 6: The Standard Measures for the Learning Algorithms (HU=10)

Algorithm	$\overline{\text{MSE}}$	Reg. $\overline{\text{MSE}}$	SPEC%	SENS%	MCC
[BPROP-GD]	0.0673	0.0446	64.11	55.77	0.340
[BPROP-GD-M]	0.0554	0.0578	65.33	55.63	0.336
[BPROP-GD-AL]	0.0161	0.0214	77.03	55.35	0.327
[BPROP-GD-AL-M]	0.0156	0.0163	77.63	53.14	0.251
[BPROP-RPROP]	0.0569	0.0368	63.61	57.52	0.235
[BPROP-CGF]	0.0117	0.0310	76.33	53.53	0.266
[BPROP-SCG]	0.0124	0.0264	77.37	52.88	0.256
[BPROP-OSS]	0.0103	0.0193	89.97	52.77	0.425

The results suggest that standard backpropagation cannot map the correlation between the distribution of glycosylation sites and the EGF-like protein sequences dataset. Such results differ from those of Gupta and Brunak [49] for general proteins in which their systems have identified the glycosylated and non-glycosylated sites by 97%.

Regularization was applied to the model in this study to take the effect of large weights into account. The manual regularization coefficient was set to either 0.5 or 0.3 depending the analysis. Nonetheless, the the manual regularization could not improve the detection or prediction of glycosylation sites. Compared to general algorithms, the average error of manual regularization with ratios of 0.5 and 0.3 was higher than that of regularized algorithms by 0.9387%.

The highest specificity for [BPROP-OSS] compared to other methods emphasized its usage for the next phase of the study, as a model comparison paradigm against Bayesian learning. On the other hand, [BPROP-RPROP] had the most sensitivity, but least specificity. Moreover, the minimum MCC value was also found for [BPROP-RPROP].

When accompanied with two-layer neural networks, Quasi-Newton family of learning algorithms appeared to be more consistent that others in terms of accuracy meaurement. One example was [BPROP-SCG] (Figures 46 to 49). The nature of their line-search routine may lead to such results [14].

In the second phase, the [BPROP-BRNN] was compared with [BPROP-OSS] to determine the efficiency of Bayesian learning in pruning the redundant parameters of the feed-forward network. Table 7 shows the standard measurements for both [BPROP-BRNN] and [BPROP-OSS].

Table 7: Standard measurement for [BPROP-BRNN] and [BPROP-OSS]

Window Frame	**Assess**	**BRNN**	**OSS**
5-residue	SPEC%	66.37	38.91
	SENS%	69.22	51.00
	MCC	0.674	0.111
11-residue	SPEC%	70.99	47.11
	SENS%	68.17	54.83
	MCC	0.790	0.165
15-residue	SPEC%	66.25	43.29
	SENS%	75.23	55.17
	MCC	0.715	0.153
19-residue	SPEC%	75.16	57.15
	SENS%	70.19	60.90
	MCC	0.821	0.341
29-residue	SPEC%	77.00	50.11
	SENS%	76.19	49.35
	MCC	0.851	0.313

The measures were evaluated for HU=1 to Hu=5 and for different window frames. In this phase, the effect of window frame as a representative of prior knowledge and the number of hidden units as a criterion for model complexity was the main purpose of study.

The maximum specificity was obtained for [BPROP-BRNN] and [BPROP-OSS] along with 29- and 19-residue window frames respectively. On the other hand, 15- and 5-residue frames correspondingly associated with [BPROP-BRNN] and [BPROP-OSS] showed the least specificity. It is possible to consider a lengthy frame along with Bayesian learning as well as quasi-Newton methods. However, resulting in the least specificity, the 15-residue frame could not show any evidence for stability. [BPROP-BRNN] and [BPROP-OSS] could also detect the most positive hits when they were fed with 29- and 19-residue sequences respectively. Performance measure (MCC) was maximum in 29-residue window frames in [BPROP-BRNN]. For [BPROP-OSS] the similar result happened for 19-residue frames. In both networks the minimum MCC belonged to 5-residue window frames. The results also showed a 62.22% in the maximum MCC using automated regularization, which is a significant improvement for the feed-forward network. It was anticipated that the Bayesian learning might lead to a stronger generalization than that of quasi-Newton approach.

The results strongly suggest to use the window frames with not less than 5 amino acids around the glycosylation sites. The performance of the networks provided with enough knowledge increased according to Table 7.

3.3.2 Machine Learning Interpretations

Bayesian automated regularization was a technique used to investigate whether the gernalization of the model would be improved. Hence, quantifying the assessment of Bayesian learning was possible by introducing a simple, yet effective parameter.

Introducing the Neural Networks Reduction Factor

Reduction Factor (ρ) is the measure which calculates how much the size of the network has been reduced:

$$\rho = \frac{\Gamma_T - \gamma_E}{\Gamma_T} \tag{17}$$

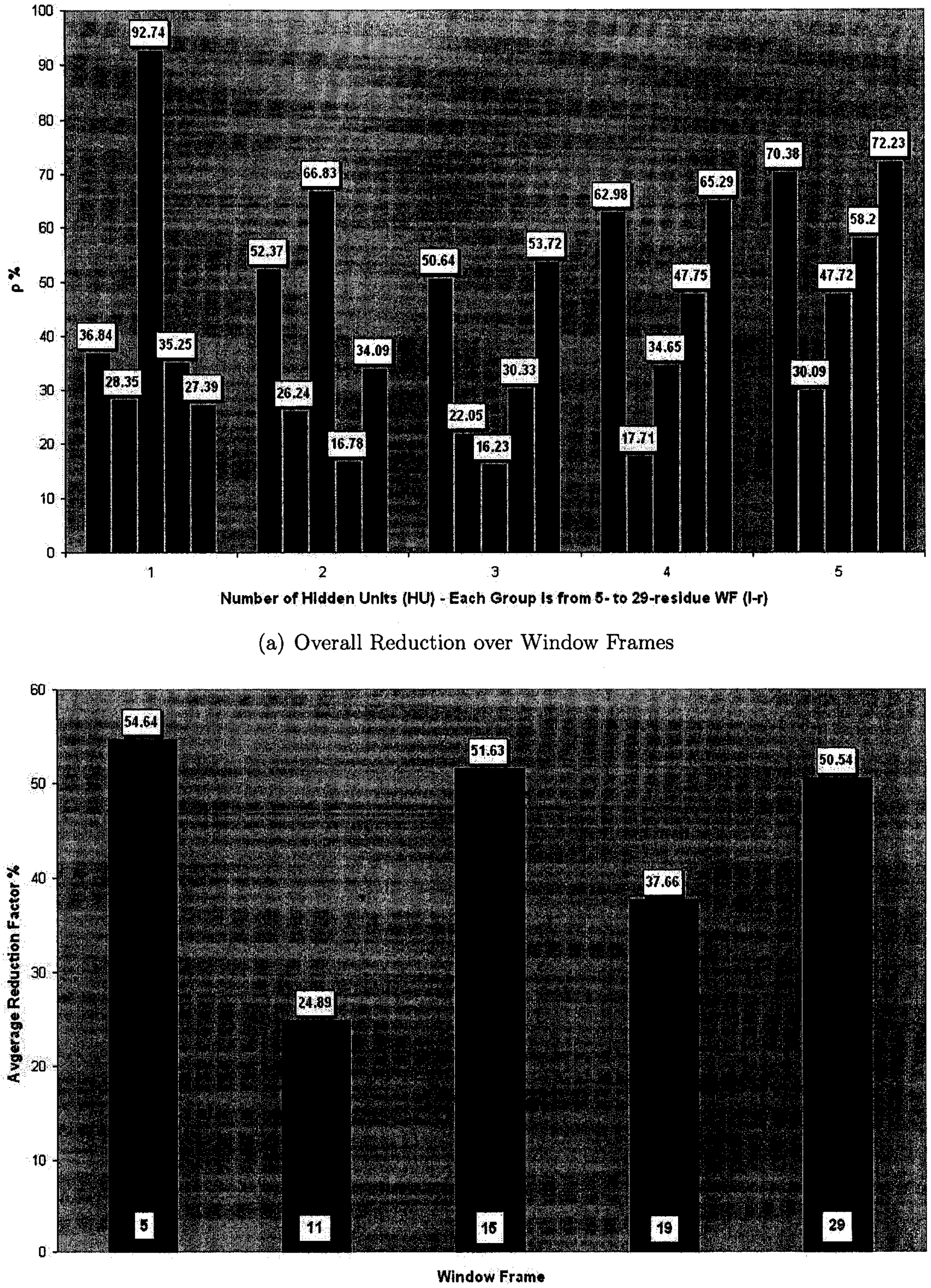
where Γ_T is the total number of network weights, and γ_E is the remaining parameters of the network after pruning process. The Reduction Factor was applied to the Bayesian framework of this study. According to the findings in Table 8, after applying the Bayesian approach, the average size of the network was reduced by 47.62%. It was revealed that the 15-residue window frame behaves chaotically. As indicated in Table 8, the 15-residue frame obtained the maximum reduction with one hidden units; on the other hand, the network parameters minimally reduced with the same window frame and using a single-layer network with 3 hidden units. Thus, the result implies that the 14 amino acids around glycosylation sites do not offer significant information. This finding is in substantial agreement with that of Section 3.3.1 regarding to 15-residue window frame.

Figure 26(a) shows the overall reduction of the network after applying Bayesian framework. Each cluster represents the group of window frames from left to right. The order is from 5- to 29-residue frame. One major point is the transient step in the Table 8. The Reduction Factor for the networks with one and two hidden always was maximum for 15-residue window frame. Instead, for the networks with 3, 4 and 5 hidden units, there was more clear pattern. For example, there was more reduction for 5- and 29-residue frames in the single-layer networks with 4 and 5 hidden units.

Figure 26(b) displays the average reduction over the studied window frames. The minimum and maximum reduction belonged to 5- and 29-residue frames respectively.

Table 8: [BPROP-BRNN] Reduction Size Analysis

WF	HU	$\mathbf{\Gamma}_T$	γ_E	$\rho\% = \frac{\mathbf{\Gamma}_T - \gamma_E}{\mathbf{\Gamma}_T} \times 100\%$
5	1	106	66.95	36.84
	2	212	100.98	52.37
	3	318	156.97	50.64
	4	424	156.97	62.98
	5	530	156.97	70.38
11	1	232	166.22	28.55
	2	464	342.25	26.24
	3	696	542.51	22.05
	4	928	763.64	17.71
	5	1160	811.00	30.09
15	1	316	22.93	92.74
	2	632	209.61	66.83
	3	948	794.15	16.23
	4	1264	826.00	34.65
	5	1580	826.00	47.72
19	1	400	258.99	35.25
	2	800	665.79	16.78
	3	1200	836.00	30.33
	4	1600	835.97	47.75
	5	2000	836.00	58.20
29	1	610	442.93	27.39
	2	1220	804.19	34.09
	3	1830	846.99	53.72
	4	2440	847.00	65.29
	5	3050	847.00	72.23



(a) Overall Reduction over Window Frames

(b) Average Reduction over Window Frames

Figure 26: BRNN Reduction Factor Statistics

3.4 The Comparison with Existing Systems

The performance of Bayesian regularization neural network was compared to that of the exsiting systems, NetNGlyc and NetOGlyc.

The test set were given to Bayesian regulariation neural network, NetNGlyc and NetOglyc. The reports of the existing systems and Bayesian network were utilized to calculate the number of true and false postives as well as true and false negatives, and then evaluated in terms of the assessment measures of this study.

Table 9: The responses of the Bayesian neural network and the existing systems

System	SPEC%	SENS%	MCC
NetNGlyc-NetOGlyc (Average)	37.10	49.70	0.538
Bayesian Regularization Neural Network	81.11	85.05	0.889

Table 9 represents the model responses for both the Bayesian neural network of this study and the existing systems. For the context of this study, Bayesian regularization neural network could improve the detection and prediction of glycosylation sites by 54.25%. Moreover, the accuracy of the response in The Bayesian model was higher by 41.56%. Bayesian neural network was more consistent than the average response of the existing models by 39.48%.

Chapter 4

Conclusion

There are noteworthy points learned through carrying out this study. In the first section, those points will be briefly reviewed. Finally, the possible future works and suggestions will be discussed.

4.1 The Lessons Learned from the Project

Artificial neural networks have been used to find a non-linear mapping between the intrinsic characterstics and sequences of protein(s) during the post-translational modifications (PTMs). As the most common and complex modification, protein glycosylation was focused in this study. Among the growth factors superfamily, Epidermal Growth Factor-like (EGFL) repeats were the protein studied in this research. The accurate detection and prediction of the glycosylation sites in EGF-like oncoproteins was the ultimate goal of this study.

One class of neural networks, feed-forward networks or multi-layer perceptrons, were employed to succeed the goal. As reviewed in Chapter 1, most of the existing techniques use standard backpropagation as the learning paradigm with feed-forward network. It was learned that the standard backpropagation was not sufficient for concerned perception of non-linear functionality in EGFLs. Furthermore, it was anticipated that the topology of the feed-forward network may influence the process of knowledge extraction from the EGF-like sequences. As a result, Bayesian framework was established to investigate that hypothesis. Exploited as a quantitative approach, the concept of Reduction Factor was introduced to support the investigation.

The results of Bayesian regularization neural network was compared to the average results of the existing systems. For EGFL proteins, the Bayesian network was more consistent by 39.48%. In addition, the specificity and sensitivity of Bayesian neural network was higher by 54.25% and 41.56% respectively.

In the first phase of the study, different learning algorithms were used along with the feed-forward network. [BPROP-OSS], a quasi-Newton learning algorithm, was found to be more consistent than others in terms of accuracy and reliability. In the second phase, the Bayesian automated learning was utilized. The network was initially implemented by the maximizing the posterior probabilities. After that, the network parameters were pruned such that the less important weights and biases were neglected. Bayesian learning outperformed the quasi-Newton algorithm in terms of both accuracy and consistency over the networks parameters as well as model response. The neural network with both Bayesian and quasi-Newton learning approaches was employed to detect the glycosylation sites of the epidermal growth factor-like repeat proteins. The true positive hits were much higher in the network trained with Bayesian learning. This would suggest applying this framework for knowledge inference from large proteomic data. In fact, with enough prior information, it is possible to estimate the model parameters even with large number of protein sequences.

4.2 Future Works

Evolutionary-related EGFL sequences may affect the accuracy of the model due to an unavoidable similarity among those sequences. Removing the sequences with the same origin from the dataset introduces less prior knowledge to the model whereas keeping them may influence the model response. Therefore, the trade-off between keeping and removing evolutionary-related EGFL sequences is a challenging issue for further studies.

Bayesian learning is expensive and computationally complex. Using other encoding schemes such as adaptive encoding [67] may suggest a solution to overcome that disadvantage of [BPROP-BRNN]. It appears that the lengthy sequences of EGF-like domains lead to a more consistent Bayesian framework. Thus, it is suggested that the correlation between various lengths longer than 14 residues and the model response be studied and evaluated.

The network parameters as well as the parameters indirectly affecting the networks may play a role in extracting knowledge from such studied models. For example, the number of folds in the process of cross validation and the nature of transfer functions are two of those criteria. It has been demonstrated that the application of Radial Basis Functions to single-layer networks can direct to improve the convergence of the network [76].

It is proposed that the approach outlined in this study be replicated in or extended to other specific superfamilies of proteins to find a standard reference for such studies.

Finally, it is recommended that the results of such *in silico* analysis be validated by biologists to advance the reliability of such connectionist models.

Bibliography

[1] C. Abeijon and C.B. Hirschberg. Topography of Glycosylation Reactions in the Endoplasmic Reticulum. *Trends Biochem Sci.*, 17(1):32–36, 1992.

[2] E Appella, E.A. Robinson, S.J. Ullrich, M.P. Stoppelli, A. Corti, G. Cassani, and F. Blasi. The receptor-binding sequence of urokinase. a biological function for the growth-factor module of proteases. *Journal of Biological Chemistry*, 262(10):4437–4440, 1987.

[3] E. Appella, I.T. Weber, and F. Blasi. Structure and Function of Epidermal Growth Factor-Like Regions in Proteins. *FEBS Letters*, 231(1):1–4, 1988.

[4] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.L. Yeh. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, 32(suppl_1):D115–119, 2004.

[5] R. Apweiler, H. Hermjakob, and N. Sharon. On the Frequency of Protein Glycosylation, as Deduced from Analysis of the SWISS-PROT Database. *Biochimica et Biophysica Acta*, 1473(1):4–8, 1999.

[6] B. Arpinar, K. Giriloganathan, and B. Aleman-Meza. Ontology Quality by Detecting of Conflicts in Metadata. In *WWW 2006*, Edinburgh, UK, May 2006.

[7] C.L. Arteaga. Epidermal Growth Factor Receptor Dependence in Human Tumors: More Than Just Expression? *Oncologist*, 7(90004):31–39, 2002.

[8] A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.L. Yeh. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(suppl_1):D154–159, 2005.

[9] P. Baldi, S. Brunak, Y. Chauvin, C.A. F. Andersen, and H. Nielsen. Assessing the Accuracy of Prediction Algorithms for Classification: An Overview. *Bioinformatics*, 16(5):412–424, 2000.

[10] L.E. Ball, M.N. Berkaw, and M.G. Buse. Identification of the Major Site of O-Linked beta-N-Acetylglucosamine Modification in the C Terminus of Insulin Receptor Substrate-1 . *Molecular and Cellular Proteomics*, 5(2):313–323, 2006.

[11] P.L. Bartlett and R.C. Williamson. The VC Dimension and Pseudodimension of Two-Layer Neural Networks with Discrete Inputs. *Neural Computation*, 8(3):625–628, 1996.

[12] T. Battiti. First and Second-Order Methods for Learning: Between Steepest Descent and Newton's Method. *Neural Computation*, 4(2):141–166, 1992.

[13] M. Bernacki and P. Włodarczyk. *Principles of Training Multi-layer Neural Network Using Backpropagation Algorithm.* World Wide Web, `http://galaxy.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html`, 2005.

[14] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, Nashua, NH, U.S.A., 1999.

[15] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Inc., New York, NY, USA, 1995.

[16] G. Blobel and D.D. Sabatini. Ribosome-membrane Interaction in Eucaryotic Cells. In L. A. Manson, editor, *Biomembranes*, volume 2, pages 193–195, New York, U.S.A., 1971. Plenum Publishing Corp.

[17] N. Blom, S. Gammeltoft, and S. Brunak. Sequence- and Structure-based Prediction of Eukaryotic Protein Phosphorylation Sites. *Journal of Biological Chemistry*, 294(5):1351–1362, 1999.

[18] N. Blom, T. Sicheritz-Pontén, R. Gupta, S. Gammeltoft, and S. Brunak. Prediction of Post-translational Glycosylation and Phosphorylation of Proteins from the Amino Acid Sequence. *Proteomics*, 4(6):1633–1649, 2004.

[19] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.

[20] A. Bohne-Lang and C.-W. von der Lieth. GlyProt: *In Silico* Glycosylation of Proteins. *Nucleic Acids Research*, 33(suppl_2):W214–219, 2005.

[21] T. Buskas, S. Ingale, and G.-J. Boons. Glycopeptides as Versatile Tools for Glycobiology. *Glycobiology*, 16(8):113R–136, 2006.

[22] Y.D. Cai, H. Yu, and K.C. Chou. Artificial Neural Network method for Predicting the Specificity of GalNAc-transferase. *Journal of Protein Chemistry*, 16(7):689–700, 1997.

[23] G. Carpenter and S. Cohen. Epidermal Growth Factor. *Journal of Biological Chemistry*, 265(14):7709–7712, 1990.

[24] Z. Chen and S. Haykin. On Different Facets of Regularization Theory. *Neural Computing*, 14(12):2791–2846, 2002.

[25] C.A. Cooper, M.J. Harrison, J.M. Webster, M.R. Wilkins, and N.H. Parker. Data Standardisation in GlycoSuiteDB. *Pac. Symp. Biocomput.*, pages 297–309, 2002.

[26] C.A. Cooper, H.J. Joshi, M.J. Harrison, M.R. Wilkins, and N.H. Packer. GlycoSuiteDB: A Curated Relational Database of Glycoprotein Glycan Structures and Their Biological Sources. 2003 update. *Nucleic Acids Research*, 31(1):511–513, 2003.

[27] C.A. Cooper, K.L. Wilkins, M.R.and Williams, and N.H. Packer. BOLD-A Biological O-linked Glycan Database. *Electrophoresis*, 20(18):3589–3598, 1999.

[28] P.M. Cutinho and B. Henrissat. Carbohydrate-Active Enzymes: An Integrated Database Approach. In H.J. Gilbert, G. Davies, B. Henrissat, and B. Svensson, editors, *Recent Advances in Carbohydrate Engineering*, pages 3–12. The Royal Society of Chemistry, Cambridge, UK, 1999.

[29] G. Cybenko. Approximations by Superpositions of a Sigmoid Function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

[30] P. Dayan and I.F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* MIT Press, Cambridge, MA, U.S.A., 2001.

[31] J.W. Dennis, Granovsky, M., and C.E. Warren. Protein glycosylation in development and disease. *Bioessays*, 21(5):412–421, 1999.

[32] J.W. Dennis, M. Granovsky, and C.E. Warren. Glycoprotein, Glycosylation and Cancer Progression. *Biochimica et Biophysica Acta*, 1473(1):21–34, 1999.

[33] A Dricu, M Carlberg, M Wang, and O Larsson. Inhibition of N-linked glycosylation using tunicamycin causes cell death in malignant cells: role of down-regulation of the insulin-like growth factor 1 receptor in induction of apoptosis. *Cancer Research*, 57(3):543–548, 1997.

[34] B Eisenhaber, P Bork, and F Eisenhaber. Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. *Protein Engineering*, 11(12):1155–1161, 1998.

[35] N. Farriol-Mathis, J.S. Garavelli, B. Boeckmann, S. Duvaud, E. Gasteiger, A. Gateau, A.L. Veuthey, and A. Bairoch. Annotation of Post-Translational Modifications in the Swiss-Prot Knowledge base. *Proteomics*, 6(4):1537–1550, 2004.

[36] L. Fausett. *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1994.

[37] R.E. Favoni and A. De Cupis. The Role of Polypeptide Growth Factors in Human Carcinomas: New Targets for a Novel Pharmacological Approach. *Pharmacol. Rev.*, 52(2):179–206, 2000.

[38] L. Felten and R. Jozefowicz. *Netter's Atlas of Human Neuroscience.* Elsevier Science Ltd., Oxford, UK, 2003.

[39] R.D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer, and A. Bateman. Pfam: Clans, Web Tools and Services. *Nucleic Acids Research*, 34(suppl_1):D247–251, 2006.

[40] M.J. Fitch, L. Campagnolo, F. Kuhnert, and H. Stuhlmann. EGFL7, a Novel Epidermal Growth Factor-Domain Gene Expressed in Endothelial Cells. *Developmental Dynamics*, 230(2):316–324, 2004.

[41] S. Gamou and N. Shimizu. Glycosylation of the Epidermal Growth Factor Receptor and Its Relationship to Membrane Transport and Ligand Binding. *J Biochem (Tokyo)*, 104(3):388–396, 1988.

[42] J.S. Garavelli. The RESID Database of Protein Modifications: 2003 developments. *Nucleic Acids Research*, 31(1):499–501, 2003.

[43] V.I Glivenko. Sur quelques points de la logique de M. Brouwer. *Acadḗmie royale de Belgique, Bulletin de la classe des sciences*, 15(5):183–188, 1929.

[44] D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley Professional, Boston, MA, U.S.A., January 1989.

[45] J. Gorodkin. Comparing two K-category Assignments by a K-category Correlation Coefficient. *Computational Biology and Chemistry*, 28(5-6):367–374, 2004.

[46] S. Grunewald, G. Matthijs, and J. Jaeken. Congenital Disorders of Glycosylation: A Review. *Pediatr. Res.*, 52(5):618–624, 2002.

[47] J.R. Gulcher, D.E. Nies, L.S. Marton, and K. Stefansson. An Alternatively Spliced Region of the Human Hexabrachion Contains a Repeat of Potential N-glycosylation Sites. *Proc. Natl. Acad. Sci. U S A.*, 86(5):1588–1592, 1989.

[48] R. Gupta, H. Birch, K. Rapacki, S. Brunak, and J.E. Hansen. O-GLYCBASE Version 4.0: A Revised Database of O-glycosylated Proteins. *Nucleic Acids Research*, 27(1):370–372, 1999.

[49] R. Gupta and S. Brunak. Prediction of Glycosylation Across the Human Proteome and the Correlation to Protein Function. *Pac. Symp. Biocomput.*, pages 310–322, 2002.

[50] R. Gupta, J. Hansen, and S. Brunak. Identifying Intracellular O-(beta)-GlcNAc 'yin-yang' Switches in the Available Human Proteome. Manuscript in preparation.

[51] R Gupta, E Jung, AA Gooley, KL Williams, S Brunak, and J Hansen. Scanning the Available *Dictyostelium discoideum* Proteome for O-linked GlcNAc Glycosylation Sites Using Neural Networks. *Glycobiology*, 9(10):1009–1022, 1999.

[52] M Hagan and M. Menhaj. Training Feedforward Networks with the Marquardt Algorithm. *IEEE Transactions on Neural Networks*, 5(6):989–993, 1994.

[53] M.T. Hagan and H.B. Demuth. Neural Networks for Control. In *American Control Conference*, volume 3, pages 1642–1656, 1999.

[54] S Hakomori. Tumor Malignancy Defined by Aberrant Glycosylation and Sphingo(glyco) Lipid Metabolism. *Cancer Research*, 56(23):5309–5318, 1996.

[55] S. Hakomori. Glycosylation Defining Cancer Malignancy: New Wine in an Old Bottle. *Proc. Natl. Acad. Sci. U S A.*, 99(16):10231–10233, 2002.

[56] Haltiwanger Laboratory, Department of Biochemistry and Cell Biology, State University of New York (SUNY) at Stony Brook , 2005.

[57] R.S. Haltiwanger and J.B. Lowe. Role of Glycosylation in Development. *Annual Review of Biochemistry*, 73(1):491–537, 2004.

[58] H.C. Hang and C.R. Bertozzi. The chemistry and biology of mucin-type o-linked glycosylation. *Bioorganiz and Medical Chemistry*, 36(47):5021–5034, 2005.

[59] R.J. Harris and M.W. Spellman. O-Linked Fucose and Other Post-Translational Modifications Unique to EGF Modules. *Glycobiology*, 3(3):219–224, 1993.

[60] P. Heitzler and P. Simpson. Altered Epidermal Growth Factor-like Sequences Provide Evidence for a Role of Notch as a Receptor in Cell Fate Decisions. *Development*, 117(3):1113–1123, 1993.

[61] A. Helenius. How N-linked Oligosaccharides Affect Glycoprotein Folding in the Endoplasmic Reticulum. *Molecular Biology of the Cell*, 5:253–265, 1994.

[62] H.B.K. Henrick and H. Nakamura. Announcing the Worldwide Protein Data Bank. *Nature Structural and Molecular Biology*, 10(12):980, 2003.

[63] J.L. Hermann, R. Delahay, A. Gallagher, B. Robertson, and D. Young. Analysis of Post-translational Modification of Mycobacterial Proteins Using a Cassette Expression System. *FEBS Letters*, 473(3):358–362, 2000.

[64] K. Hornik. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, 4(2):251–257, 1991.

[65] K. Hornik, M. Stinchcombe, and H. White. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5):359–366, 1989.

[66] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. Langendijk-Genevaux, M. Pagni, and C. J. A. Sigrist. The PROSITE database. *Nucleic Acids Research*, 34(suppl_1):D227–230, 2006.

[67] B. Jagla and J. Schuchhardt. Adaptive Encoding Neural Networks for the Recognition of Human Signal Peptide Cleavage Sites. *Bioinformatics*, 16(3):245–250, 2000.

[68] L.J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, , C. Kesmir, H. Nielsen, H.H. Stæfeldt, K. Rapacki, C. Workman, C.A.F. Andersen, S. Knudsen, A. Krogh, A. Valencia, and S. Brunak. Prediction of human protein function from post-translational modification and localization features. *Journal of Molecular Biology*, 319(5):1257–1265, 2002.

[69] L.J. Jensen, M. Skovgaard, and S. Brunak. Prediction of Novel Archaeal Enzymes from Sequence-derived Features. *Protein Science*, 11(12):2894–2898, 2002.

[70] K. Julenius, A. Mølgaard, R. Gupta, and S. Brunak. Prediction, Conservation Analysis, and Structural Characterization of Mammalian Mucin-type O-glycosylation Sites. *Glycobiology*, 15(2):153–164, 2005.

[71] H. Kadokura, F. Katzen, and J. and Beckwith. Protein Disulfide Bond Formation in Prokaroytes. *Annual Review of Biochemistry*, 72(1):111–135, 2003.

[72] M. Karpinski and A. Macintyre. Polynomial bounds for VC dimension of sigmoidal neural networks. In *STOC '95: Proceedings of the Twenty-seventh Annual ACM Symposium on Theory of Computing*, pages 200–208, New York, NY, USA, 1995. ACM Press.

[73] A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of Sets in Functional Space. *American Mathematical Society Translations (2)*, 17:277–364, 1961.

[74] M. A. Kon and L. Plaskota. Information Complexity of Neural Networks. *Neural Networks*, 13(3):365–375, 2000.

[75] F.J. Krambeck and M.J. Betenbaugh. A Mathematical Model of N-linked Glycosylation. *Biotechnology and Bioengineering*, 92(6):711–728, 2005.

[76] A. Krzyżak and T. Linder. Radial Basis Function Networks and Complexity Regularization in Function Learning. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 197–203, Cambridge, MA, U.S.A., 1996. MIT Press.

[77] M. A. Kukuruzinska and K. Lennon. Protein N-glycosylation: Molecular Genetics and Functional Significance. *Crit Rev Oral Biol Med*, 9(4):415–448, 1998.

[78] W. Li, L. Jaroszewski, and A. Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, 2001.

[79] H. Lin, M. Stacey, C. Saxby, V. Knott, Y. Chaudhry, D. Evans, S. Gordon, A.J. McKnight, P. Handford, and S. Lea. Molecular Analysis of the Epidermal Growth Factor-like Short Consensus Repeat Domain-mediated Protein-Protein Interactions. DISSECTION OF THE CD97-CD55 COMPLEX. *Journal of Biological Chemistry*, 276(26):24160–24169, 2001.

[80] K. Lin, A.C.W. May, and W.R. Taylor. Amino Acid Encoding Schemes from Protein Structure Alignments: Multi-dimensional Vectors to Describe Residue Types. *Journal of Theoretical Biology*, 216(3):361–365, 2002.

[81] H. Lis and N. Sharon. Protein Glycosylation: Structural and Functional Aspects. *European Journal of Biochemistry*, 218(1):1–27, 1993.

[82] T. Lütteke, A. Bohne-Lang, A. Loss, T. Goetz, M. Frank, and C.-W. von der Lieth. GLYCOSCIENCES.de: an Internet Portal to Support Glycomics and Glycobiology Research. *Glycobiology*, 16(5):71R–81, 2006.

[83] T. Lütteke, M. Frank, and C.-W. von der Lieth. Carbohydrate Structure Suite (CSS): Analysis of Carbohydrate 3D Structures Derived from the PDB. *Nucleic Acids Research*, 33(suppl_1):D242–246, 2005.

[84] W. Maass. Vapnik-Chervonenkis Dimension of Neural Networks. In Michael A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 522–526. MIT Press, Cambridge, Massachusetts, 1995.

[85] D. J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computing*, 4(3):448–472, 1992.

[86] D. J. C. MacKay. Bayesian Interpolation. *Neural Computing*, 4(3):415–447, 1992.

[87] R.D. Marshall. Glycoproteins. *Annual Review of Biochemistry*, 41(1):673–702, 1972.

[88] MATLAB, The Mathworks, Inc. World Wide Web, `http://www.mathworks.com`.

[89] B.W. Matthews. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta*, 405(2):441–451, 1975.

[90] H. Mayer, R. Huber, and R. Schwaiger. Lean Artificial Neural Networks - Regularization Helps Evolution. In *Nordic Workshop on Genetic Algorithms*, pages 163–172, 1996.

[91] W.S. McCulloch and W. Pitts. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[92] M.L. Minsky and S.A. Papert. *Perceptrons.* MIT Press, Cambridge, MA, USA, 1965.

[93] M. Møller. A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. *Neural Networks*, 6(4):525–533, 1993.

[94] T.J. Monica, D.C. Andersen, and C.F. Goochee. A Mathematical Model of Sialylation of N-linked Oligosaccharides in the Trans-Golgi Network. *Glycobiology*, 7(4):515–521, 1997.

[95] N.J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A. N. Nikolskaya, S. Orchard, M. Pagni, C. P. Ponting, E. Quevillon, J. Selengut, C. J. A. Sigrist, V. Silventoinen, D. J. Studholme, R. Vaughan, and C.H. Wu. InterPro, Progress and Status in 2005. *Nucleic Acids Research*, 33(suppl_1):D201–205, 2005.

[96] D. Näf, D.M. Krupke, J.P. Sundberg, J.T. Eppig, and C.J. Bult. The Mouse Tumor Biology Database: A Public Resource for Cancer Genetics and Pathology of the Mouse. *Cancer Research*, 62(5):1235–1240, 2002.

[97] D. Nguyen and B. Widrow. Improving the Learning Speed of 2-Layer Neural Networks by Choosing initial Values of the Adaptive Weights. In *International Joint Conference on Neural Networks*, volume III, pages 21–26, San Diego, CA, U.S.A., 1990.

[98] N. Normanno and F. Ciardiello. EGF-related Peptides in the Pathophysiology of the Mammary Gland. *Journal of Mammary Gland Biolology Neoplasia*, 2(2):143–151, 1997.

[99] J.M.C. Oh, F. Brichory, E. Puravs, R. Kuick, C. Wood, J.M. Rouillard, J. Tra, D. Beer S. Kardia, and S. Hanash. A Database of Protein Expression in Lung Cancer. *Proteomics*, 1(10):1303–1319, 2001.

[100] B.C. Paria, K. Elenius, M. Klagsbrun, and S.K. Dey. Heparin-binding EGF-like Growth Factor Interacts with Mouse Blastocysts Independently of ErbB1: A Possible Role for Heparan Sulfate Proteoglycans and ErbB4 in Blastocyst Implantation. *Development*, 126(9):1997–2005, 1999.

[101] A.J. Petrescu, A.L. Milac, S.M. Petrescu, R.A. Dwek, and M.R. Wormald. Statistical Analysis of the Protein Environment of N-glycosylation Sites: Implications for Occupancy, Structure, and Folding. *Glycobiology*, 14(2):103–114, 2004.

[102] E. Jung R. Gupta and S. Brunak. Prediction of N-glycosylation Sites in Human Proteins. In preparation, 2004.

[103] M. Riedmiller and H. Braun. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In *The IEEE International Conference on Neural Networks*, pages 586–591, San Francisco, CA, 1993.

[104] S.K. Riis and A. Krogh. Improving Prediction of Protein Secondary Structure Using Structured Neural Networks and Multiple Sequence Alignments. *Journal of Computational Biology*, 3:163–183, 1996.

[105] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in*

the Microstructure of Cognition, Vol. 1: Foundations, volume 1, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

[106] S. Sahoo, A. Sheth, W. York, and J.A. Miller. Semantic Web Services for N-glycosylation Process. In *International Symposium on Web Services for Computational Biology and Bioinformatics*, Blacksburg, VA, U.S.A., May 2005.

[107] D.S. Salomon, R. Brandt, F. Ciardiello, and N. Normanno. Epidermal growth factor-related peptides and their receptors in human malignancies. *Critical Reviews in Oncology/Hematology*, 19(3):183–232, 1995.

[108] L.E. Scales. *Introduction to Non-linear Optimization.* Springer Verlag, New York, NY, U.S.A., 1985.

[109] M. Scanlan, G. Ritter, B.W.T. Yin, C. Williams Jr., L.S. Cohen, S.R. Fortunato, D. Frosina, S-.Y. Lee, A.E. Murray, R. Chua, V.V. Filonenko, E. Sato, L.J. Old, and A.A. Jungbluth. Glycoprotein a34, a novel target for antibody-based cancer immunotherapy. *Cancer Immunity*, 6:2–10, 2006.

[110] A. Shaneh and G. Butler. The Neural Detection of Glycosylation Sites of the Epidermal Growth Factor-Like Repeat Proteins (EGFL) of Mammalian Cells. In *Mathematical Programming and Data Mining (MPDM)*, Hamilton, ON, Canada, 2005. The Fields Institute / IBM / MITAC Workshop.

[111] A. Shaneh and G. Butler. Bayesian Learning for Feed-Forward Neural Network with Application to Proteomic Data: The Glycosylation Sites Detection of the Epidermal Growth Factor-Like Proteins Associated with Cancer as a Case Study. In L. Lamontagne and M. Marchand, editors, *Advances in Artificial Intelligence*, volume 4013 of *Lecture Notes in Artificial Intelligence*, pages 110–121. Springer Verlag, 2006.

[112] L. Shao, Y. Luo, D.J. Moloney, and Haltiwanger R. O-glycosylation of EGF repeats: identification and initial characterization of a UDP-glucose: protein O-glucosyltransferase. *Glycobiology*, 12(11):763–770, 2002.

[113] R.G. Spiro. Protein Glycosylation: Nature, Distribution, Enzymatic Formation, and Disease Implications of Glycopeptide Bonds. *Glycobiology*, 12(4):43R–56, 2002.

[114] S.R. Sunyaev, F. Eisenhaber, I.V. Rodchenkov, B. Eisenhaber, V.G. Tumanyan, and E.N. Kuznetsov. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Engineering*, 12(5):387–394, 1999.

[115] P. Umaña and J.E. Bailey. A Mathematical Model of N-linked Glycoform Biosynthesis. *Biotechnology and Bioengineering*, 55:890–908, 2000.

[116] P. Van den Steen, P.M. Rudd, R.A. Dwek, and G. Opdenakker. Concepts and Principles of O-linked Glycosylation. *Critical Review of Biochemistry and Molecular Biology*, 33(3):151–208, 1998.

[117] V.N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[118] Y. Wang, L. Shao, S. Shi, R.J. Harris, M.W. Spellman, P. Stanley, and R.S. Haltiwanger. Modification of Epidermal Growth Factor-like Repeats with O-Fucose. *Journal of Biological Chemistry*, 276(43):40338–40345, 2001.

[119] P.J. Werbos. Backpropagation: Past and Future. In *International Conference on Neural Networks*, volume 1, pages 343–354, San Diego, CA, U.S.A., 1988.

[120] S.A. Whelan and G.W. Hart. Proteomic Approaches to Analyze the Dynamic Relationships Between Nucleocytoplasmic Protein Glycosylation and Phosphorylation. *Circulation Research*, 93(11):1047–1058, 2003.

[121] B.M. Wilamowski, S. Iplikci, O. Kaynak, and M.O. Efe. An Algorithm for Fast Convergence in Training Neural Networks. In *International Joint Conference on Neural Networks*, volume 3, pages 1778–1782. Neural Networks, 2001.

[122] K.P. Willey. An Elusive Role for Glycosylation in the Structure and Function of Reproductive Hormones. *Human Reproduction Update*, 5(4):330–355, 1999.

[123] D.A. Winkler and F.R. Burden. Application of Neural Networks to Large Dataset QSAR, Virtual Screening, and Library Design. In L.B. English, editor, *Combinatorial Library Methods and Protocols*, volume 201 of *Methods in Molecular Biology*, chapter 21, pages 325–368. Humana Press, Inc., Totowa, NJ, U.S.A., August 2002.

[124] M.A. Wouters, I. Rigoutsos, C.K. Chu, L.L. Feng, D.B. Sparrow, and S.L. Dunwoodie. Evolution of Distinct EGF Domains with Specific Functions. *Protein Science*, 14(4):1091–1103, 2005.

[125] A. Yan and W.J. Lennarz. Unraveling the Mechanism of Protein N-Glycosylation. *Journal of Biological Chemistry*, 280(5):3121–3124, 2005.

[126] L. Zadeh. Fuzzy sets. *Journal of Information and Control*, 8:338–353, 1965.

[127] Y. Zhen, R.M. Caprioli, and J.V. Staros. Characterization of Glycosylation Sites of the Epidermal Growth Factor Receptor. *Biochemistry*, 42(18):5478–5492, 2003.

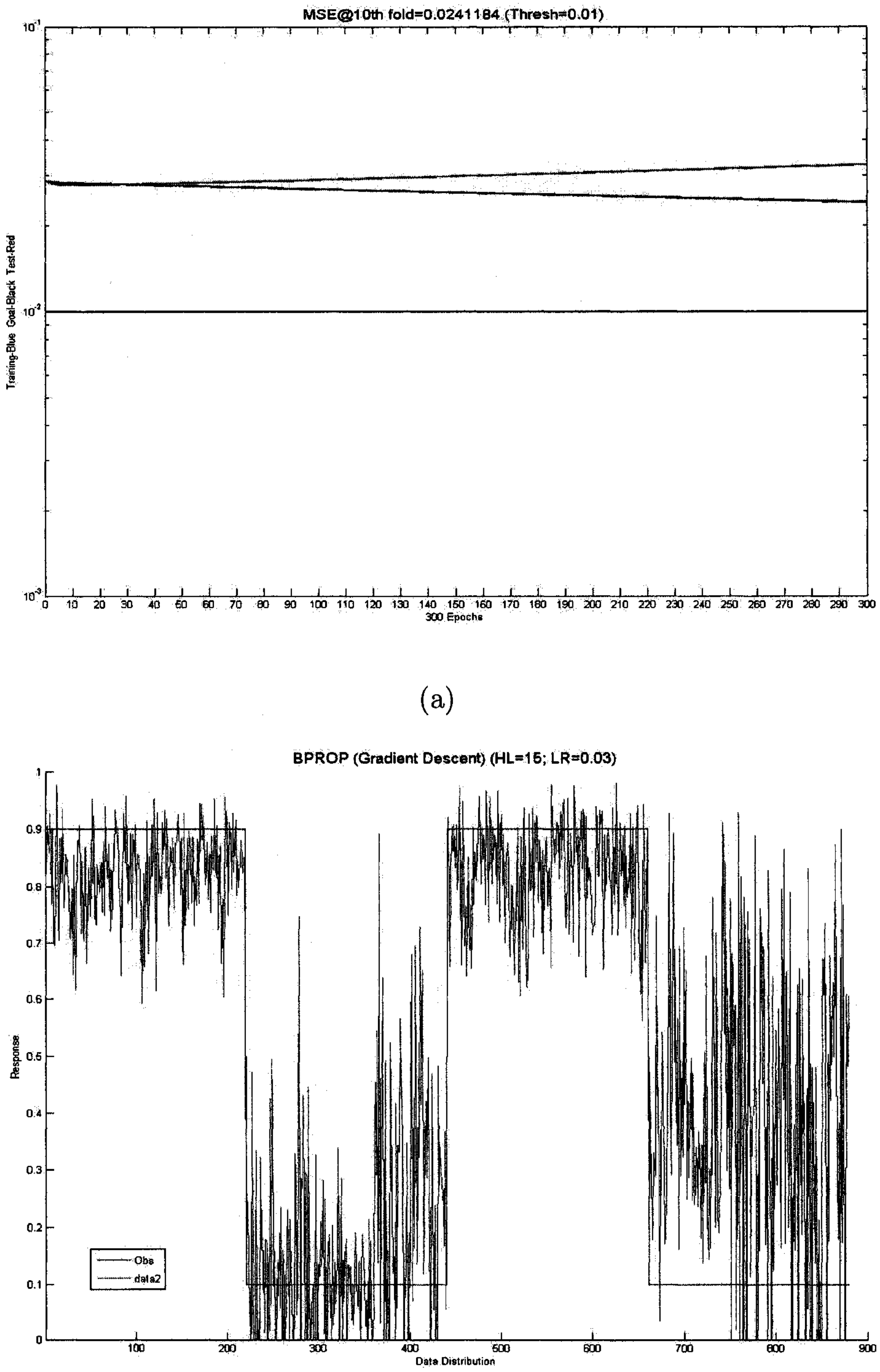
Appendix A

The Responses of Learning Algorithms of This Study

The results for comparing various learning algorithms with feed-forward neural network have been shown in Figures 27 to 56. [BPROP-OSS] was more consistent than others in terms of stability and performance [110]. Table 10 lists the designations of graph and algorithms.

Table 10: The Nomenclature Used for Learning Algorithms

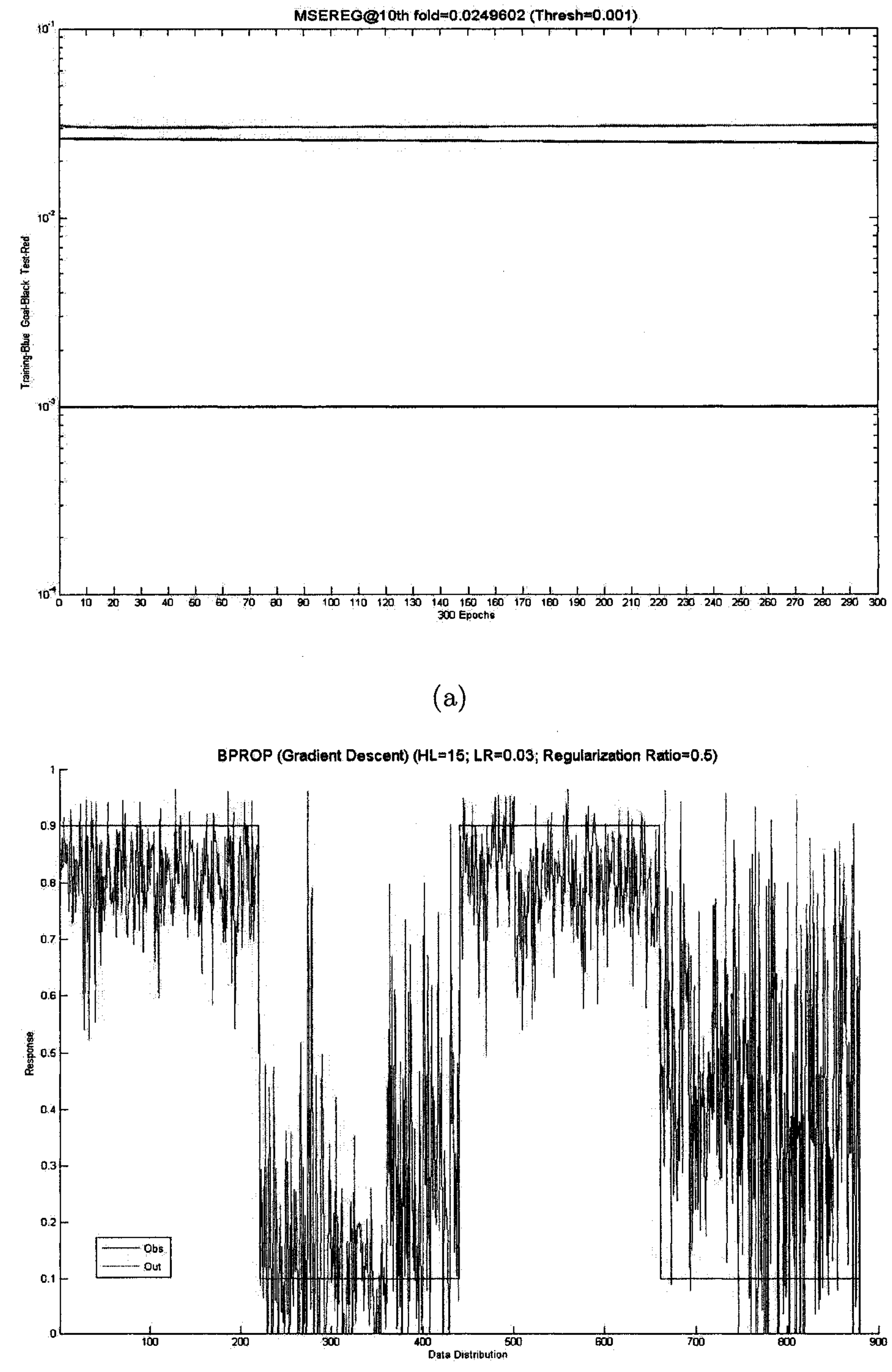
Symbol	Description
HU or *HL*	Number of hidden units in a layer
LR_i	The initial value for learning rate
LR	Learning rate
RR	Regularization ratio
$HLnm$	Two-layer neural network with n and m hidden units for each layer
MC	Momentum constant
MSE	Mean Squared Error
thresh or *goal*	The error tolerance
Performance	The greatest real number less than goal
WF	Window frame
OUT	Model response distribution
OBS	Real test set distribution



(a)

(b)

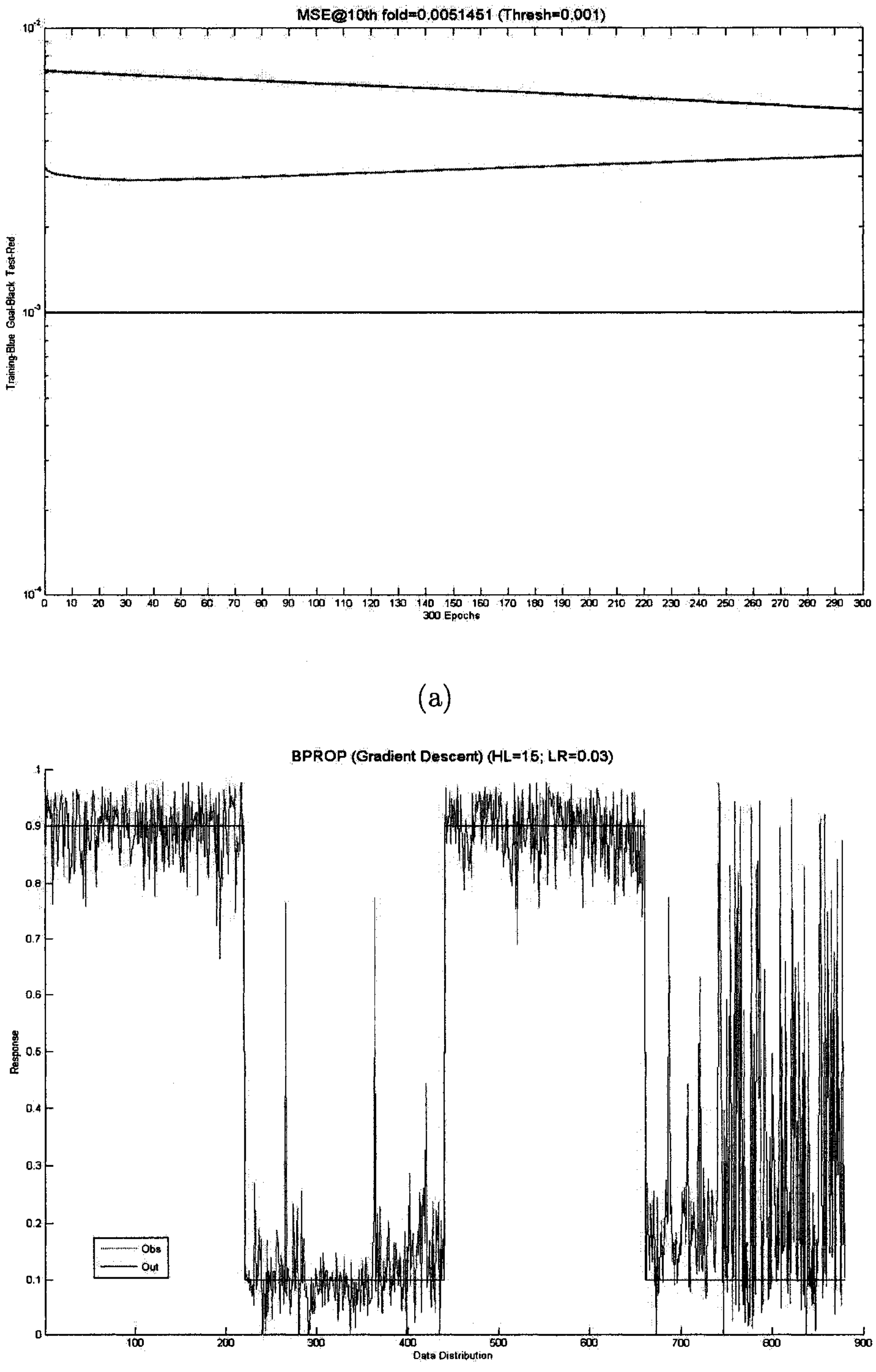
Figure 27: (a) [BPROP-GD] WF=29, HU=15 (b) Model Response



(a)

(b)

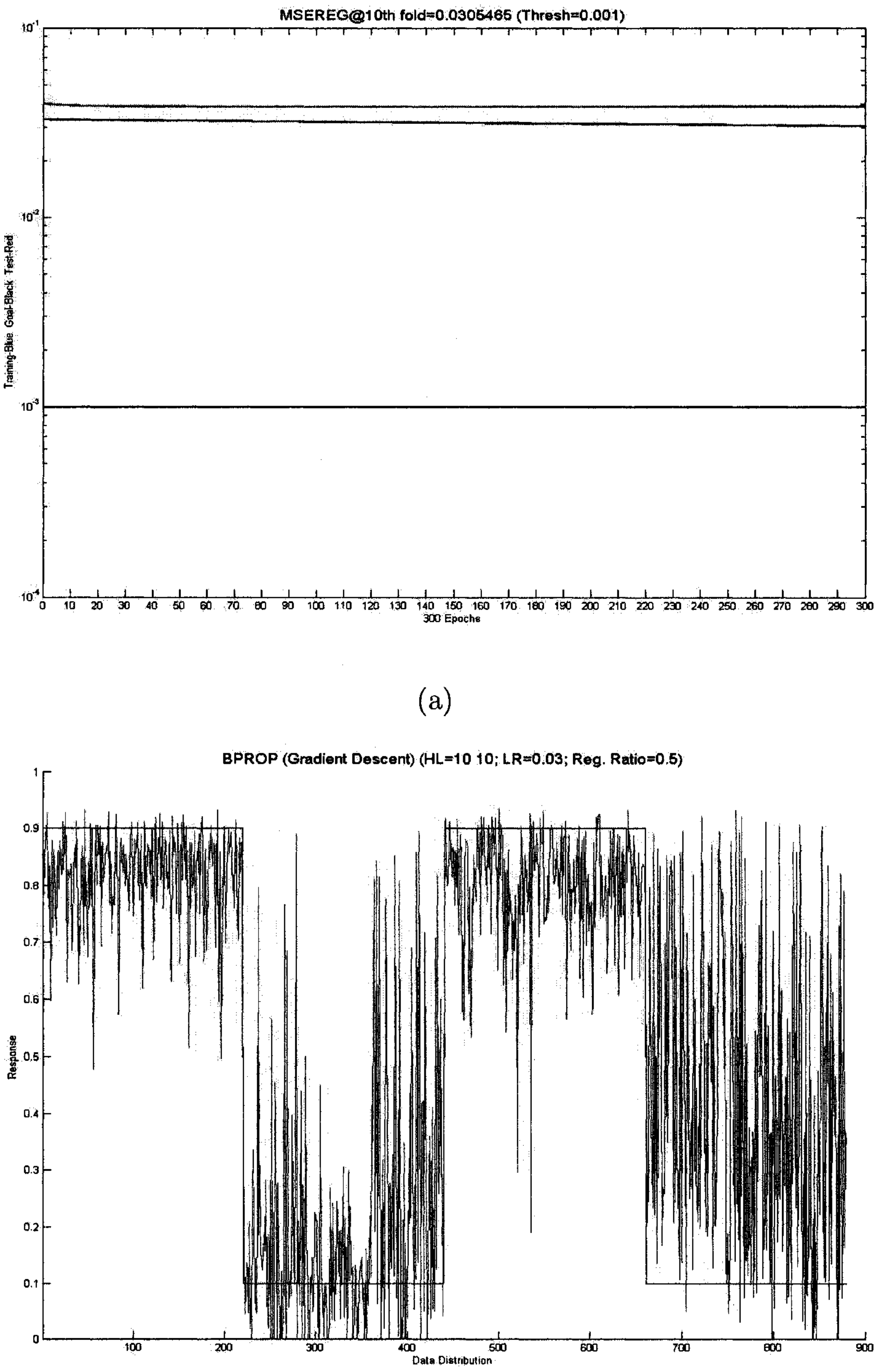
Figure 28: (a) [BPROP-GD] WF=29, HU=15, RR=0.5, LR_i=0.03 (b) Model Response



(a)

(b)

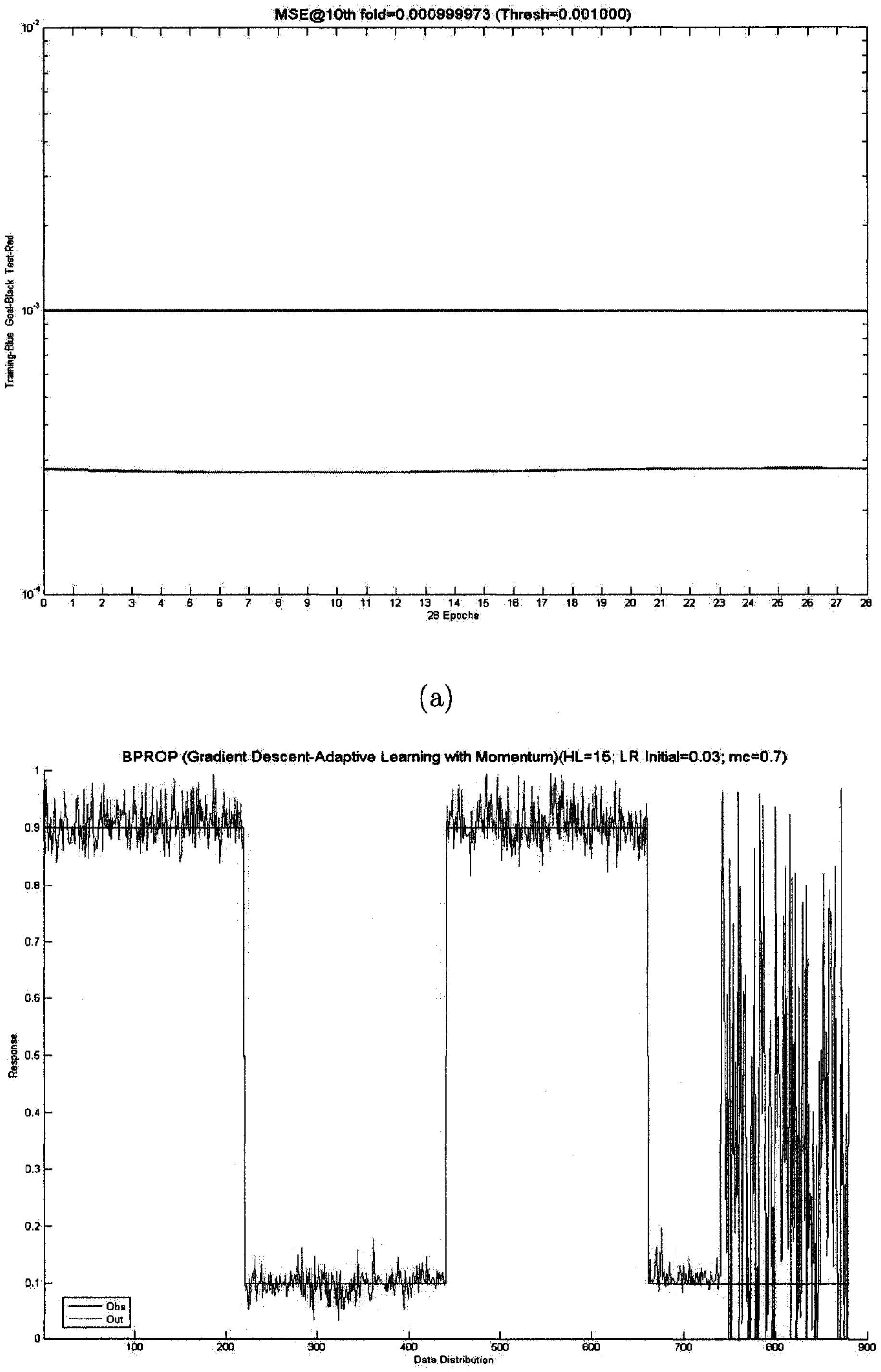
Figure 29: (a) [BPROP-GD] WF=29,HU=10 10,LR_i=0.03 (b) Model Response



(a)

(b)

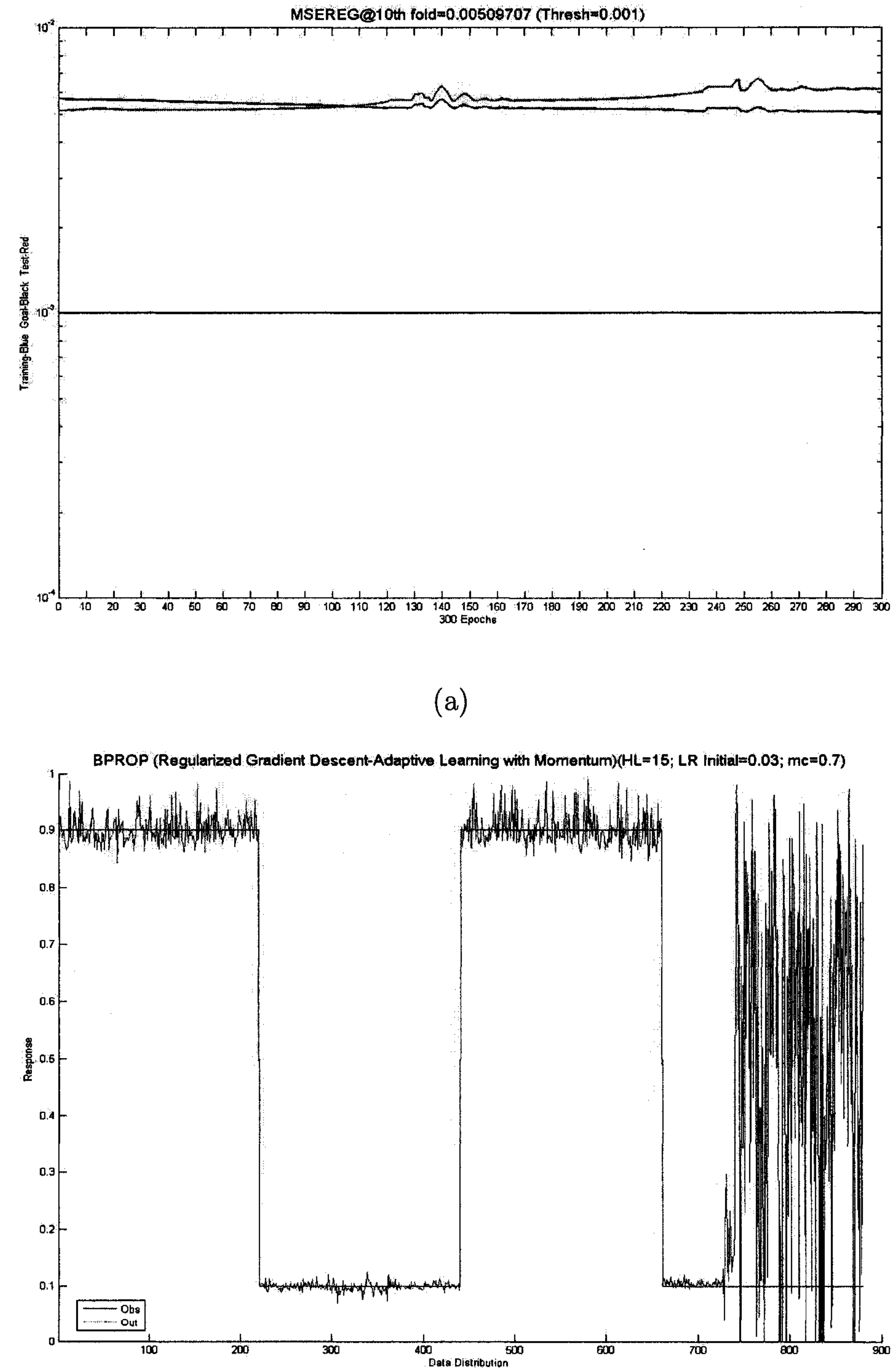
Figure 30: (a) [BPROP-GD] WF=29,HU=10 10,RR=0.5 (b) Model Response



(a)

(b)

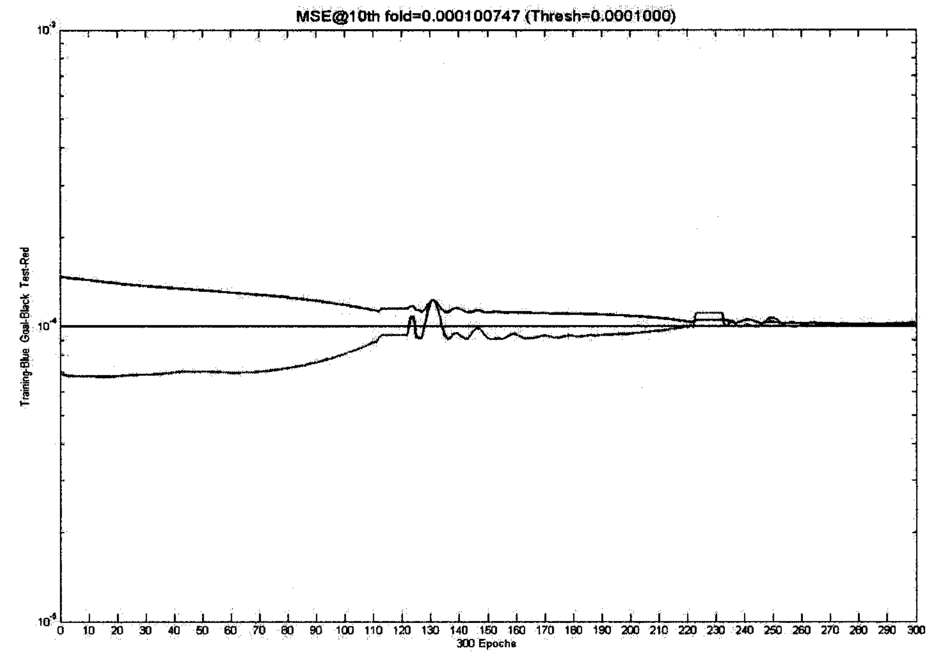
Figure 31: (a) [BPROP-GD-X] WF=29, HU=15 (b) Model Response



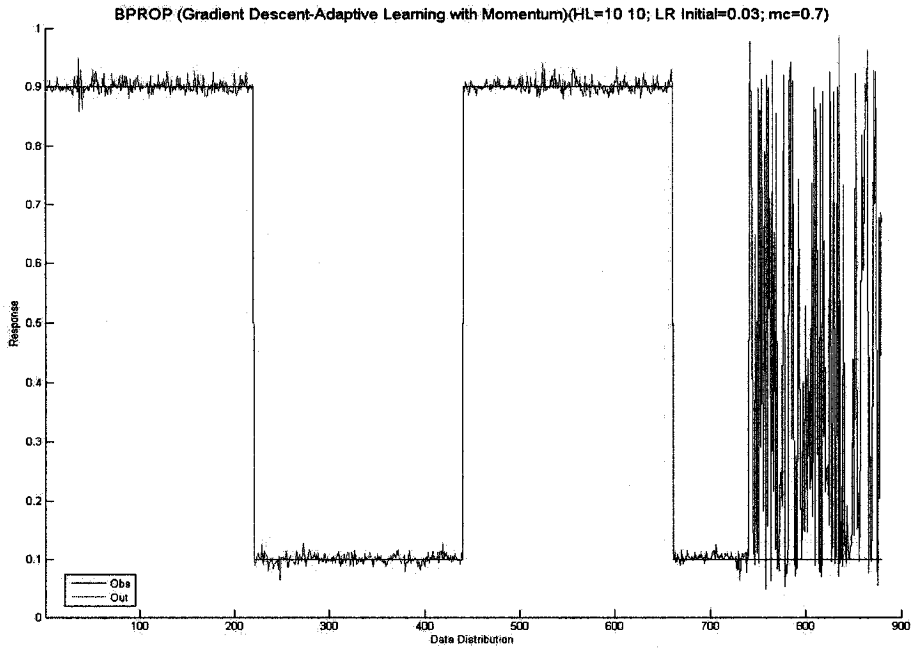
(a)

(b)

Figure 32: (a) [BPROP-GD-X] WF=29,HU=15,RR=0.9 (b) Model Response

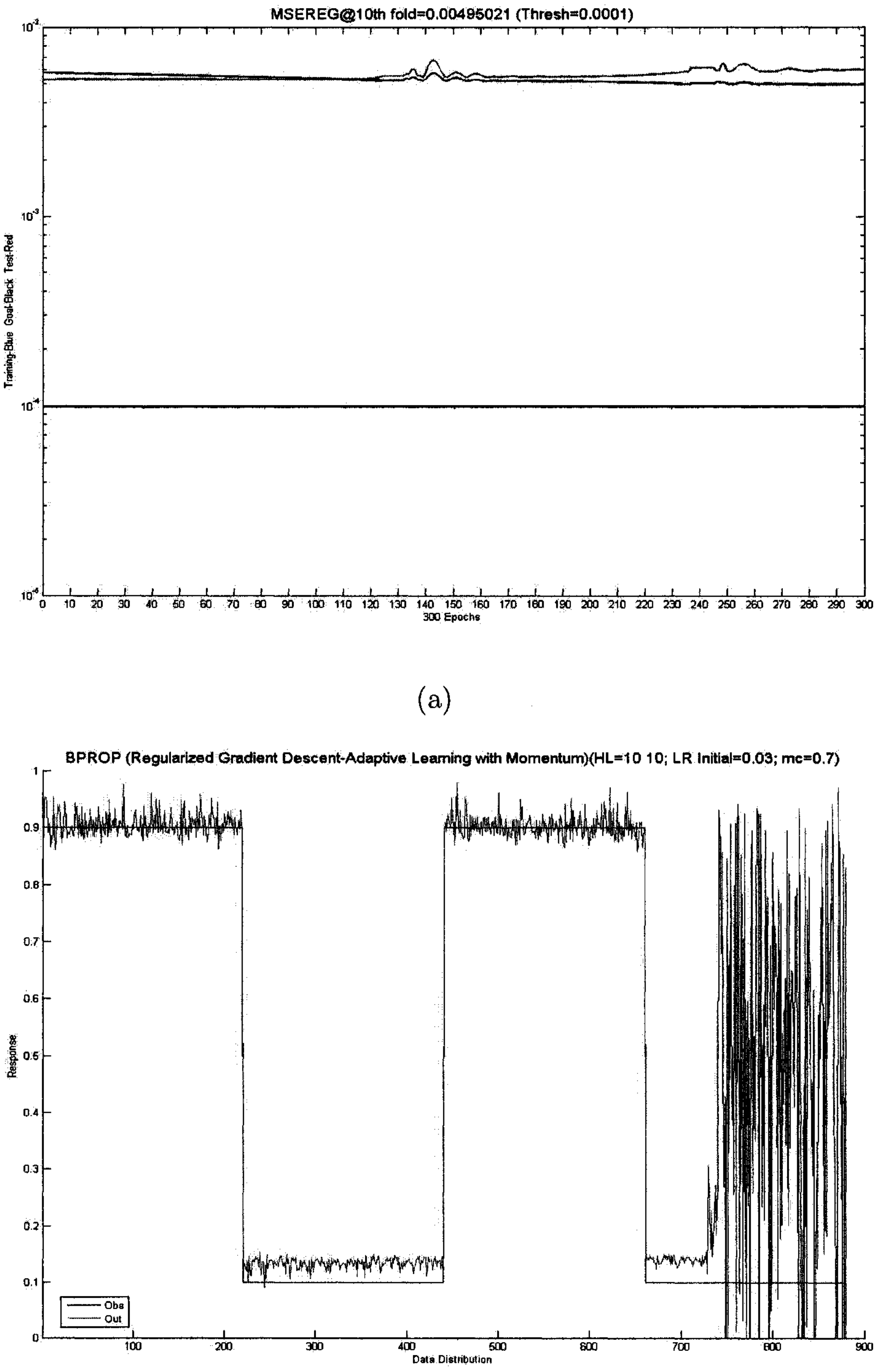


(a)



(b)

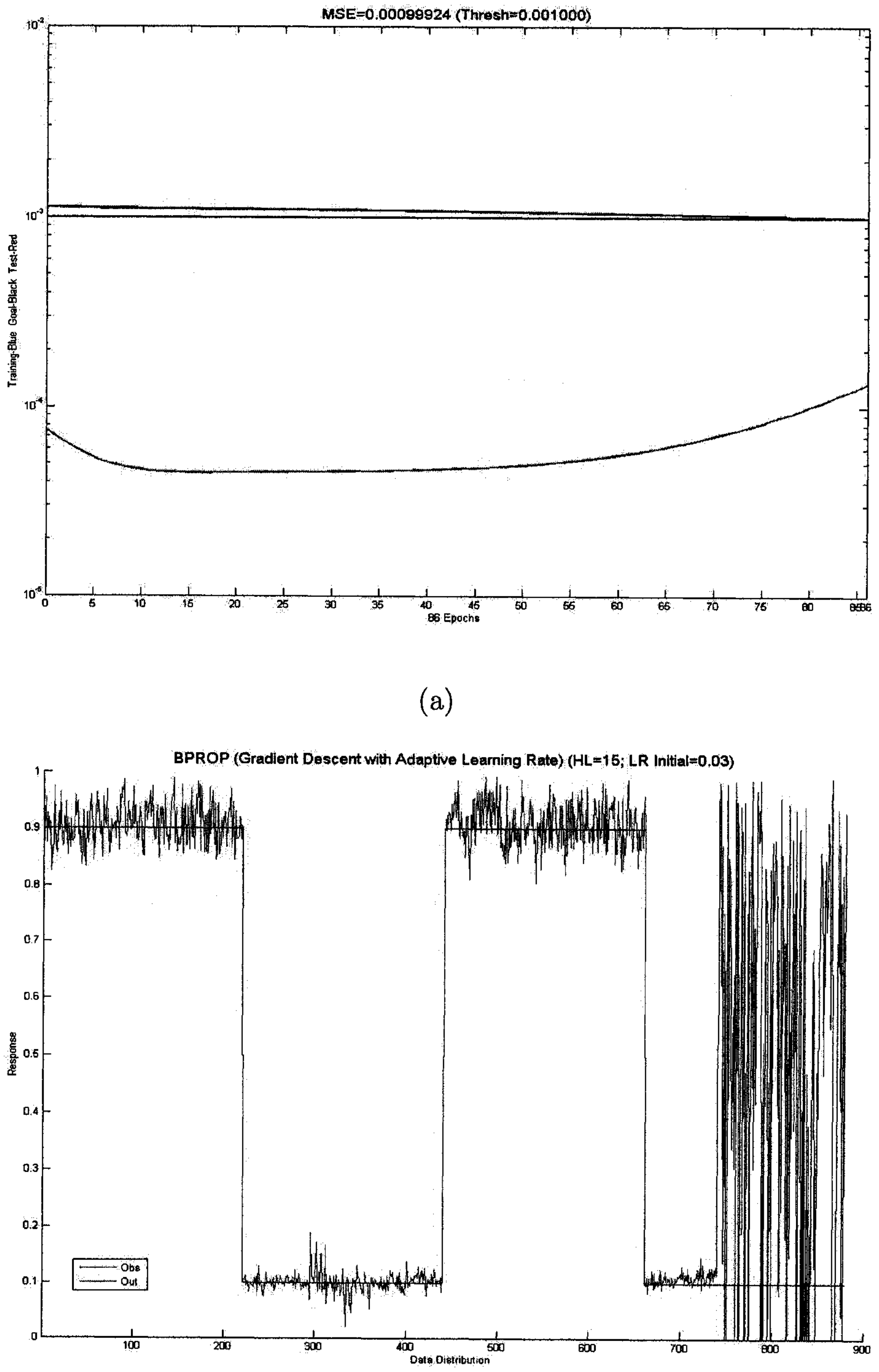
Figure 33: (a) [BPROP-GD-X] WF=29,HU=10 10 (b) Model Response



(a)

(b)

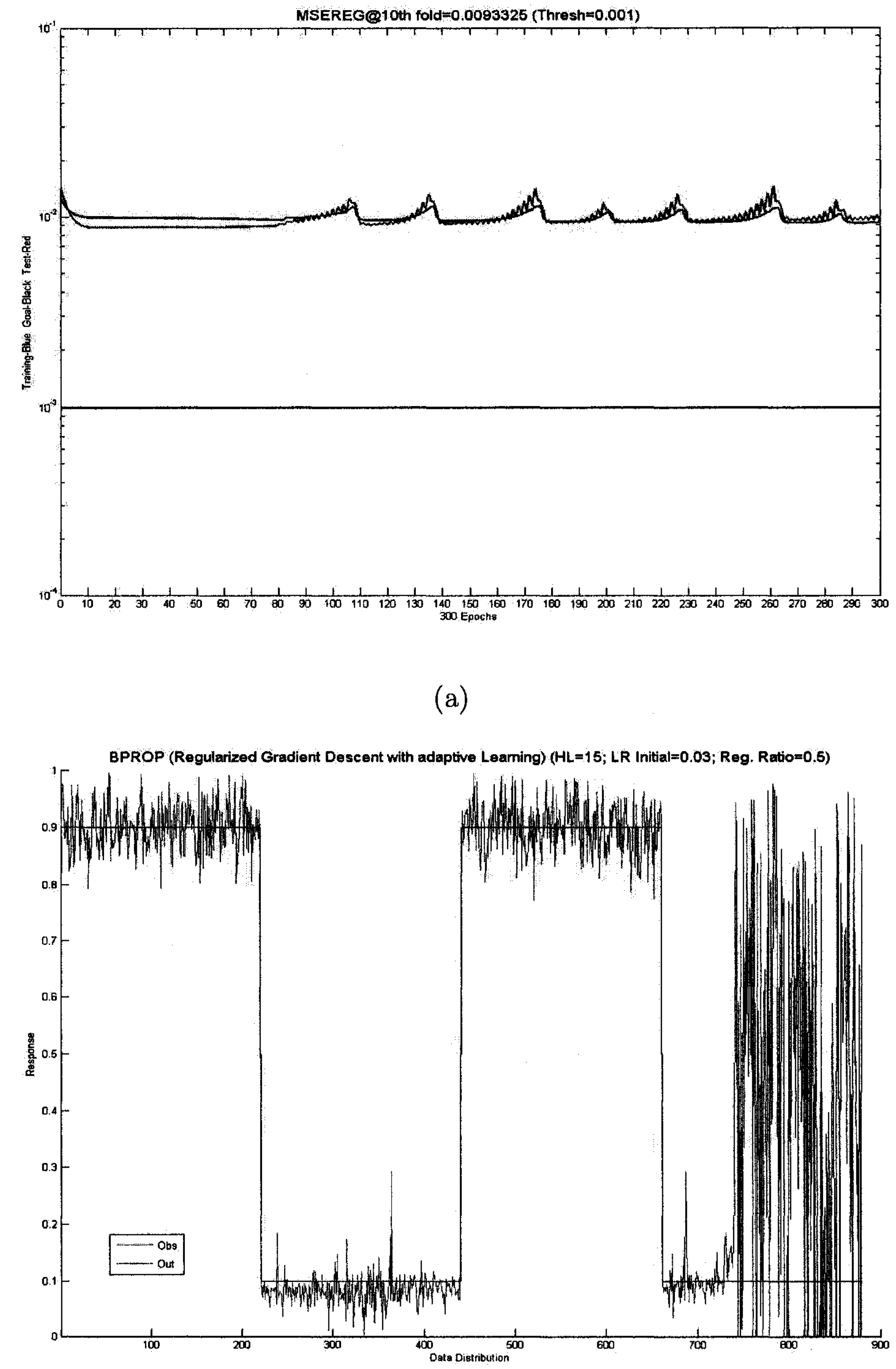
Figure 34: (a) [BPROP-GD-X] WF=29, HU=10 10 (b) Model Response



(a)

(b)

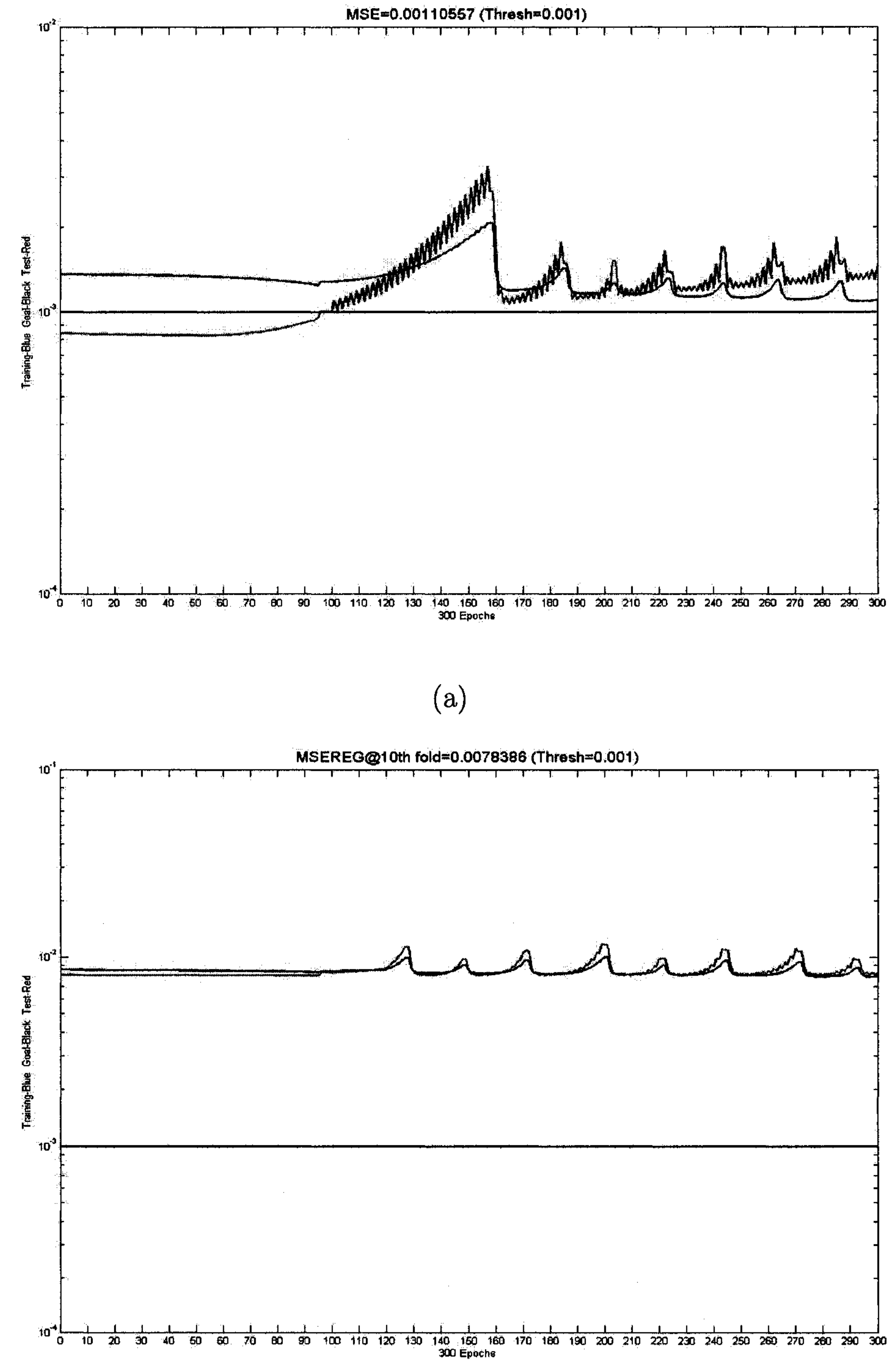
Figure 35: (a) [BPROP-GD-A] WF=29, HU=15 (b) Model Response



(a)

(b)

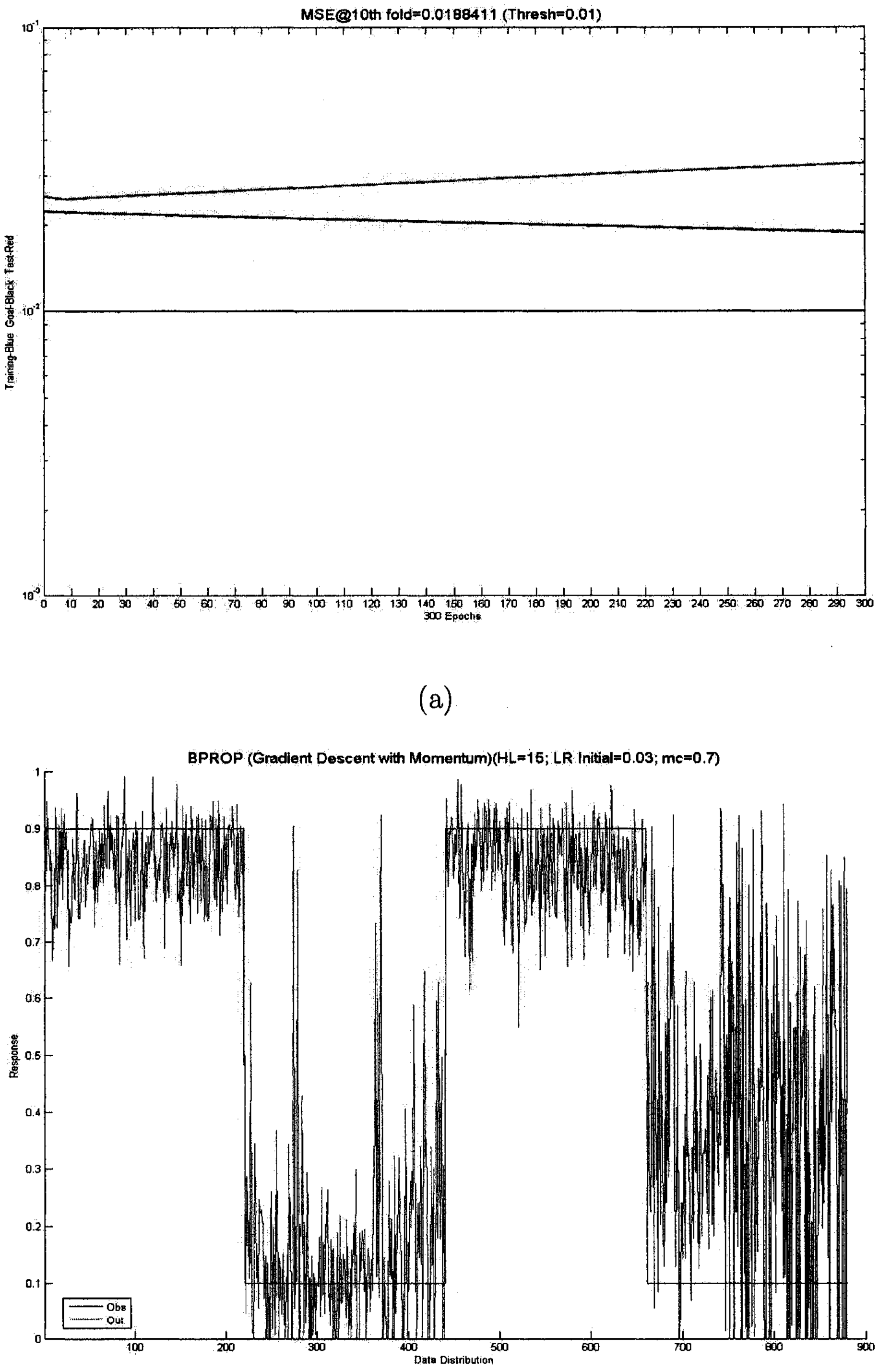
Figure 36: (a) [BPROP-GD-A] WF=29,HU=15,RR=0.5 (b) Model Response



(a)

(b)

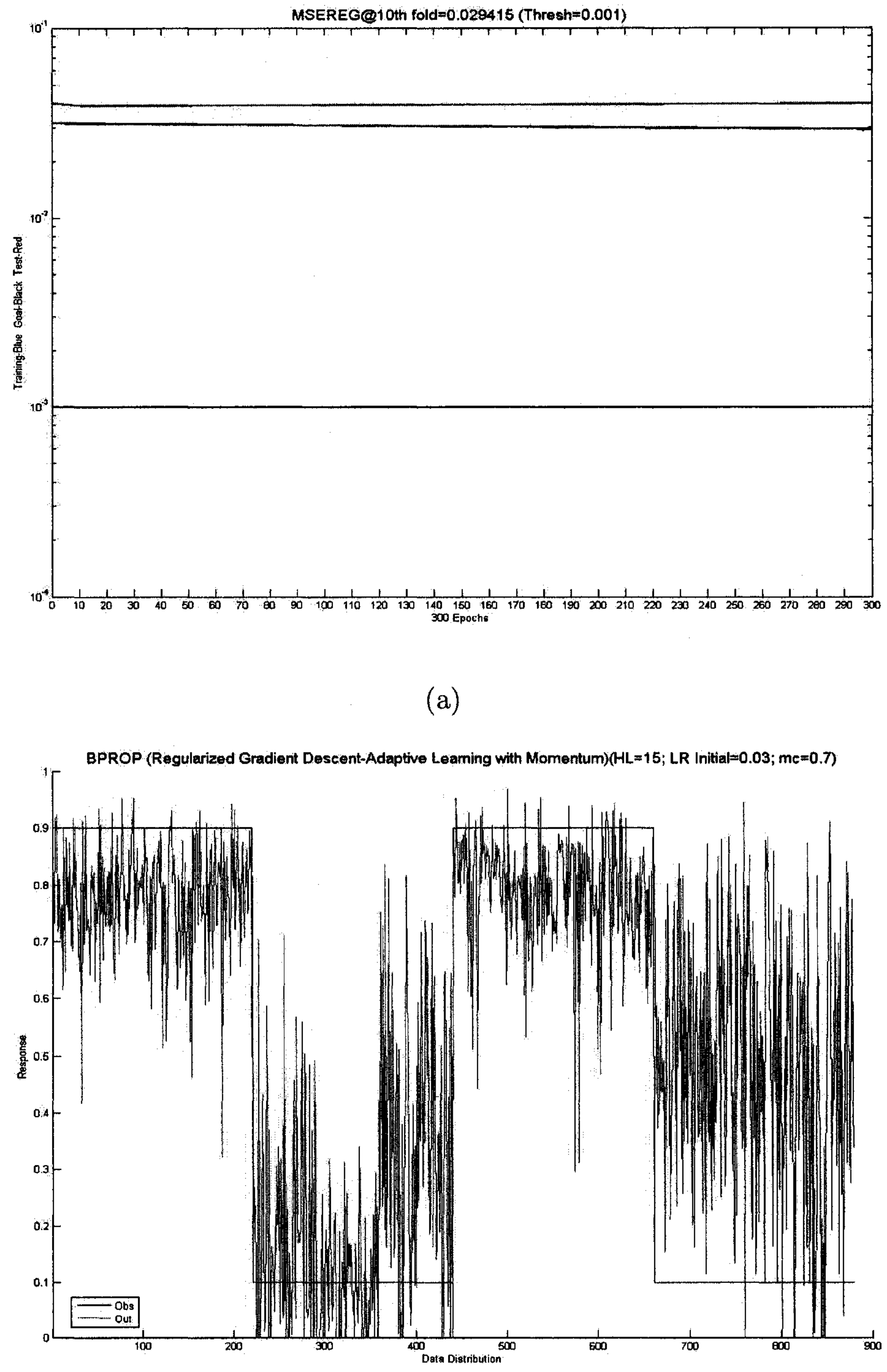
Figure 37: (a) [BPROP-GD-A] WF=29, HU=10 10 (b) Model Response



(a)

(b)

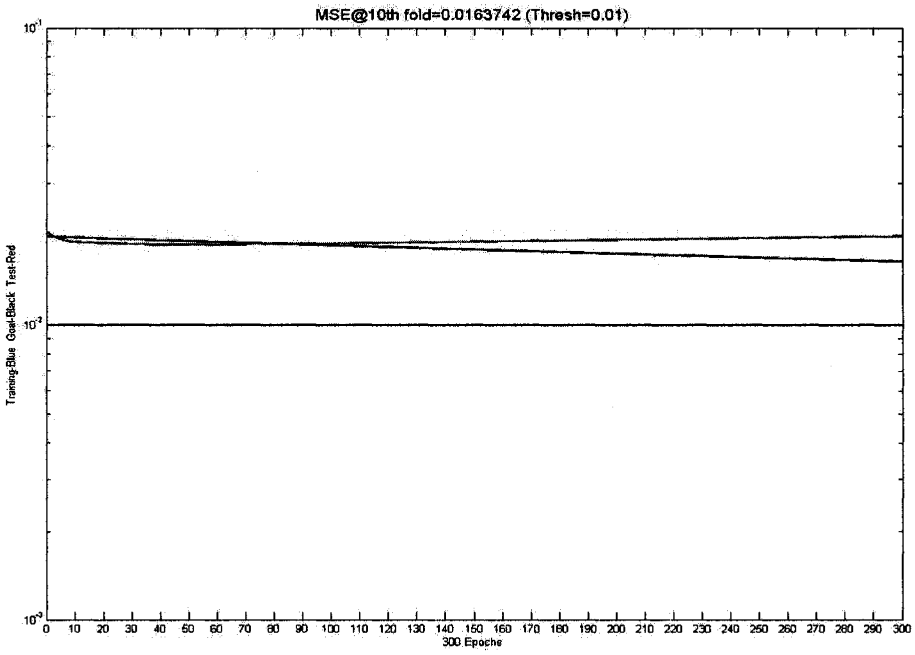
Figure 38: (a) [BPROP-GD-M] WF=29, HU=15 (b) Model Response



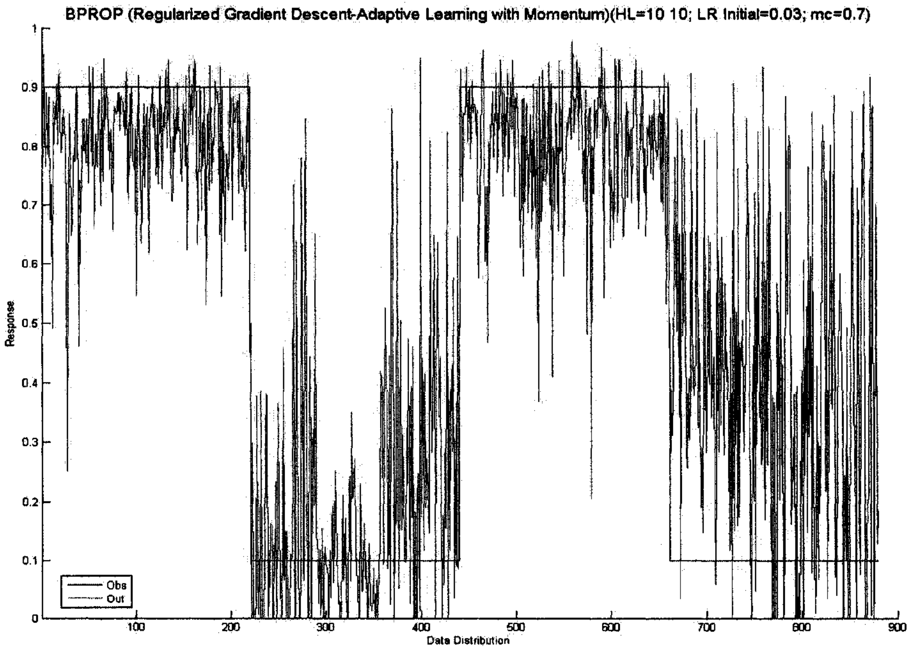
(a)

(b)

Figure 39: (a) [BPROP-GD-M] WF=29,HU=15,RR=0.5 (b) Model Response

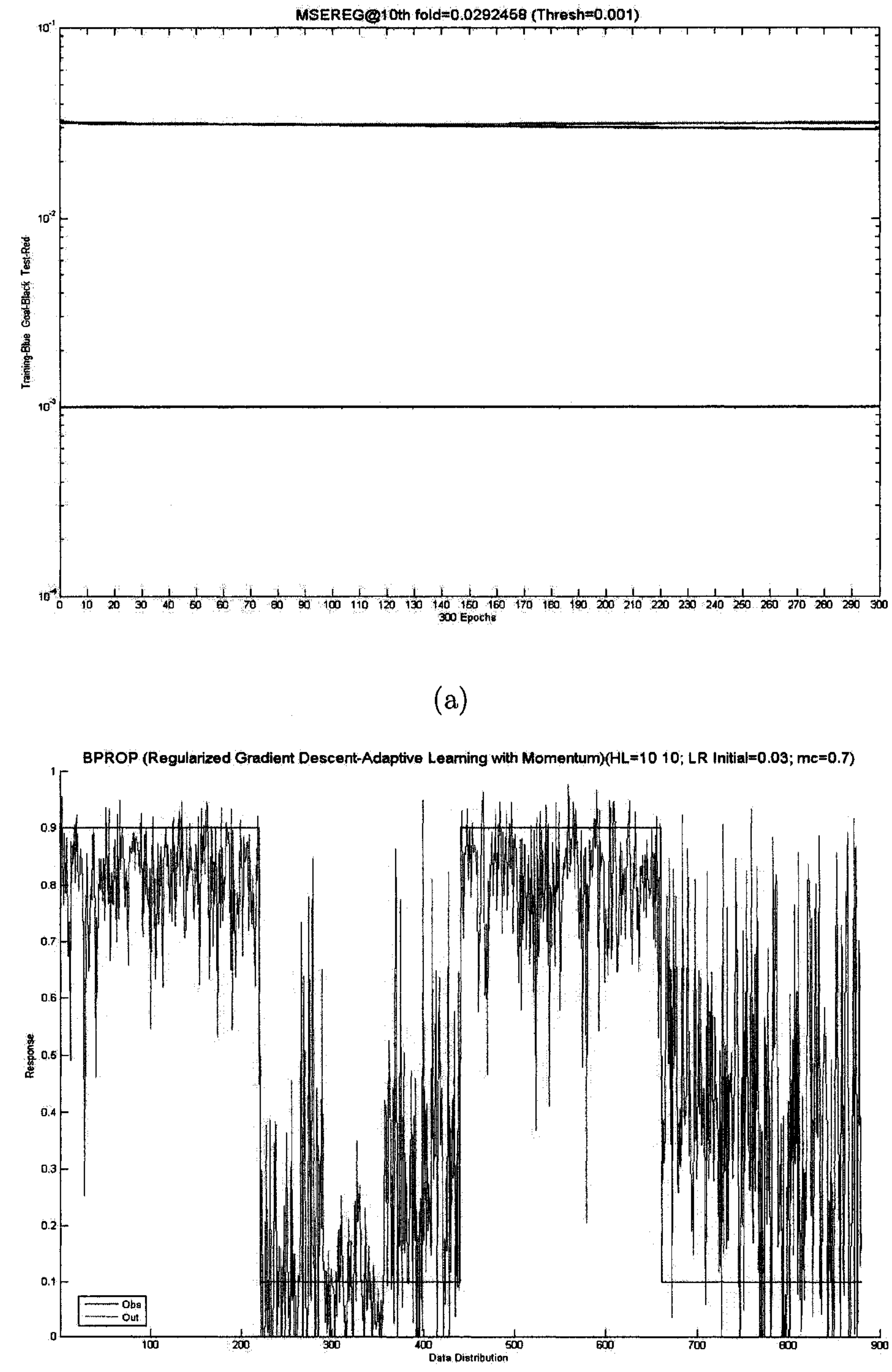


(a)



(b)

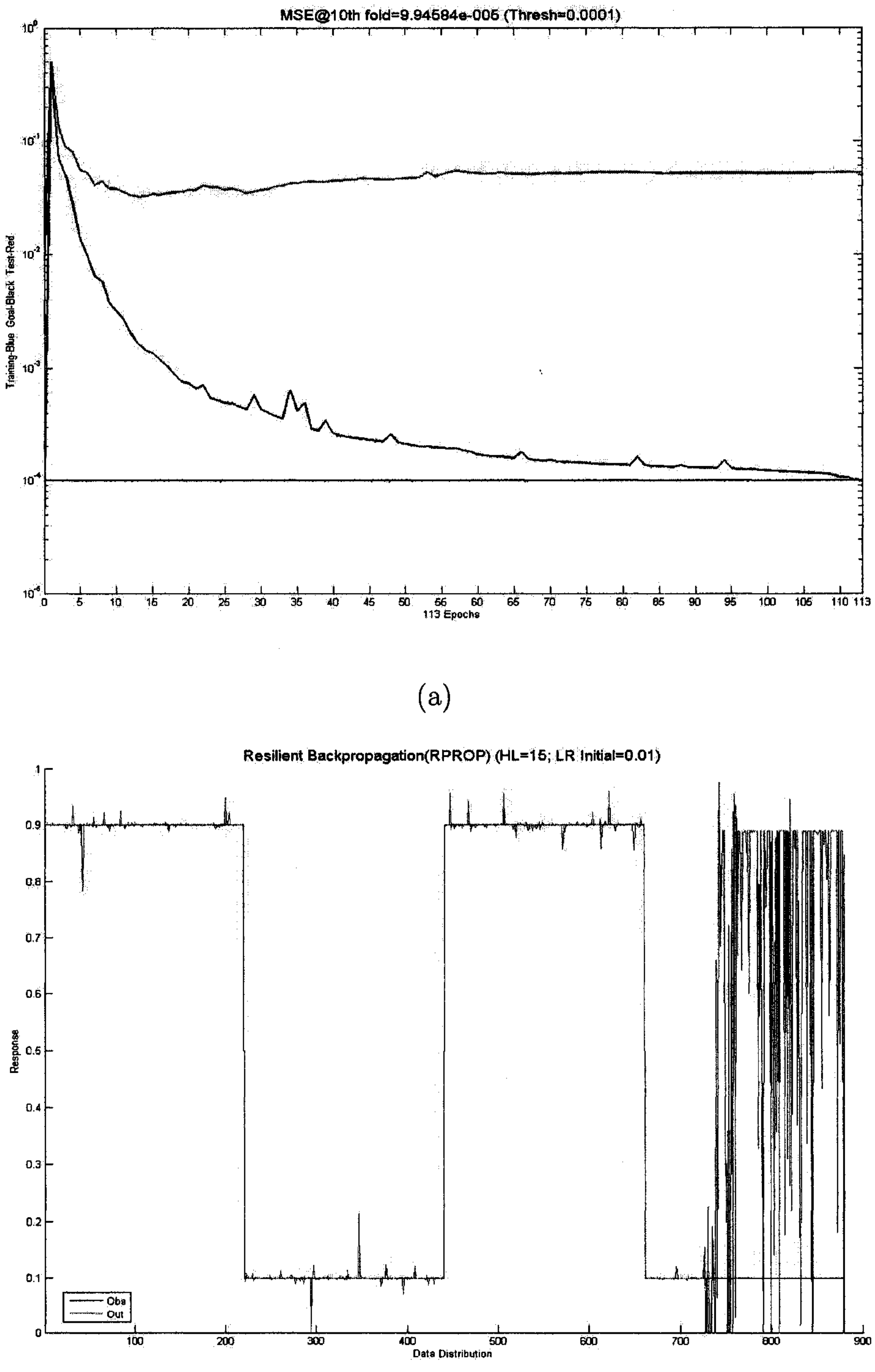
Figure 40: (a) [BPROP-GD-M] WF=29,HU=10 10 (b) Model Response



(a)

(b)

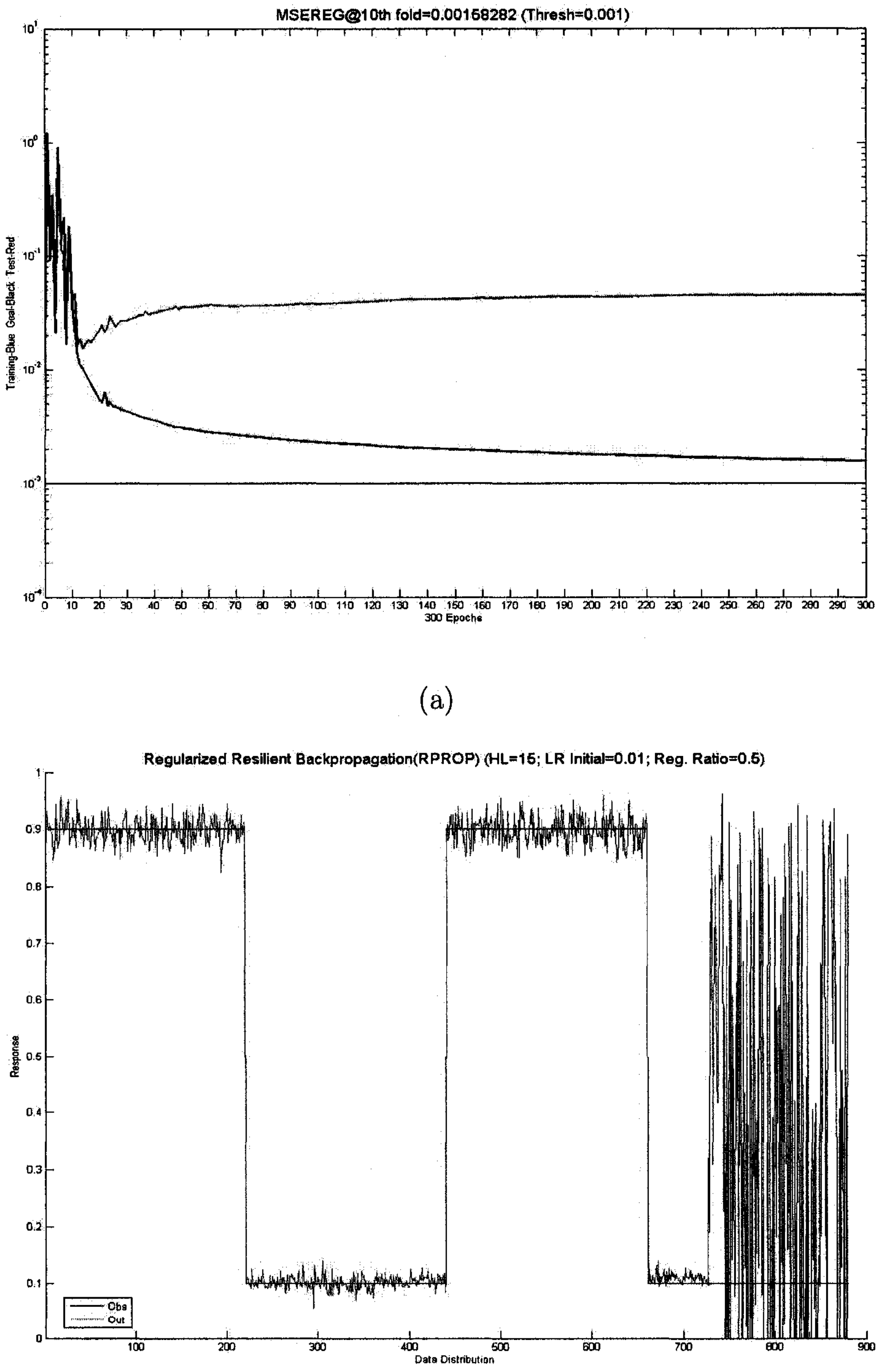
Figure 41: (a) [BPROP-GD-M] WF=29,HU=10 10 (b) Model Response



(a)

(b)

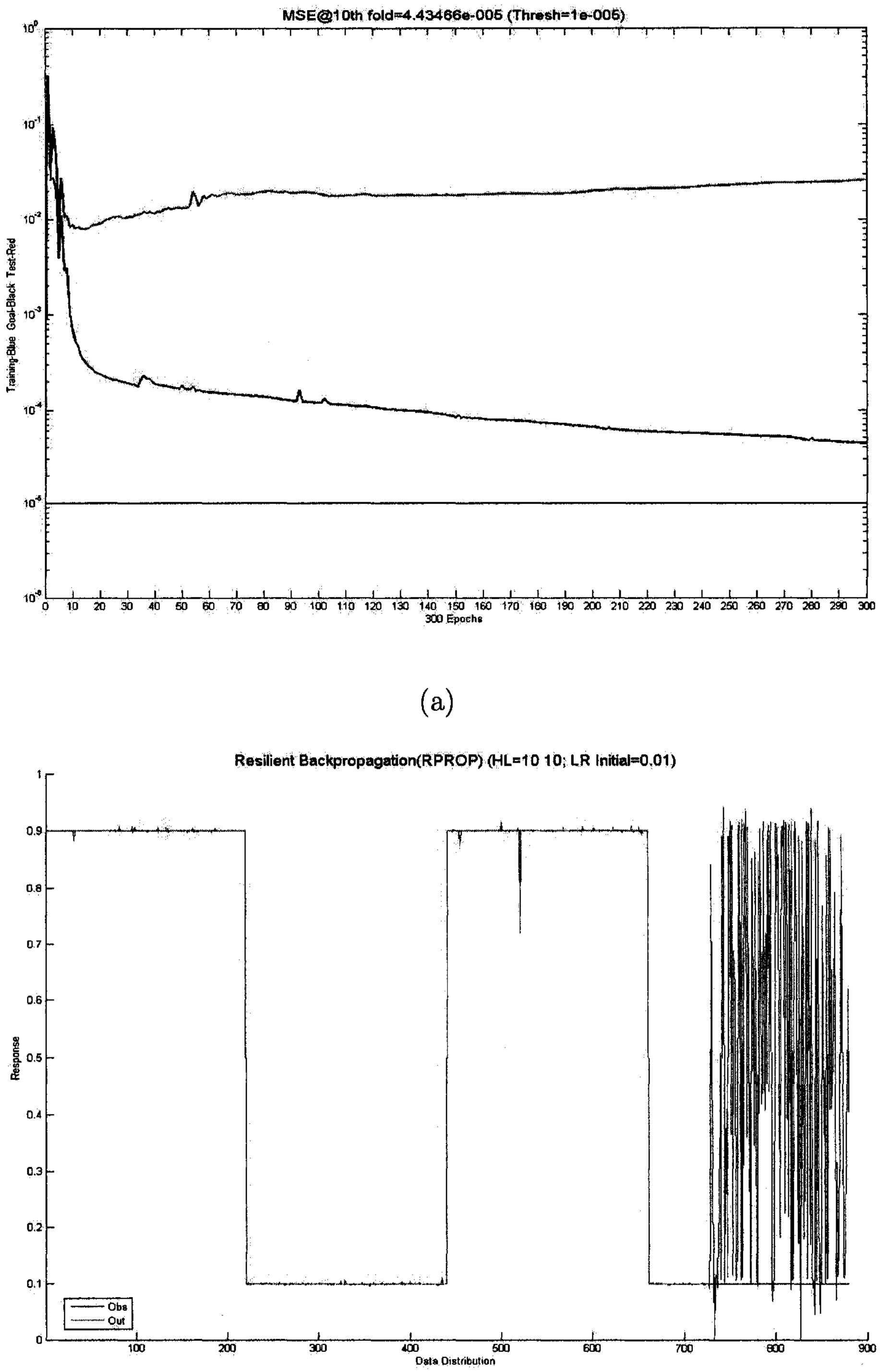
Figure 42: (a) [BPROP-RPROP] WF=29,HU=15 (b) Model Response



(a)

(b)

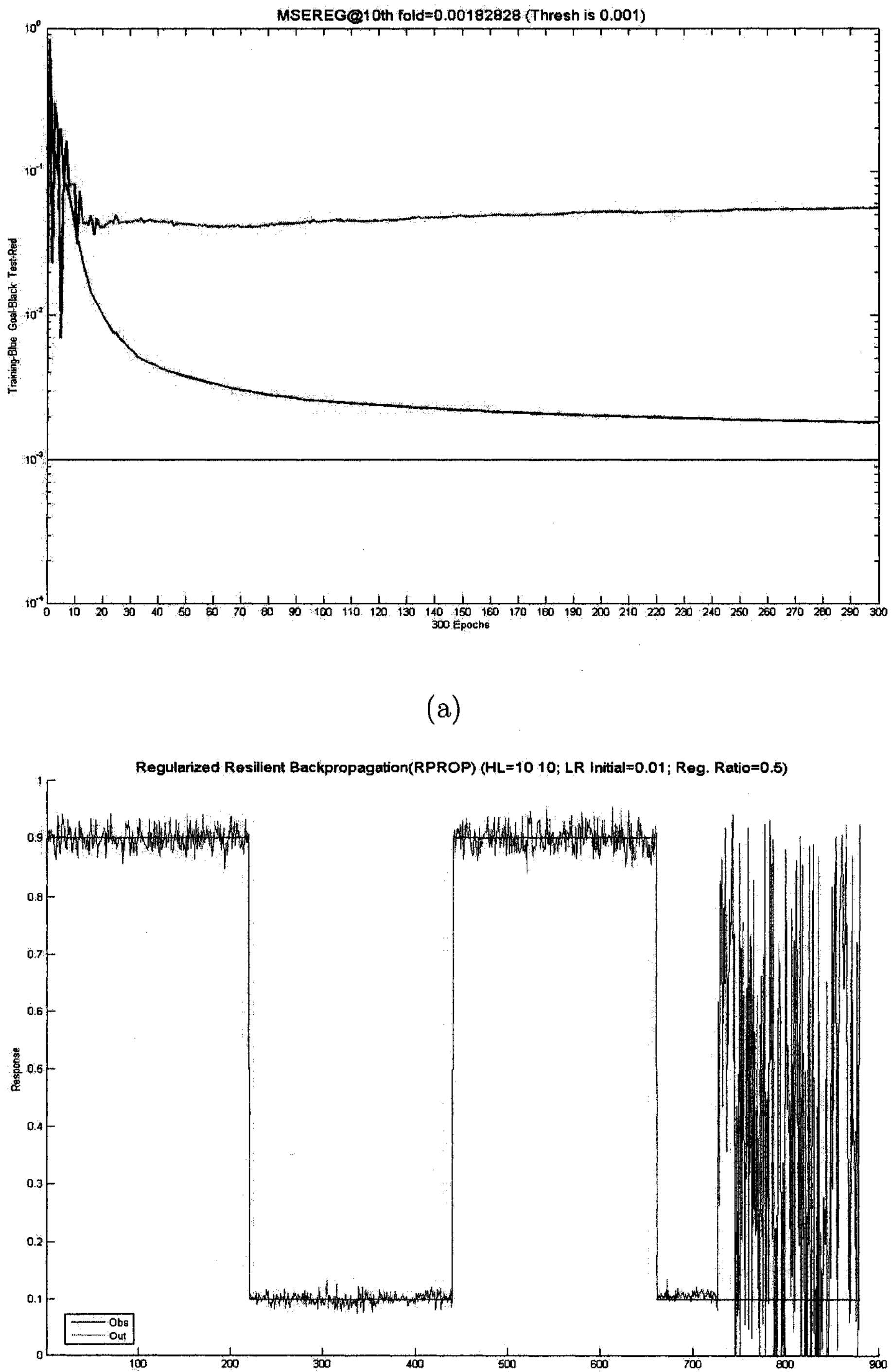
Figure 43: (a) [BPROP-RPROP] WF=29,HU=15,RR=0.5 (b) Model Response



(a)

(b)

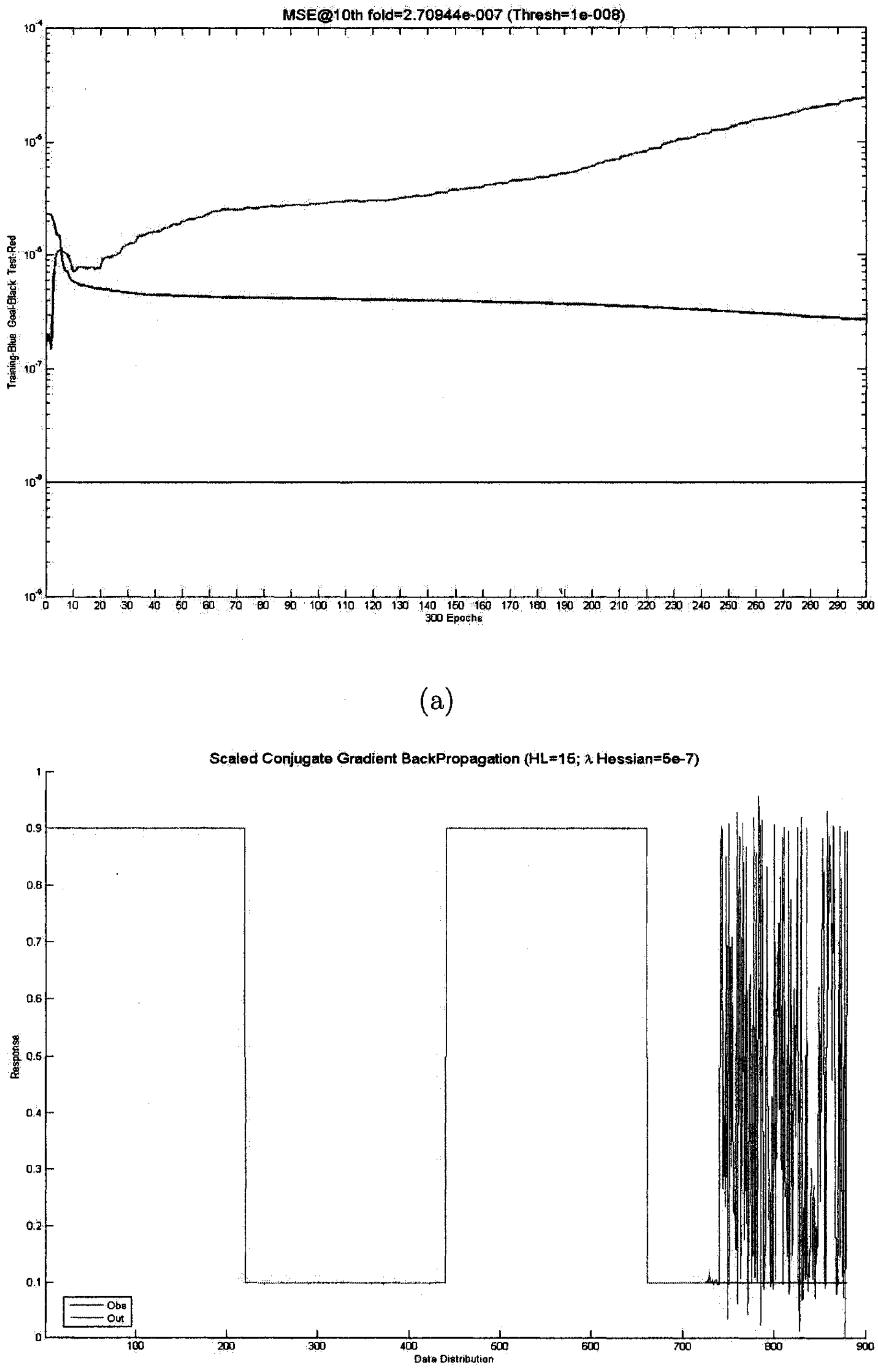
Figure 44: (a) [BPROP-RPROP] WF=29,HU=10 10 (b) Model Response



(a)

(b)

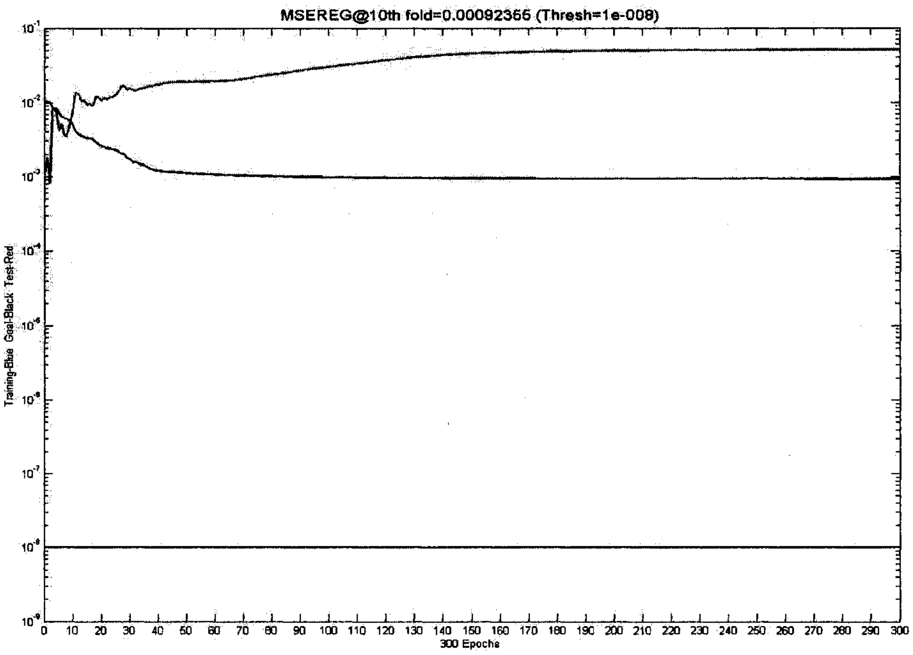
Figure 45: (a) [BPROP-RPROP] WF=29,HU=10 10, RR=0.5 (b) Model Response



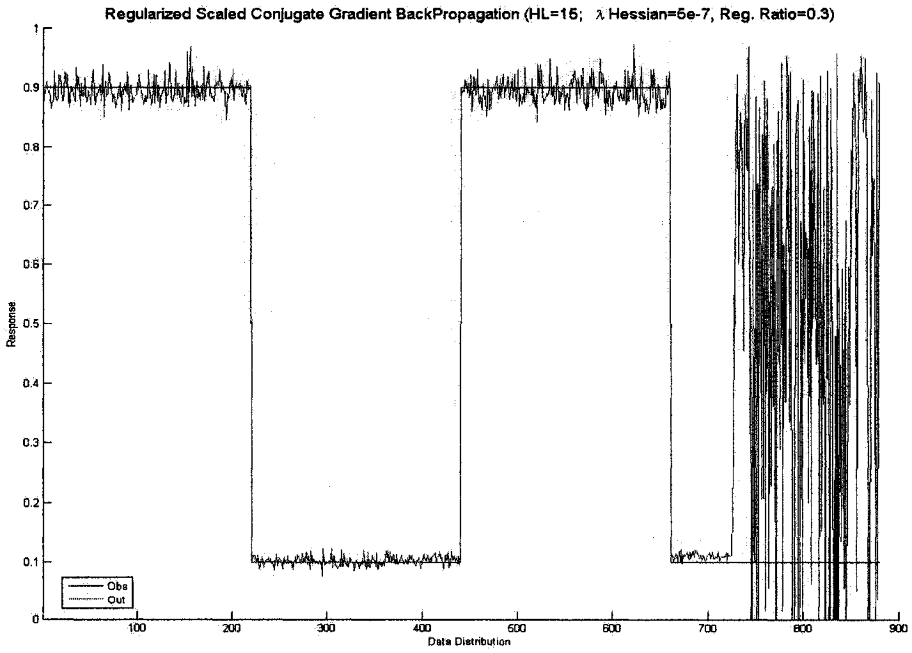
(a)

(b)

Figure 46: (a) [BPROP-SCG] WF=29,HU=15 (b) Model Response

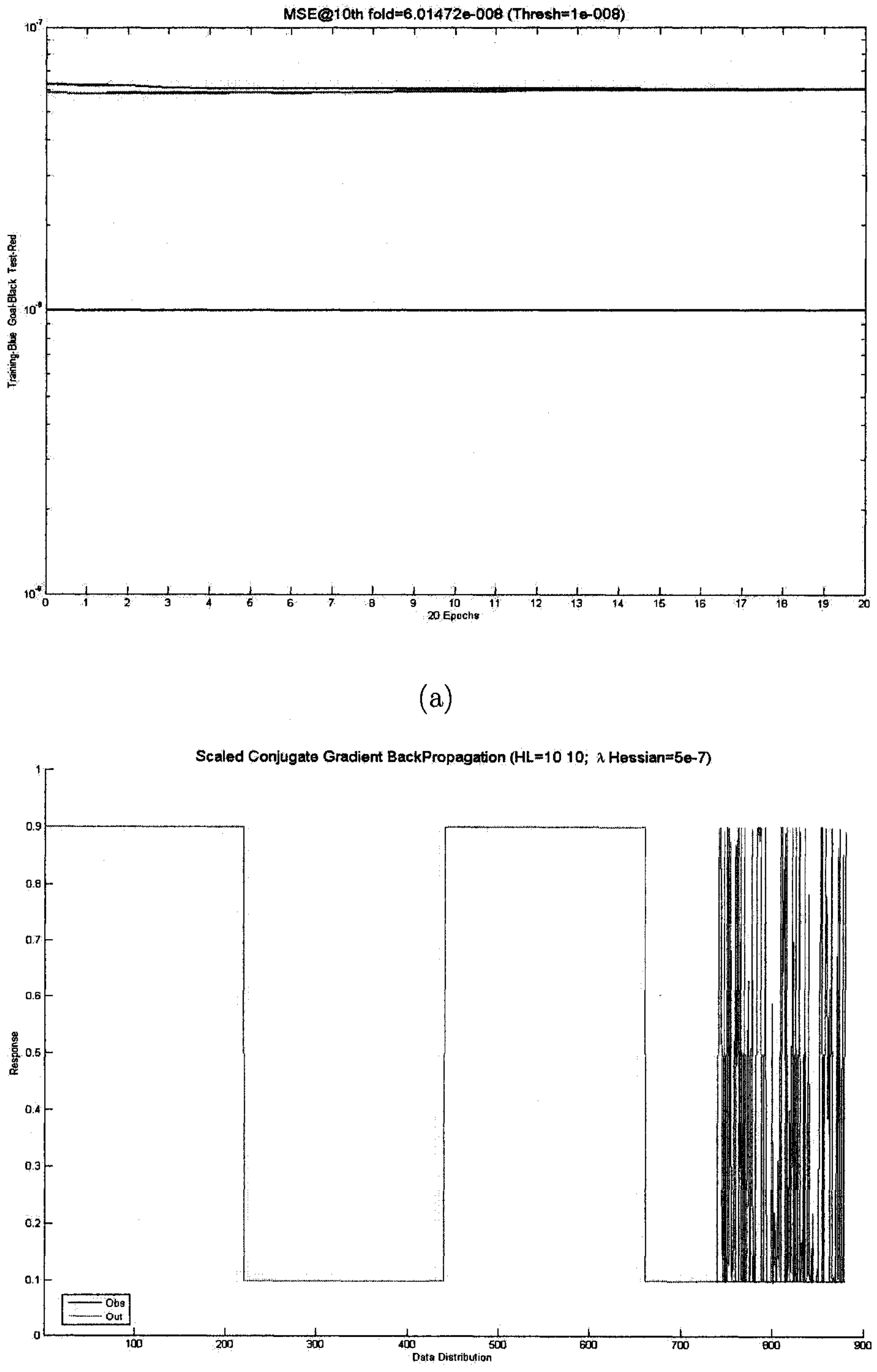


(a)



(b)

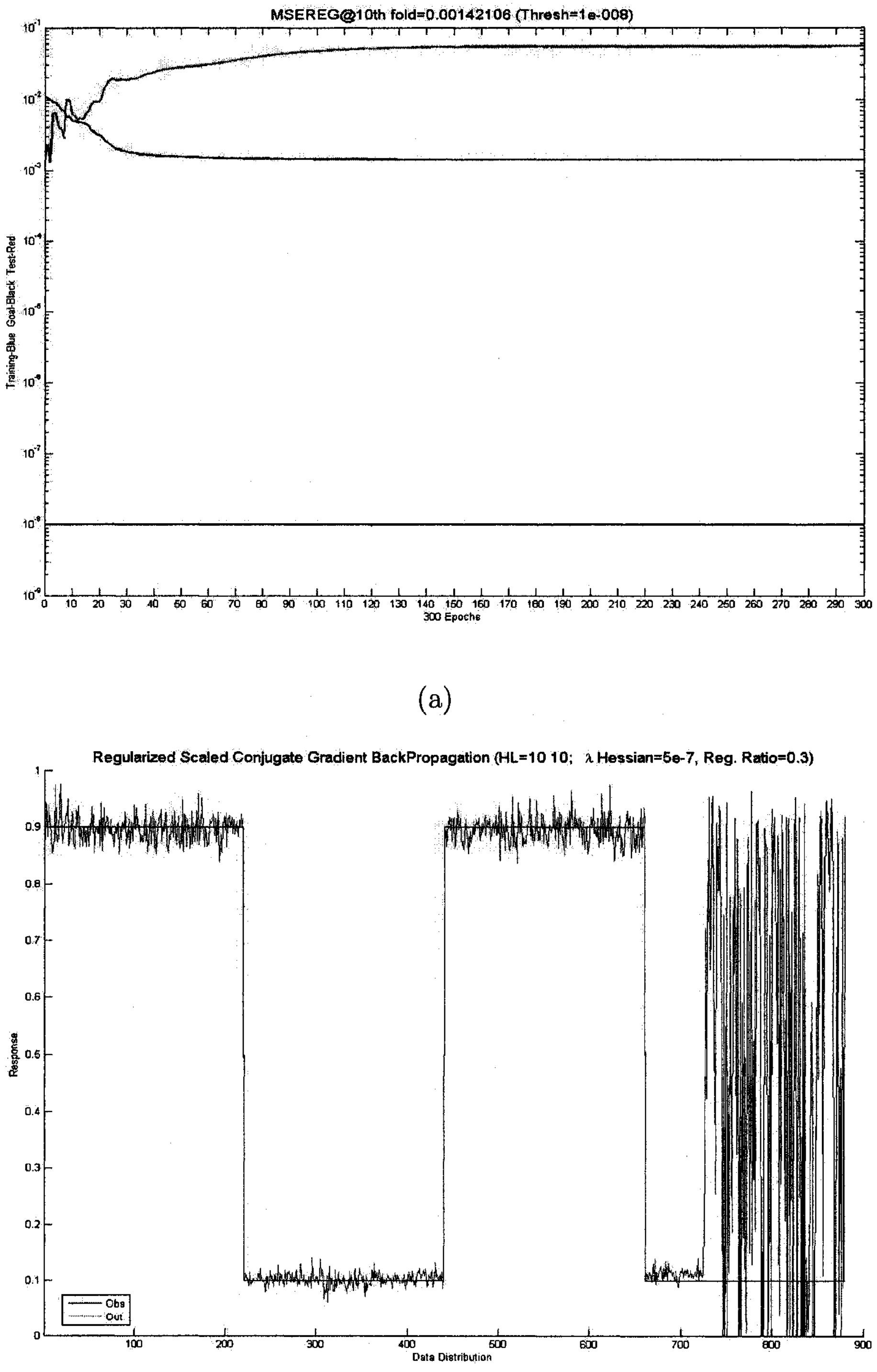
Figure 47: (a) [BPROP-SCG] WF=29,HU=15,RR=0.3 (b) Model Response



(a)

(b)

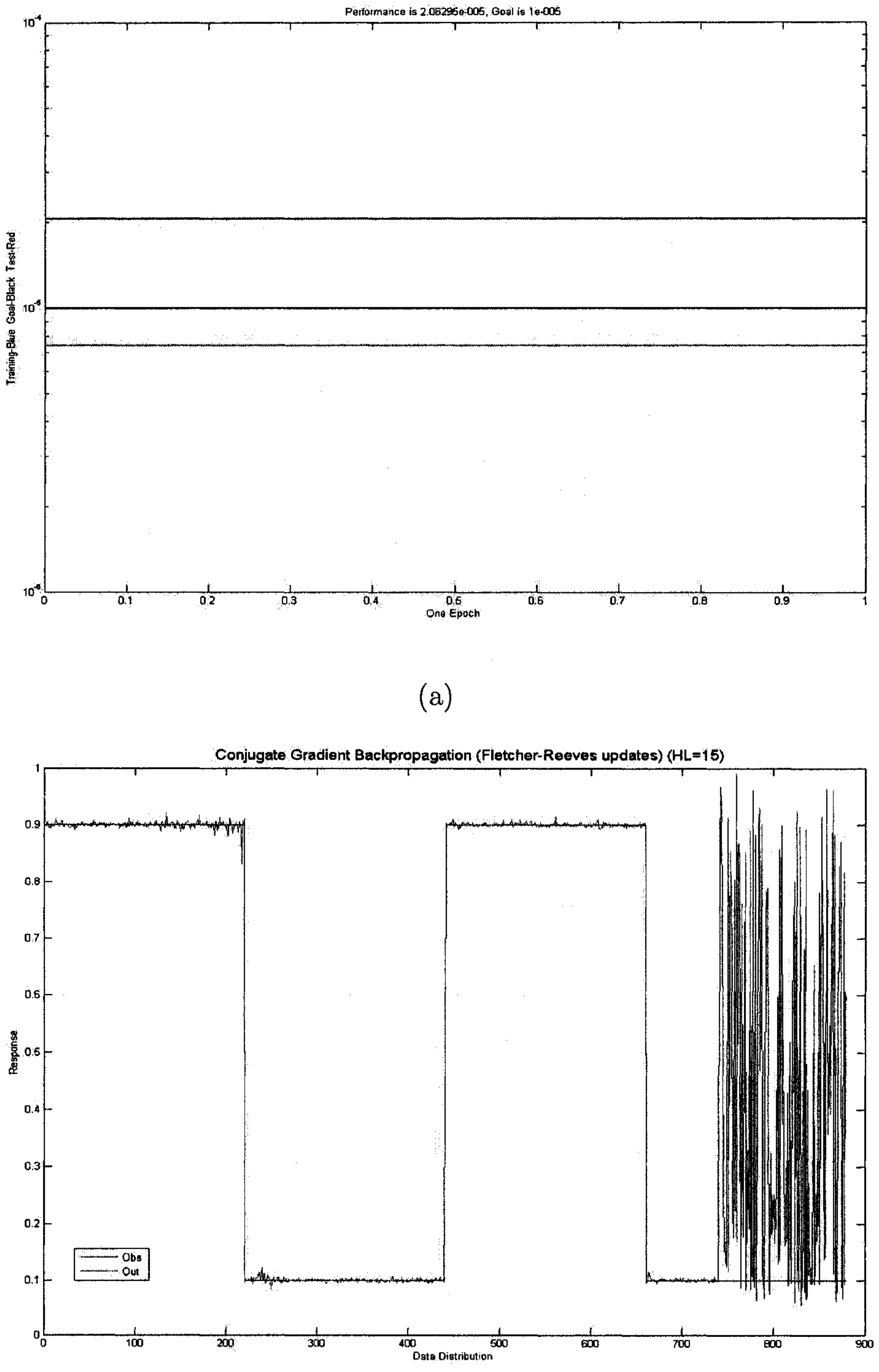
Figure 48: (a) [BPROP-SCG] WF=29,HU=10 10 (b) Model Response



(a)

(b)

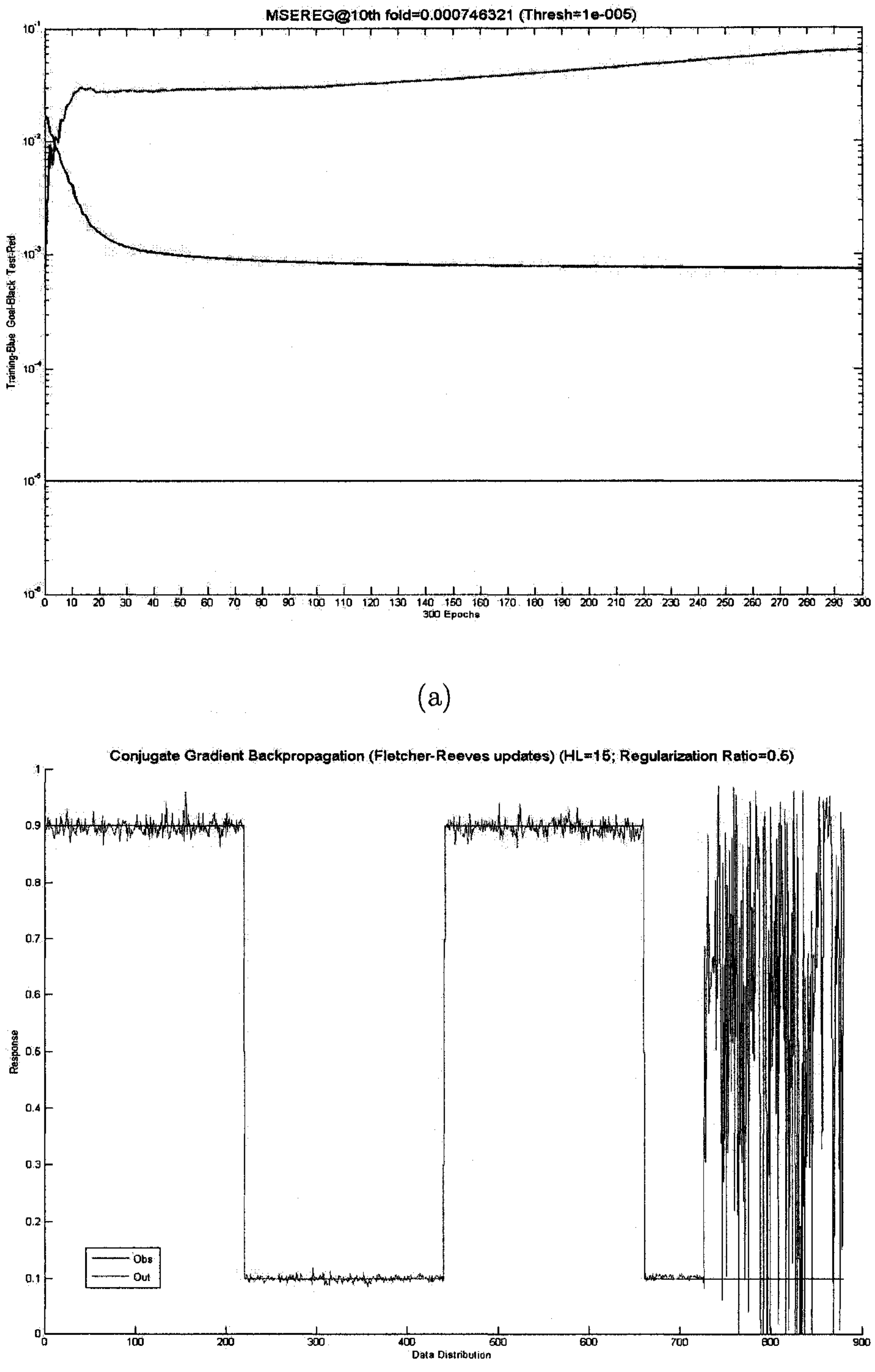
Figure 49: (a) [BPROP-SCG] WF=29,HU=10 10,RR=0.3 (b) Model Response



(a)

(b)

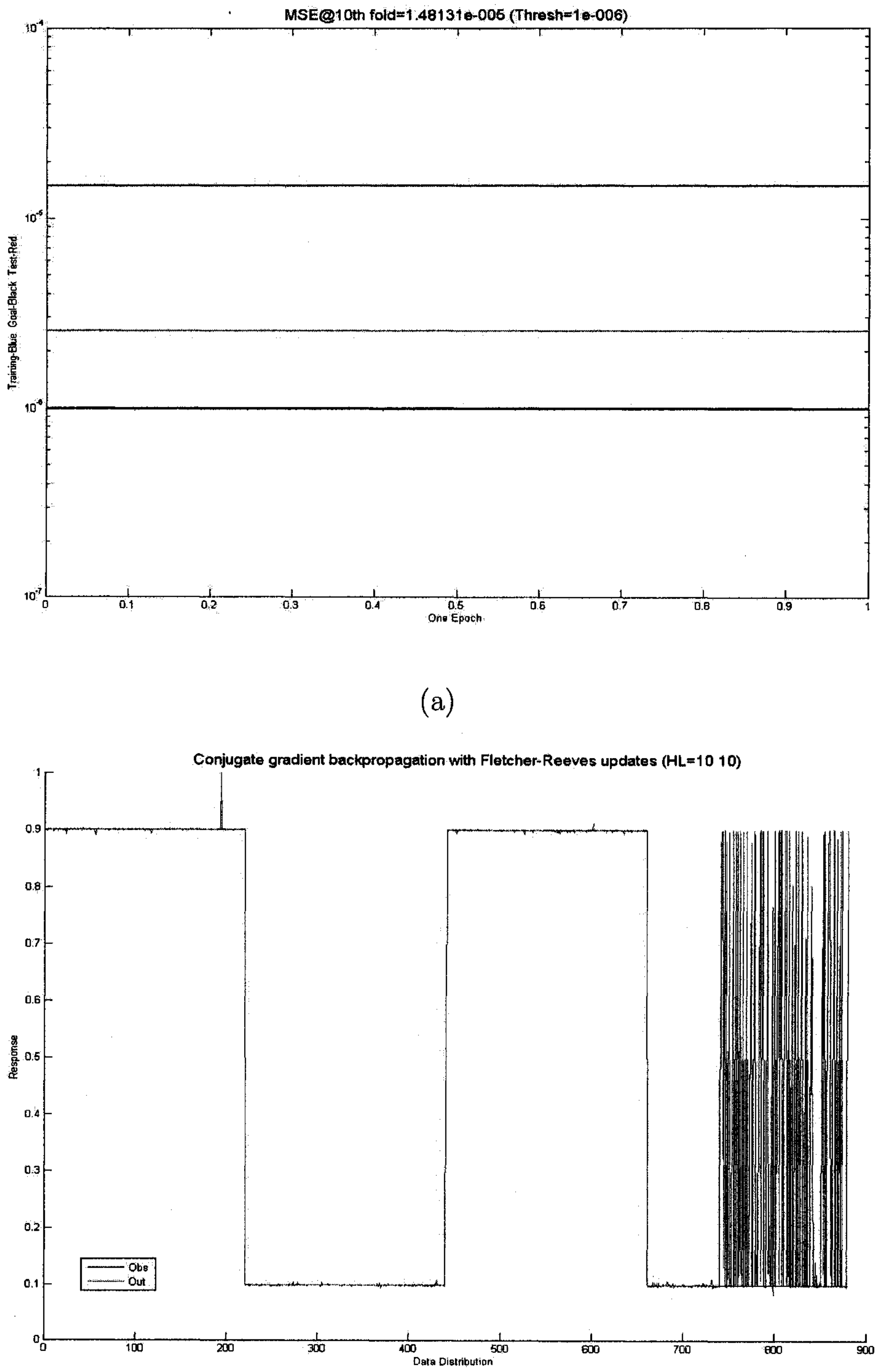
Figure 50: (a) [BPROP-CGF] WF=29,HU=15 (b) Model Response



(a)

(b)

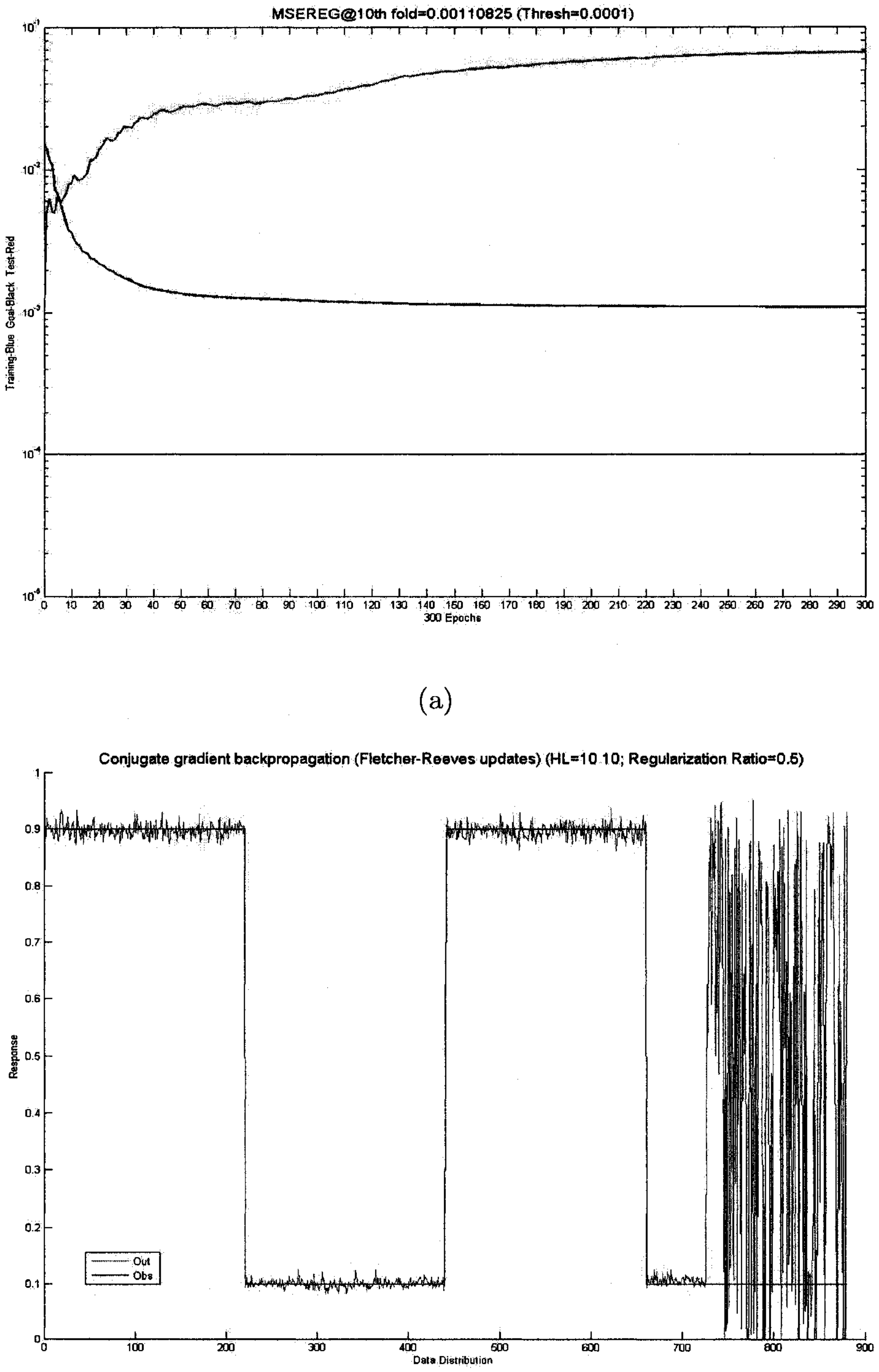
Figure 51: (a) [BPROP-CGF] WF=29,HU=15,RR=0.5 (b) Model Response



(a)

(b)

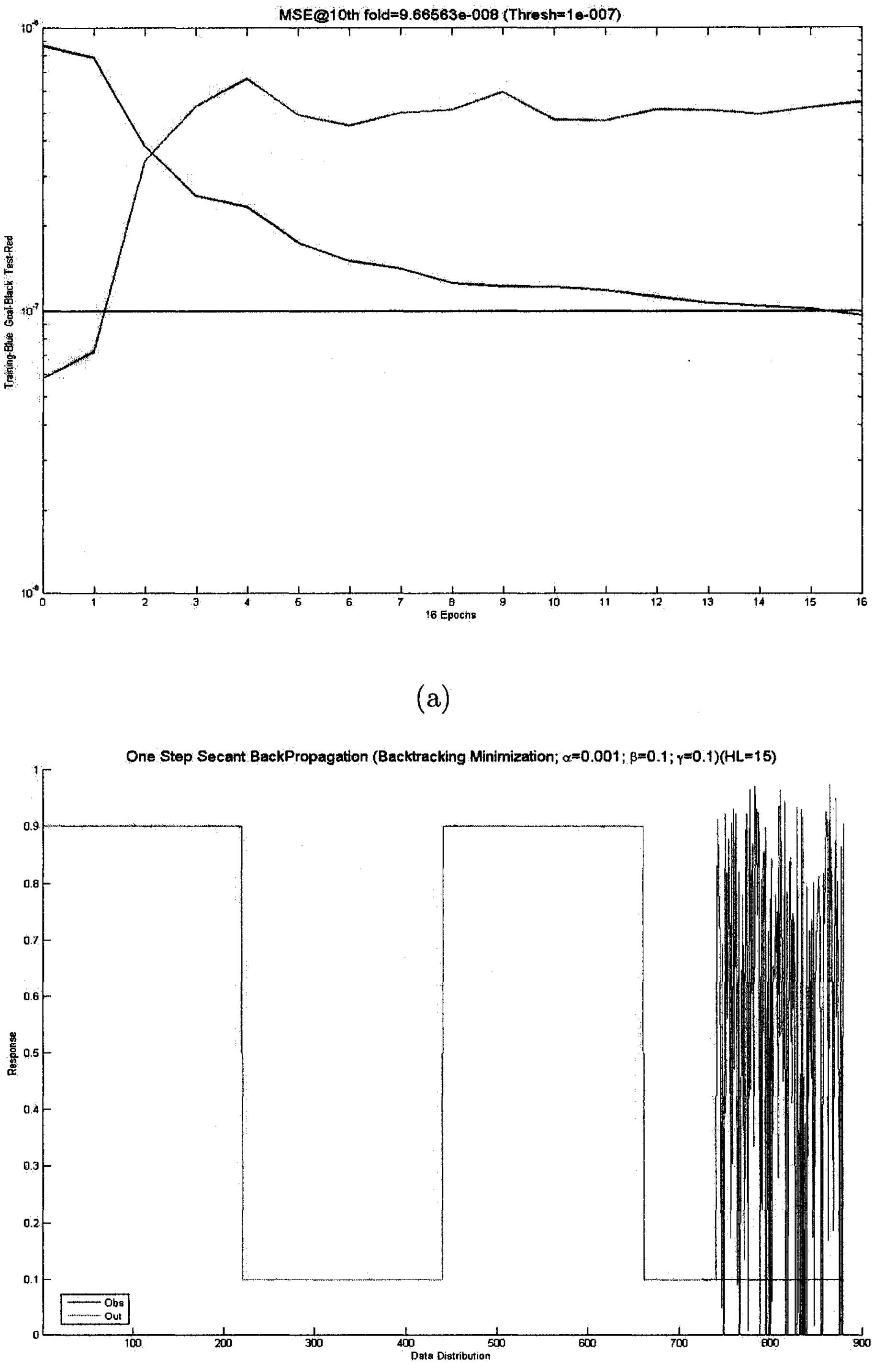
Figure 52: (a) [BPROP-CGF] WF=29,HU=10 10 (b) Model Response



(a)

(b)

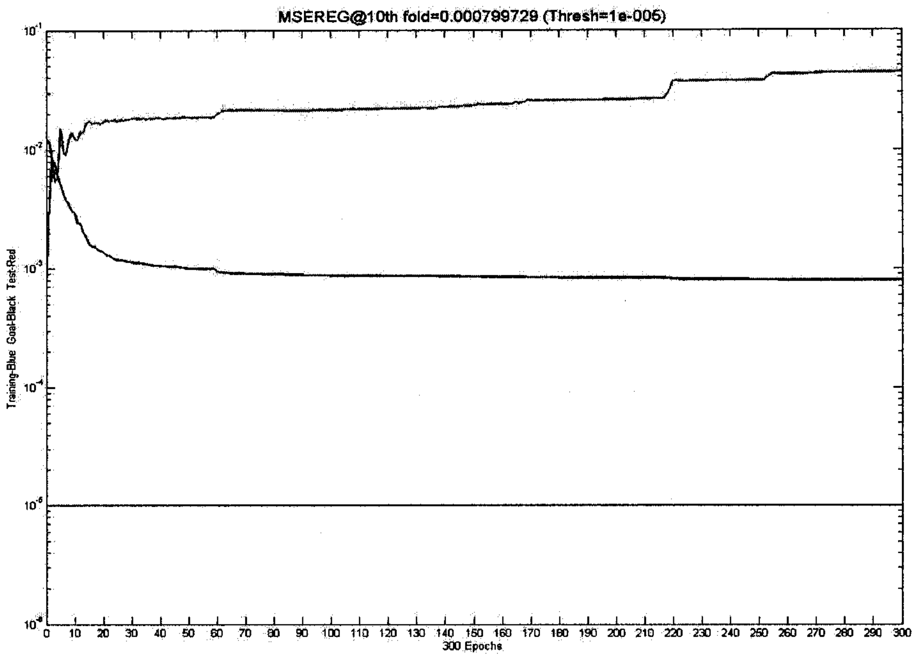
Figure 53: (a) [BPROP-CGF] WF=29,HU=10 10,RR=0.5 (b) Model Response



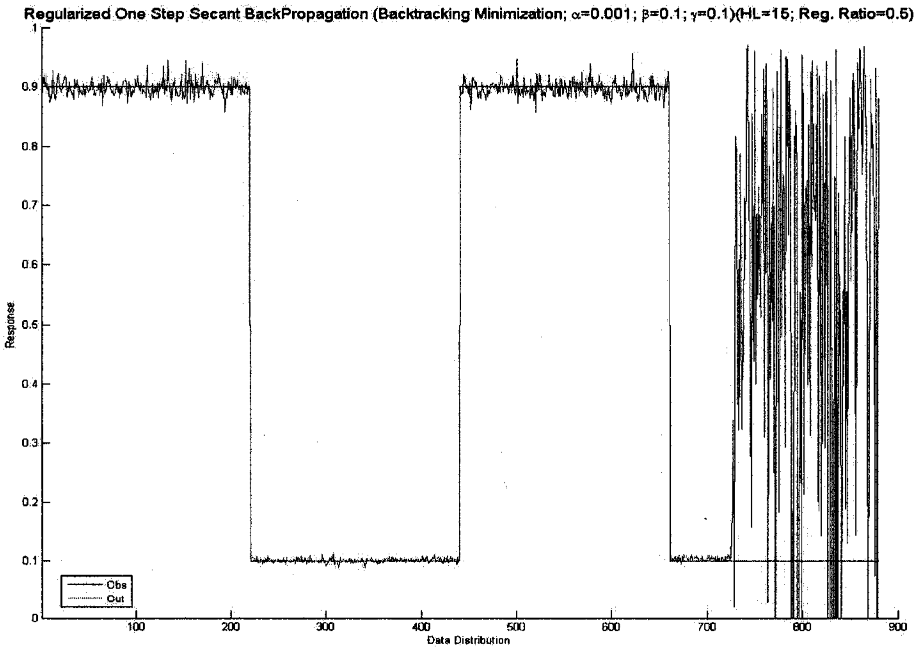
(a)

(b)

Figure 54: (a) [BPROP-OSS] WF=29,HU=15 (b) Model Response

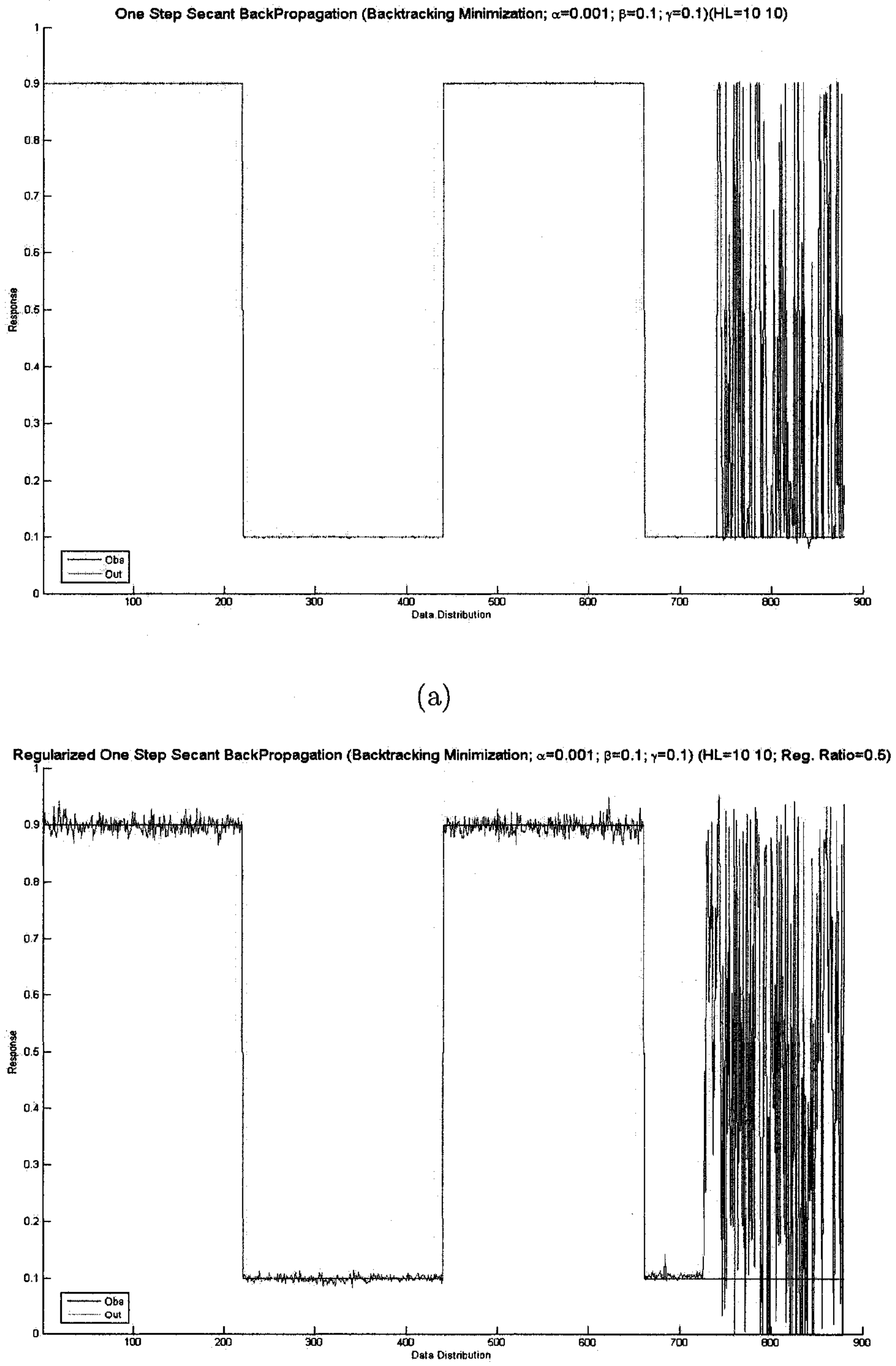


(a)



(b)

Figure 55: (a) [BPROP-OSS] WF=29,HU=15,RR=0.5 (b) Model Response



(a)

(b)

Figure 56: (a) [BPROP-OSS] WF=29,HU=15 (b) RR=0.5

Appendix B

Dataset A

No.	Accession	No.	Accession	No.	Accession	No.	Accession
1	Q9R158	34	Q923W9	67	Q923W9	100	Q9HCU4
2	O75078	35	P70505	68	P70505	101	Q9R0M0
3	Q9R1V4	36	O77780	69	O77780	102	Q9QYP2
4	Q9PSZ3	37	Q60411	70	Q60411	103	Q9NYQ7
5	O43184	38	Q99965	71	Q99965	104	Q91ZI0
6	Q61824	39	Q28478	72	Q28478	105	O88278
7	Q13444	40	Q60718	73	Q60718	106	Q9I8Q3
8	O88839	41	Q28660	74	Q28660	107	Q9GZR3
9	Q9QYV0	42	Q63202	75	Q63202	108	P97766
10	Q9Y3Q7	43	P78325	76	P78325	109	Q86T13
11	Q95194	44	Q05910	77	Q05910	110	Q8VCP9
12	Q9R157	45	Q13443	78	Q13443	111	P78357
13	P97776	46	Q61072	79	Q61072	112	O54991
14	Q9H013	47	Q60813	80	Q60813	113	P97846
15	O35674	48	Q8R534	81	Q8R534	114	Q9UHC6
16	O43506	49	P25371	82	P25371	115	Q9CPW0
17	Q9UKJ8	50	P31696	83	P31696	116	Q5RD64
18	Q9JI76	51	Q90404	84	Q90404	117	Q9BZ76
19	Q9P0K1	52	O00468	85	O00468	118	Q9C0A0
20	Q9R1V6	53	P25304	86	P25304	119	Q99P47
21	O42596	54	Q01594	87	Q01594	120	P13671
22	O75077	55	Q41233	88	Q41233	121	P61134
23	Q9R1V7	56	P31756	89	P31756	122	P61135
24	Q9R160	57	P31757	90	P31757	123	Q811M5
25	Q9R159	58	Q6UW56	91	Q6UW56	124	P10643
26	Q9UKQ2	59	Q6PGD0	92	Q6PGD0	125	Q9TUQ3
27	Q9XSL6	60	P41990	93	P41990	126	Q5RAD0
28	Q9JLN6	61	P15514	94	P15514	127	P07357
29	Q9UKF5	62	P31955	95	P31955	128	Q8K182
30	Q9UKF2	63	P24338	96	P24338	129	P98136
31	Q8TC27	64	O14525	97	O14525	130	P07358
32	Q8K410	65	Q61137	98	Q61137	131	Q8BH35
33	Q9BZ11	66	O75882	99	O75882	132	Q90X85

No.	Accession	No.	Accession	No.	Accession	No.	Accession
133	Q9PVW7	183	Q61483	233	P00743	283	P98133
134	P98137	184	P97677	234	P25155	284	P35555
135	P55314	185	Q9NYJ7	235	P00742	285	Q61554
136	Q3MHN2	186	O88516	236	O88947	286	Q9TV36
137	P79755	187	O88671	237	O19045	287	P35556
138	P48770	188	Q9NR61	238	Q63207	288	Q61555
139	P02748	189	Q9JI71	239	Q4QXT9	289	Q75N90
140	P06683	190	Q6DI48	240	P98140	290	P10079
141	P06682	191	O57409	241	Q04962	291	P15216
142	P48747	192	Q9IAT6	242	P00748	292	P49013
143	Q62930	193	Q8UWJ4	243	O97507	293	Q25464
144	P49747	194	P10041	244	P22457	294	Q14393
145	Q9R0G6	195	O43854	245	P08709	295	Q61592
146	P35444	196	O35474	246	P70375	296	Q63772
147	Q9NQ79	197	Q9UHF1	247	Q2F9P4	297	Q76CA1
148	Q8R555	198	Q9QXT5	248	Q2F9P2	298	P13508
149	P10040	199	Q6AZ60	249	P98139	299	P25291
150	Q5EA46	200	Q99944	250	Q8K3U6	300	P55259
151	Q96HD1	201	Q6GUQ1	251	P00741	301	Q9D733
152	Q91XD7	202	Q6MG84	252	P19540	302	P19218
153	Q4V7F2	203	Q6UY11	253	Q804X6	303	Q8V307
154	P82279	204	Q8K1E3	254	Q6SA95	304	Q776B5
155	Q8VHS2	205	P00533	255	P00740	305	P08072
156	Q5IJ48	206	Q9BEA0	256	P16294	306	Q6RZT5
157	Q80YA8	207	Q95ND4	257	Q95ND7	307	P08441
158	Q8CG14	208	P01133	258	P16293	308	O57166
159	Q8CFG8	209	P01132	259	Q9VW71	309	P20494
160	P81282	210	Q00968	260	Q9NYQ8	310	Q86607
161	Q90953	211	P07522	261	O88277	311	Q9JFH4
162	P13611	212	P15217	262	Q14517	312	P01136
163	Q28858	213	Q99372	263	P33450	313	P33804
164	Q62059	214	Q9HBW9	264	O42182	314	P42287
165	Q9ERB4	215	Q923X1	265	O77469	315	Q5E9Z2
166	O14594	216	Q9ESC1	266	Q8MJJ9	316	Q14520
167	P55066	217	Q14246	267	O73775	317	Q8K0D2
168	Q5IS41	218	Q61549	268	P23142	318	Q6L711
169	P55067	219	Q9UHX3	269	Q08879	319	Q09118
170	Q9DF69	220	Q9BY15	270	P98095	320	Q99075
171	O95196	221	Q86SQ3	271	P37889	321	Q06186
172	Q71M36	222	Q91ZE5	272	Q12805	322	Q01580
173	Q9ERQ6	223	Q5EG71	273	Q7YQD7	323	Q06175
174	Q20911	224	Q6UW88	274	Q8BPB5	324	Q6QNF4
175	Q9TU53	225	Q924X1	275	O35568	325	Q04756
176	O60494	226	Q15303	276	O55058	326	Q9R098
177	Q9JLB4	227	Q61527	277	O95967	327	Q96QV1
178	O70244	228	Q62956	278	Q9WVJ9	328	Q7TN16
179	Q09165	229	O14944	279	Q5EA62	329	Q96RW7
180	P80370	230	Q61521	280	Q9UBX5	330	Q5E985
181	Q09163	231	P83370	281	Q9WVH9	331	Q12794
182	O00548	232	P81428	282	Q9WVH8	332	Q91ZJ9

No.	Accession	No.	Accession	No.	Accession	No.	Accession
333	Q6RHW4	383	P17801	433	Q924T4	483	P16109
334	Q76HN1	384	P45442	434	P79948	484	Q01102
335	Q12891	385	P25391	435	P14585	485	P98106
336	O35632	386	P19137	436	Q03345	486	P98109
337	Q9Z2Q3	387	P24043	437	Q9NZR2	487	P48740
338	Q8SQG7	388	Q60675	438	Q9JI18	488	P98064
339	O43820	389	Q16787	439	P98157	489	O00187
340	Q8VEI3	390	Q61789	440	Q07954	490	Q91WP0
341	Q6RHW2	391	Q16363	441	P98164	491	Q9JJS8
342	Q5REQ1	392	P97927	442	P98158	492	P05099
343	Q76HM9	393	O15230	443	O75096	493	P21941
344	P12606	394	Q61001	444	Q8VI56	494	P51942
345	P12607	395	Q00174	445	Q9QYP1	495	O00339
346	P53712	396	Q01635	446	O75197	496	O08746
347	P07228	397	P11046	447	Q91VN0	497	O42401
348	P53713	398	P07942	448	O75581	498	O15232
349	P05556	399	Q27262	449	O88572	499	O35701
350	P09055	400	P02469	450	Q98931	500	O95460
351	Q9GLP0	401	P55268	451	Q14114	501	O89029
352	Q5RCA9	402	Q61292	452	Q924X6	502	O75095
353	P49134	403	P15800	453	Q04833	503	Q80V70
354	P32592	404	Q13751	454	Q14766	504	O88281
355	P05107	405	Q61087	455	Q8CG19	505	Q7Z7M0
356	P11835	406	Q01636	456	P22064	506	P60882
357	P53714	407	Q25092	457	Q8CG18	507	Q9QYP0
358	P05106	408	P15215	458	Q00918	508	Q9H1U4
359	O54890	409	P11047	459	Q28019	509	Q8BH27
360	P16144	410	P02468	460	Q14767	510	Q16819
361	Q64632	411	Q8HZI9	461	O08999	511	P28825
362	P18084	412	Q13753	462	O35806	512	Q64230
363	O70309	413	Q61092	463	Q9NS15	513	Q16820
364	Q07441	414	Q9Y6N6	464	Q61810	514	Q61847
365	Q8SQB8	415	Q9R0B6	465	Q8K4G1	515	P28826
366	P18563	416	Q18823	466	P98131	516	Q95114
367	P18564	417	Q21313	467	P14151	517	Q08431
368	Q9Z0T9	418	Q99087	468	Q95198	518	P21956
369	P26010	419	Q99088	469	P18337	519	P79385
370	P26011	420	P01131	470	Q95237	520	P70490
371	P26012	421	P35950	471	Q28768	521	Q13201
372	P26013	422	P01130	472	Q95235	522	P19598
373	Q27591	423	P35951	473	P30836	523	P04934
374	P11584	424	Q28832	474	P98107	524	P13819
375	Q90Y57	425	P20063	475	P33730	525	P04932
376	Q90Y54	426	P35952	476	Q95LG1	526	P08569
377	P78504	427	Q26422	477	P16581	527	P50495
378	Q9QXX0	428	P28175	478	Q00690	528	P04933
379	Q63722	429	Q8JHF2	479	P98110	529	P34576
380	Q9Y219	430	O12971	480	P27113	530	Q9H3R2
381	Q9QYE5	431	Q8NES3	481	P98105	531	P19467
382	P97607	432	O09010	482	P42201	532	P97881

No.	Accession	No.	Accession	No.	Accession	No.	Accession
533	Q8BJ48	583	Q01705	633	P14650	683	Q9YHB3
534	Q21180	584	Q07008	634	P98160	684	Q9R1U9
535	Q21178	585	Q04721	635	Q05793	685	P79949
536	Q21179	586	O35516	636	P13608	686	P18168
537	Q21181	587	Q9QW30	637	Q28343	687	O75093
538	Q22396	588	Q9UM47	638	P07898	688	Q80TR4
539	Q22398	589	Q61982	639	P16112	689	O88279
540	Q7Z0M7	590	Q9R172	640	Q61282	690	O94813
541	Q93542	591	Q99466	641	Q28062	691	Q9R1B9
542	Q20459	592	P31695	642	Q96GW7	692	O750945
543	Q22710	593	P07207	643	Q61361	693	Q9WVB4
544	O17264	594	P21783	644	P55068	694	O88280
545	P98061	595	Q05199	645	P23219	695	P24014
546	Q20958	596	Q02297	646	P22437	696	Q15491
547	Q9N2V2	597	P43322	647	O97554	697	Q98930
548	Q7JLI1	598	O93383	648	Q63921	698	Q92673
549	O16977	599	O14511	649	P05979	699	O88307
550	P55114	600	P56974	650	O62698	700	Q95209
551	Q21059	601	O35569	651	P70682	701	Q07929
552	P98060	602	P56975	652	P27607	702	P98068
553	Q61EX6	603	O35181	653	O19183	703	Q01083
554	Q18206	604	Q8WWG1	654	P35354	704	Q96GP6
555	Q93243	605	Q9WTX4	655	Q05769	705	P59222
556	Q20942	606	Q28146	656	O62725	706	Q14162
557	Q20176	607	Q9DDD0	657	O02768	707	P98167
558	Q92832	608	Q9ULB1	658	P35355	708	Q2PC93
559	Q62919	609	Q9CS84	659	P79208	709	Q8CG65
560	Q99435	610	Q63372	660	P00745	710	Q700K0
561	Q61220	611	Q9P2S2	661	Q28278	711	Q9NY15
562	Q62918	612	Q63374	662	P04070	712	Q8R4Y4
563	Q90827	613	Q9Y4C0	663	P33587	713	Q8WWQ8
564	Q90922	614	Q07310	664	Q9GLP2	714	Q8R4U0
565	O95631	615	Q94887	665	Q28661	715	Q8CFM6
566	O09118	616	Q9Y2I2	666	P31394	716	Q9V5N8
567	Q90923	617	Q8R4G0	667	P07224	717	Q26627
568	O00634	618	Q96CW9	668	P07225	718	P13385
569	Q9R1A3	619	Q8R4F1	669	Q28520	719	P51865
570	Q9HB63	620	Q04620	670	Q08761	720	P51864
571	Q9JI33	621	P13829	671	P98118	721	O75443
572	Q5RB89	622	P13401	672	P53813	722	Q9UIK5
573	Q24567	623	P19455	673	P00744	723	Q9QYM9
574	Q24568	624	Q05439	674	P22891	724	P10039
575	P14543	625	Q27874	675	Q9CQW3	725	P24821
576	P10493	626	P14222	676	O93574	726	Q80YX1
577	P08460	627	P10820	677	P78509	727	Q29116
578	Q14112	628	P35763	678	Q60841	728	Q9UQP3
579	O88322	629	Q8HYB7	679	P58751	729	Q80Z71
580	P46530	630	P07202	680	O12972	730	Q00546
581	P46531	631	P35419	681	Q9Y644	731	Q92752
582	Q01705	632	P09933	682	O09009	732	Q8BYI9

No.	Accession	No.	Accession	No.	Accession	No.	Accession
733	Q05546	783	Q9Z1T2	833	Q6IN38	883	Q8N780
734	P22105	784	P49744	834	Q9W6F8	884	Q2FA44
735	P01135	785	Q06441	835	O44443	885	Q8NHD4
736	P55244	786	Q76DT2	836	Q19981	886	Q5XG84
737	P48030	787	P58911	837	P41950	887	Q8IXD0
738	Q06922	788	Q06561	838	P98163	888	Q1HK36
739	P98138	789	P34710	839	P34554	889	Q5GFL6
740	P01134	790	Q05589	840	Q5UR16	890	Q5TEW6
741	P98135	791	P15120	841	Q7T6Y2	891	Q70UZ8
742	Q06805	792	P00749	842	Q9Y493	892	Q4V305
743	P35590	793	P06869	843	O88799	893	Q5RGR9
744	Q06806	794	P16227	844	Q28983	894	Q6UY05
745	Q06807	795	P04185	845	P57999	895	Q71S64
746	O73791	796	P29598	846	Q2U1S0	896	Q71S65
747	Q02763	797	Q5DID0	847	Q4ICJ6	897	Q8IYR6
748	Q02858	798	Q5DID3	848	Q6C5U0	898	Q5VYQ7
749	P25723	799	P48733	849	Q2UP95	899	Q5VYQ4
750	O57460	800	Q862Z3	850	Q756R4	900	Q71S61
751	Q9DER7	801	P07911	851	Q5KEN3	901	Q6NSY8
752	O43897	802	Q91X17	852	Q6FKY0	902	Q5SQD3
753	Q62381	803	Q5R5C1	853	Q2UPN5	903	Q71S63
754	Q8JI28	804	P27590	854	Q5B3W3	904	Q71S69
755	Q9Y6L7	805	P98119	855	Q5B5P0	905	Q71S62
756	Q9WVM6	806	P15638	856	Q55NL5	906	Q5ST74
757	O57382	807	P98121	857	Q9NY77	907	Q71S67
758	Q9HCN3	808	O75445	858	Q59HD7	908	Q5VU27
759	Q9ESN3	809	Q2QI47	859	Q9P273	909	Q5VYE7
760	Q28198	810	Q8K3K1	860	Q59HF7	910	Q5SW66
761	P00750	811	Q9J524	861	Q86YS9	911	Q5VUP1
762	P11214	812	Q6EMK4	862	Q5VVG4	912	Q3MI86
763	P19637	813	Q9CZT5	863	Q5T669	913	Q71S66
764	P06579	814	Q6DF55	864	Q5RGS1	914	Q5TI48
765	Q5W7P8	815	Q94918	865	Q59FL3	915	Q9GZZ2
766	P07204	816	P98165	866	Q4UJ74	916	Q71S68
767	P15306	817	P98155	867	Q5T1H1	917	Q5VYH3
768	Q71U07	818	P98156	868	Q5RGU6	918	Q5T8V6
769	Q28178	819	P35953	869	Q59H36	919	Q5VW17
770	P07996	820	P98166	870	Q59GI8	920	Q5VYQ6
771	P35441	821	P93026	871	O75441	921	Q13086
772	P35448	822	P93484	872	Q69YJ3	922	Q5TI47
773	Q95116	823	O22925	873	Q6ZS56	923	Q5T7C8
774	P35440	824	O80977	874	Q75QY0	924	Q5VUM2
775	P35442	825	Q56ZQ3	875	Q9H557	925	Q5TI49
776	Q03350	826	O64758	876	Q9ULU2	926	Q53FU9
777	Q8JHW2	827	Q9FYH7	877	Q5SZI8	927	Q5SR68
778	Q1L8P7	828	Q8L7E3	878	Q5ICN7	928	O95938
779	P49746	829	Q9W6F9	879	Q16519	929	Q5TI50
780	Q05895	830	Q9W3W5	880	Q6P192	930	Q5VTD0
781	Q8JGW0	831	Q9Y5W5	881	Q5SZI7	931	Q5SXM3
782	P35443	832	Q9WUA1	882	Q96FY1	932	Q53XQ0

No.	Accession	No.	Accession	No.	Accession	No.	Accession
933	Q5T7C7	983	Q5H9P5	1033	Q6MZK8	1083	Q6PJ72
934	Q5JZ17	984	Q5HYM8	1034	Q6L9N4	1084	Q76E14
935	O95965	985	Q5JVF1	1035	Q96M80	1085	Q7LC53
936	Q5T938	986	Q5T2Y7	1036	Q5JTP4	1086	Q7Z387
937	Q96IB3	987	Q5T3T9	1037	Q4LE67	1087	Q96K89
938	Q6PJA5	988	Q5W9F7	1038	Q96MS7	1088	Q9NP01
939	Q9BS56	989	Q5W9G8	1039	Q53R09	1089	Q9UKW9
940	Q8WTR8	990	Q5Y190	1040	Q5CZB3	1090	Q9P2P4
941	Q7Z7K9	991	Q6P2G0	1041	Q9NSD0	1091	Q6VU69
942	Q5TI44	992	Q6P3V5	1042	Q2M1N2	1092	Q8NAL2
943	Q9BTL9	993	Q86TI6	1043	O14637	1093	Q8TEP7
944	Q5SZ10	994	Q8N8N5	1044	Q8N3T8	1094	Q6ZYK7
945	Q8WUM6	995	Q9UN94	1045	Q59GF2	1095	Q6PJ75
946	Q5TI43	996	Q6ZS39	1046	Q59G97	1096	Q96SQ3
947	Q5SW65	997	Q9UN95	1047	Q53TP7	1097	Q5EBL7
948	Q8WWY1	998	Q8NBV0	1048	Q53TA7	1098	Q9NTF1
949	Q5VW18	999	Q9UKZ4	1049	Q53RD9	1099	Q9NQ15
950	Q9UJ43	1000	Q8IZZ5	1050	Q4AC85	1100	Q2I7G5
951	Q5SPL1	1001	Q8N1E9	1051	Q2NL86	1101	Q6UXI9
952	Q8TBU7	1002	Q5TI75	1052	Q53H93	1102	Q6NVV9
953	Q9NY09	1003	Q5IEC1	1053	Q4W0V0	1103	Q5VZK1
954	Q5SW67	1004	Q59FG2	1054	Q5JRP1	1104	Q53HU9
955	Q5VUP0	1005	Q59EG0	1055	O00508	1105	Q4VB91
956	Q5VTE4	1006	Q4KMR2	1056	Q96MJ5	1106	Q9UKM2
957	Q5VY43	1007	Q9NPR0	1057	Q86YZ7	1107	Q59FG9
958	Q5TI45	1008	Q6ICV5	1058	Q8TAS6	1108	Q9H7M4
959	Q5SR69	1009	Q96AA0	1059	Q6ZSN4	1109	Q86TV4
960	O43686	1010	Q9H3Q7	1060	Q6QBS1	1110	Q6VU68
961	Q5VV63	1011	Q9H481	1061	Q6NUL9	1111	Q53F54
962	Q8IX30	1012	Q5TCP6	1062	Q96JS2	1112	Q2VPA1
963	Q5TI46	1013	Q9NY75	1063	Q4ZG02	1113	Q6PK61
964	Q8NFT8	1014	Q5JP23	1064	Q5SSX3	1114	Q5RI52
965	Q969Y6	1015	Q4LE33	1065	Q53S73	1115	Q59H46
966	Q5VZK2	1016	Q9NT67	1066	Q9NY76	1116	Q5D094
967	Q5T3T8	1017	Q6ULR8	1067	Q8IV28	1117	Q59HB9
968	Q5R336	1018	Q59HG2	1068	Q86UC0	1118	Q59EX8
969	Q5JYJ8	1019	Q99533	1069	O14549	1119	Q53RA0
970	Q8NAN7	1020	Q8IV34	1070	O75440	1120	Q53QM8
971	Q8N124	1021	Q86SW0	1071	O95898	1121	Q4VB90
972	Q7Z7L6	1022	Q7Z3V1	1072	Q53SD8	1122	Q6P9E3
973	Q8WWZ8	1023	Q9NZL7	1073	Q54A24	1123	Q5IEC3
974	Q541U7	1024	Q9H195	1074	Q59F71	1124	Q9UKK5
975	Q6IMN1	1025	Q68DE5	1075	Q5IEC6	1125	Q9UKM3
976	Q52LZ6	1026	Q5PY49	1076	Q5JRP0	1126	Q8IUI0
977	Q7Z3S9	1027	Q59EE6	1077	Q5JVE8	1127	Q6N062
978	Q5U643	1028	Q9H3D5	1078	Q5KTZ5	1128	Q695G9
979	O60283	1029	Q96KG6	1079	Q5STJ3	1129	Q5U026
980	Q53QT2	1030	Q9P0Z7	1080	Q5SVA1	1130	Q8WUQ9
981	Q53RR6	1031	Q86WJ0	1081	Q659B4	1131	Q4U3E1
982	Q53SG1	1032	Q6ZTM2	1082	Q6PIA2	1132	Q9NPK9

No.	Accession	No.	Accession	No.	Accession	No.	Accession
1133	Q8WYH1	1183	Q8IUI1	1233	Q4VB88	1283	Q8WW79
1134	Q8IXB8	1184	Q8TDF8	1234	Q4ZFV5	1284	Q5IEC8
1135	Q6UXJ1	1185	Q59H72	1235	Q53FR6	1285	Q5FBE1
1136	Q5TEL3	1186	Q59FQ1	1236	Q53X47	1286	O75079
1137	Q5BKT8	1187	Q53TC0	1237	Q5T3U0	1287	Q9UKM1
1138	Q59E99	1188	Q8TDC9	1238	Q5T7T8	1288	Q96NT6
1139	Q9UHI2	1189	O14651	1239	Q5TCU2	1289	Q9UDV4
1140	Q7Z5C1	1190	Q8TF19	1240	Q6IAL4	1290	Q6ZP86
1141	Q4ZG84	1191	Q8N2D6	1241	Q6ZMN9	1291	Q6UWB0
1142	Q96JW2	1192	O95127	1242	Q7RTW4	1292	Q8N610
1143	Q8IV29	1193	Q8N9G0	1243	Q7Z3G3	1293	Q5XG79
1144	Q86UZ9	1194	Q86XN2	1244	Q8IZA9	1294	Q5VTG9
1145	Q53FK6	1195	Q6UXH9	1245	Q8N2S1	1295	Q5T3U1
1146	Q53GP0	1196	Q5SSY7	1246	Q8NBT9	1296	Q5IEC2
1147	Q53S76	1197	Q8NHD5	1247	Q8NHD3	1297	Q5CZI2
1148	Q53SK7	1198	Q566Q1	1248	Q8TB42	1298	Q58A83
1149	Q59F90	1199	Q8NAU9	1249	Q9NY67	1299	Q53TQ5
1150	Q5H8W3	1200	Q9H4D8	1250	Q9UES7	1300	Q9UDM2
1151	Q5JVF5	1201	Q8ND91	1251	Q9UFK6	1301	Q86TP7
1152	Q5JVF6	1202	Q5H8X1	1252	Q6VU67	1302	Q6QBS2
1153	Q5T7S3	1203	Q8IVT0	1253	Q99740	1303	Q6IQ50
1154	Q68D96	1204	Q75N89	1254	Q6L9N5	1304	Q5TI73
1155	Q6LBN5	1205	Q6ZWJ7	1255	Q6AZ94	1305	Q5STG5
1156	Q6R267	1206	Q6P3V1	1256	Q5TI74	1306	Q5JVE7
1157	Q7RTW3	1207	Q8TEK2	1257	Q5TBB9	1307	Q59H37
1158	Q8N2G3	1208	Q53SG3	1258	Q5HYM1	1308	Q53TB8
1159	Q8NHD2	1209	Q4VX26	1259	Q53TL0	1309	Q63HQ2
1160	Q8TDW7	1210	Q75N88	1260	Q53SF3	1310	Q96DN2
1161	Q8WU63	1211	Q96KG7	1261	Q53R88	1311	Q6ZWI1
1162	Q8WUL3	1212	Q8WYK1	1262	Q45KX0	1312	O75413
1163	Q9NQ36	1213	Q8IY13	1263	Q9UN93	1313	Q9UKD4
1164	Q9UH51	1214	Q2VYF6	1264	Q7RTV8	1314	Q96PQ8
1165	Q9ULI3	1215	Q9H1R1	1265	Q5T7T7	1315	Q6LE94
1166	Q8N369	1216	Q8N7Y0	1266	Q59ED8	1316	Q5TI72
1167	Q8IUV8	1217	Q8NBS4	1267	Q8NC23	1317	Q5T5Y9
1168	Q8IUE3	1218	Q9UF98	1268	Q9Y3V7	1318	Q9UMJ6
1169	Q86WX2	1219	Q86YZ8	1269	Q6FH69	1319	Q5IEC4
1170	Q5JP22	1220	Q86V58	1270	Q8IXK1	1320	Q5D044
1171	Q6ZUL9	1221	Q7Z5C0	1271	Q5IEC7	1321	Q53S74
1172	Q96JU7	1222	Q6V0I7	1272	Q5IEC5	1322	Q53RG4
1173	Q8WY28	1223	Q6ULR6	1273	Q503B0	1323	Q53HT9
1174	Q8TER0	1224	Q5U4N9	1274	Q3HY29	1324	O17494
1175	Q59EB6	1225	Q5JPI4	1275	Q5VVF5	1325	Q8I499
1176	Q53HS5	1226	Q2VP98	1276	Q8IUX8	1326	Q22W77
1177	Q4LDE5	1227	Q8WX98	1277	Q8N197	1327	Q8IQH0
1178	Q59EV4	1228	Q59ES6	1278	Q3KP23	1328	Q7PT06
1179	Q6T256	1229	Q4AC86	1279	Q96TF5	1329	Q5TQF9
1180	Q4W5L1	1230	Q14487	1280	Q9H3Q6	1330	Q247E2
1181	O75412	1231	Q336F5	1281	Q8IWY4	1331	Q86AL9
1182	Q8IYT0	1232	Q3HY28	1282	Q6FH22	1332	Q22WG6

No.	Accession	No.	Accession	No.	Accession	No.	Accession
1333	Q86AZ3	1383	Q25979	1433	Q5CCS9	1483	Q6LBT0
1334	Q23II7	1384	Q25980	1434	Q5CS87	1484	Q6T3J7
1335	Q26194	1385	Q25981	1435	Q5EKX6	1485	Q6W3C6
1336	Q9TVY6	1386	Q25984	1436	Q5EKX7	1486	Q6X0I2
1337	Q61UE4	1387	Q26043	1437	Q5EKX8	1487	Q70HE9
1338	Q249Q6	1388	Q26184	1438	Q5EKX9	1488	Q70HF0
1339	Q247C3	1389	Q26661	1439	Q5EKY0	1489	Q70HF2
1340	Q237H2	1390	Q2VF47	1440	Q5EKY1	1490	Q70HF3
1341	Q9U5D0	1391	Q3MLL6	1441	Q5EKY2	1491	Q70HF5
1342	Q7QCP4	1392	Q45HK1	1442	Q5EKY3	1492	Q70HF6
1343	Q86KU4	1393	Q4ABF0	1443	Q5EKY4	1493	Q70HF7
1344	Q5TX06	1394	Q4QQB7	1444	Q5EKY5	1494	Q70HF8
1345	Q25976	1395	Q4X6G1	1445	Q5EKY6	1495	Q70HG0
1346	Q7RAT4	1396	Q50ZK2	1446	Q5EKY9	1496	Q70HG4
1347	Q8T1W0	1397	Q56R22	1447	Q5EKZ0	1497	Q75K29
1348	Q247E3	1398	Q56R24	1448	Q5EKZ1	1498	Q75S85
1349	Q22JN8	1399	Q56R25	1449	Q5EKZ2	1499	Q764K8
1350	Q22HI5	1400	Q56R28	1450	Q5EKZ6	1500	Q764L3
1351	Q9VU94	1401	Q56R31	1451	Q5EL01	1501	Q7KJP5
1352	Q86AL8	1402	Q56R32	1452	Q5EL04	1502	Q7PFQ7
1353	O01768	1403	Q56R33	1453	Q5EL09	1503	Q7PM69
1354	O18366	1404	Q56R36	1454	Q5EL11	1504	Q7PM82
1355	O43995	1405	Q56R37	1455	Q5ER69	1505	Q7PME7
1356	O43996	1406	Q56R39	1456	Q5ER87	1506	Q7PMF9
1357	O43997	1407	Q5BPA8	1457	Q5ER88	1507	Q7PRJ2
1358	O46131	1408	Q5BPA9	1458	Q5ER93	1508	Q7PSM9
1359	O61307	1409	Q5BPB1	1459	Q5K5W5	1509	Q7PSQ4
1360	O62055	1410	Q5BPB3	1460	Q5K5W6	1510	Q7PT75
1361	O76727	1411	Q5BPB5	1461	Q5K5W8	1511	Q7Q1L7
1362	O96444	1412	Q5BPB7	1462	Q5K5X0	1512	Q7Q1L9
1363	P91363	1413	Q5BPC1	1463	Q5K5X1	1513	Q7Q1M9
1364	Q03999	1414	Q5BPC5	1464	Q5MQ65	1514	Q7Q1N2
1365	Q04901	1415	Q5BPC6	1465	Q5TNY8	1515	Q7Q3I9
1366	Q20979	1416	Q5BPC7	1466	Q5TQ47	1516	Q7Q440
1367	Q24551	1417	Q5BPC9	1467	Q5TV39	1517	Q7Q737
1368	Q25058	1418	Q5BPD3	1468	Q5WNK5	1518	Q7QIQ6
1369	Q25059	1419	Q5BPD4	1469	Q5XKU6	1519	Q7QK54
1370	Q25243	1420	Q5BPD7	1470	Q5XKU7	1520	Q7QQP4
1371	Q25678	1421	Q5BPD9	1471	Q60VN3	1521	Q7QT99
1372	Q25722	1422	Q5BPE0	1472	Q60VT7	1522	Q7QUF9
1373	Q25723	1423	Q5BPE2	1473	Q60VX0	1523	Q7R5J3
1374	Q25724	1424	Q5BPE4	1474	Q60Z28	1524	Q7YY60
1375	Q25726	1425	Q5BPE5	1475	Q612V7	1525	Q868H5
1376	Q25727	1426	Q5BPE6	1476	Q613J2	1526	Q868H7
1377	Q25922	1427	Q5BPE8	1477	Q615A3	1527	Q868T6
1378	Q25924	1428	Q5CCJ7	1478	Q61KA6	1528	Q868T7
1379	Q25966	1429	Q5CCS2	1479	Q61T62	1529	Q868T8
1380	Q25968	1430	Q5CCS3	1480	Q629H6	1530	Q868U1
1381	Q25974	1431	Q5CCS6	1481	Q69GU3	1531	Q868U2
1382	Q25978	1432	Q5CCS7	1482	Q69HN1	1532	Q869J5

No.	Accession	No.	Accession	No.	Accession	No.	Accession
1533	Q869S2	1583	Q9NCM9	1633	Q4D629	1683	Q5EL00
1534	Q86FL1	1584	Q9NCN0	1634	Q4R1B4	1684	Q5EL02
1535	Q86JH4	1585	Q9NCN2	1635	Q4X2L0	1685	Q5EL03
1536	Q86RD2	1586	Q9NEG1	1636	Q52VK2	1686	Q5EL05
1537	Q8I0U8	1587	Q9NFS9	1637	Q52VK3	1687	Q5EL06
1538	Q8I1K8	1588	Q9NGD4	1638	Q56R21	1688	Q5EL07
1539	Q8I1K9	1589	Q9NGD8	1639	Q56R23	1689	Q5EL08
1540	Q8I1L1	1590	Q9TX98	1640	Q56R26	1690	Q5EL10
1541	Q8I1L3	1591	Q9TYE3	1641	Q56R27	1691	Q5EL12
1542	Q8I1L8	1592	Q9TYE7	1642	Q56R29	1692	Q5EL13
1543	Q8I1L9	1593	Q9TZT5	1643	Q56R30	1693	Q5EL14
1544	Q8I1M0	1594	Q9UAI8	1644	Q56R34	1694	Q5ER64
1545	Q8I1M2	1595	Q9UB87	1645	Q56R35	1695	Q5ER74
1546	Q8I1M3	1596	Q9VNU6	1646	Q56R38	1696	Q5ER77
1547	Q8I1M4	1597	Q9VXM0	1647	Q56R40	1697	Q5ER78
1548	Q8I1M5	1598	Q9W0A0	1648	Q59E20	1698	Q5ER81
1549	Q8I1M6	1599	Q9Y0V1	1649	Q5BPB0	1699	Q5ER82
1550	Q8MQ08	1600	O61230	1650	Q5BPB2	1700	Q5ER85
1551	Q8MQJ4	1601	O61677	1651	Q5BPB4	1701	Q5ER89
1552	Q8MV49	1602	P91774	1652	Q5BPB6	1702	Q5ER92
1553	Q8MV51	1603	P92163	1653	Q5BPB8	1703	Q5I6P2
1554	Q8MV54	1604	Q02569	1654	Q5BPB9	1704	Q5K5W2
1555	Q8MY75	1605	Q19882	1655	Q5BPC0	1705	Q5K5W3
1556	Q8MY76	1606	Q20997	1656	Q5BPC2	1706	Q5K5W4
1557	Q8MY78	1607	Q21980	1657	Q5BPC3	1707	Q5K5W7
1558	Q8T5W3	1608	Q24550	1658	Q5BPC4	1708	Q5K5W9
1559	Q8T5W5	1609	Q25717	1659	Q5BPC8	1709	Q5K5X2
1560	Q8T5W7	1610	Q25718	1660	Q5BPD0	1710	Q5XKU8
1561	Q8T5X0	1611	Q25719	1661	Q5BPD1	1711	Q5XKV0
1562	Q8T5X1	1612	Q25720	1662	Q5BPD2	1712	Q5XKV1
1563	Q8T5X2	1613	Q25721	1663	Q5BPD5	1713	Q60QH2
1564	Q8T5X3	1614	Q25725	1664	Q5BPD6	1714	Q60RK8
1565	Q8T5Y8	1615	Q25728	1665	Q5BPD8	1715	Q60YX0
1566	Q93519	1616	Q25923	1666	Q5BPE1	1716	Q614N6
1567	Q95PB8	1617	Q25967	1667	Q5BPE3	1717	Q61MD6
1568	Q962W9	1618	Q25969	1668	Q5BPE7	1718	Q61PF0
1569	Q963T3	1619	Q25970	1669	Q5CCJ6	1719	Q61SB6
1570	Q964F7	1620	Q25971	1670	Q5CCS4	1720	Q61T55
1571	Q964N2	1621	Q25972	1671	Q5CCS5	1721	Q623Z8
1572	Q968Y6	1622	Q25973	1672	Q5CCS8	1722	Q659T9
1573	Q9BMG8	1623	Q25975	1673	Q5CJ96	1723	Q66NE3
1574	Q9GP66	1624	Q25977	1674	Q5CXK1	1724	Q66NE4
1575	Q9GSF3	1625	Q25982	1675	Q5EKY7	1725	Q66PY4
1576	Q9N432	1626	Q25983	1676	Q5EKY8	1726	Q66S04
1577	Q9NC90	1627	Q26183	1677	Q5EKZ3	1727	Q68QF3
1578	Q9NCM1	1628	Q2M0H4	1678	Q5EKZ4	1728	Q69GT9
1579	Q9NCM2	1629	Q3C2A0	1679	Q5EKZ5	1729	Q69GU1
1580	Q9NCM3	1630	Q3YJT7	1680	Q5EKZ7	1730	Q69GU2
1581	Q9NCM4	1631	Q45U80	1681	Q5EKZ8	1731	Q69GU4
1582	Q9NCM7	1632	Q4ABE7	1682	Q5EKZ9	1732	Q6SPF9

No.	Accession	No.	Accession	No.	Accession	No.	Accession
1733	Q6Y1L9	1783	Q8I1L5	1833	Q9TYG2	1883	Q54FR2
1734	Q70HF1	1784	Q8I1L6	1834	Q9TZT4	1884	Q54EK3
1735	Q70HF4	1785	Q8I1L7	1835	Q9U0E2	1885	Q4H3A4
1736	Q70HF9	1786	Q8I1M1	1836	Q9U483	1886	Q6AWM6
1737	Q70HG1	1787	Q8I1M7	1837	Q9UAI6	1887	Q54ZK3
1738	Q70HG2	1788	Q8I476	1838	Q9UAI7	1888	Q8T5Z1
1739	Q70HG3	1789	Q8ISC6	1839	Q9UB95	1889	Q9BIC5
1740	Q764L1	1790	Q8MN62	1840	Q9VYN8	1890	Q7QH41
1741	Q764L2	1791	Q8MQN5	1841	Q20204	1891	Q55FR0
1742	Q76NT1	1792	Q8MV50	1842	Q5ER75	1892	Q964D1
1743	Q7KJP4	1793	Q8MV52	1843	Q5I6P1	1893	Q8MVW7
1744	Q7KJP6	1794	Q8MV53	1844	Q5XKU9	1894	Q18424
1745	Q7KPY6	1795	Q8MVI9	1845	Q5XKV2	1895	Q50T88
1746	Q7PM73	1796	Q8MXN9	1846	Q6VPM9	1896	Q4Q827
1747	Q7PPC0	1797	Q8MY74	1847	Q7JWC5	1897	Q8T5Z3
1748	Q7PQ15	1798	Q8SWY0	1848	Q7Z1J0	1898	Q960R8
1749	Q7Q0M5	1799	Q8T2F8	1849	Q86PQ8	1899	Q75WG2
1750	Q7Q1M6	1800	Q8T4N9	1850	Q8I058	1900	Q628R4
1751	Q7Q1N3	1801	Q8T5U7	1851	Q8I091	1901	Q61WG8
1752	Q7Q1N4	1802	Q8T5W4	1852	Q8I0G9	1902	Q54R92
1753	Q7Q3K5	1803	Q8T5W6	1853	Q8I0K3	1903	Q519X6
1754	Q7Q6R6	1804	Q8T5W8	1854	Q8I0M9	1904	Q49BF8
1755	Q7QGY2	1805	Q8T5W9	1855	Q8I0P3	1905	Q2WBY3
1756	Q7QJQ6	1806	Q8T5X4	1856	Q8I0S8	1906	Q9W4Y3
1757	Q7QK12	1807	Q8T5Y6	1857	Q8I0U0	1907	Q8T6V0
1758	Q7QL19	1808	Q8T5Y7	1858	Q8MLX3	1908	Q61PE4
1759	Q7QR15	1809	Q8T5Y9	1859	Q8MM25	1909	Q967S8
1760	Q7QWD9	1810	Q95SP5	1860	Q8STI3	1910	Q7QAH1
1761	Q7QYW5	1811	Q964F5	1861	Q8SYF5	1911	Q75WG1
1762	Q7R369	1812	Q964F6	1862	Q93473	1912	Q61GM7
1763	Q7YW57	1813	Q964F8	1863	Q95NL3	1913	Q54N64
1764	Q7Z1J1	1814	Q964N3	1864	Q9TVG8	1914	Q869L5
1765	Q868H4	1815	Q9N657	1865	Q9VPJ0	1915	Q247C4
1766	Q868H6	1816	Q9NCM0	1866	Q9W1X5	1916	Q8MVJ7
1767	Q868T4	1817	Q9NCM5	1867	Q9XXU1	1917	Q7Z103
1768	Q868T5	1818	Q9NCM6	1868	Q54P15	1918	Q61IZ5
1769	Q868T9	1819	Q9NCM8	1869	Q54J39	1919	Q61GP9
1770	Q868U0	1820	Q9NCN1	1870	Q55F41	1920	Q248Q8
1771	Q868U3	1821	Q9NCN3	1871	Q7RGX0	1921	Q55EH0
1772	Q869J7	1822	Q9NGD3	1872	Q86RL8	1922	Q55C92
1773	Q86AJ6	1823	Q9NGD5	1873	Q9VLT6	1923	Q55FQ6
1774	Q86HL2	1824	Q9NGD6	1874	Q9BMB0	1924	Q54DI1
1775	Q86K16	1825	Q9NGD7	1875	Q7QT01	1925	Q8T3A7
1776	Q86L32	1826	Q9NGD9	1876	Q60XC0	1926	Q7PXF5
1777	Q8I1K5	1827	Q9NHX1	1877	Q5TQG1	1927	Q86L17
1778	Q8I1K6	1828	Q9TX99	1878	Q9BIA0	1928	Q7QYW9
1779	Q8I1K7	1829	Q9TYE4	1879	Q7QB67	1929	Q50PD8
1780	Q8I1L0	1830	Q9TYE5	1880	Q7Q434	1930	Q49BF9
1781	Q8I1L2	1831	Q9TYE6	1881	Q556L9	1931	Q9VB20
1782	Q8I1L4	1832	Q9TYG1	1882	Q54V92	1932	Q54UR8

No.	Accession	No.	Accession	No.	Accession	No.	Accession
1933	Q54ID8	1983	Q8WPG7	2033	Q8T0D8	2083	Q22LF8
1934	Q553F8	1984	Q75WG0	2034	O96659	2084	Q8MY80
1935	Q54ZJ9	1985	Q54FJ6	2035	Q55EA3	2085	Q22LF1
1936	Q54GR8	1986	Q55A33	2036	Q553C4	2086	Q61X43
1937	Q6NP06	1987	Q60UG3	2037	Q54ZK0	2087	Q7QE55
1938	Q75UQ6	1988	Q55G63	2038	Q54LT9	2088	Q8MSR5
1939	Q7R630	1989	Q55AP8	2039	Q54L67	2089	Q61QY1
1940	Q558U5	1990	Q60QD4	2040	Q54GR7	2090	Q9GPM8
1941	Q54U93	1991	Q6BG85	2041	Q54BI9	2091	Q54QK9
1942	Q54U77	1992	Q55A34	2042	Q517Y7	2092	Q86KZ0
1943	Q50W94	1993	Q54VZ0	2043	Q54VT3	2093	Q86KD9
1944	Q54P41	1994	Q55E84	2044	P91904	2094	Q7QSE0
1945	Q54H59	1995	Q54QK8	2045	Q8MPN3	2095	Q7QYY8
1946	Q54VT0	1996	Q5TVP3	2046	Q9N6J6	2096	Q4YJ87
1947	Q517D9	1997	Q8IK13	2047	Q9GPN0	2097	Q9TYU4
1948	Q868U4	1998	Q8T3A6	2048	Q9N6R5	2098	Q9Y151
1949	Q7PPU8	1999	Q8T5Z2	2049	Q551V5	2099	Q8I7T3
1950	Q559R8	2000	Q9BPS2	2050	Q54GX9	2100	Q54KC4
1951	Q4W4T0	2001	Q7R4H1	2051	Q5K6R7	2101	Q6AWJ8
1952	Q550E1	2002	Q59JG2	2052	Q54W81	2102	Q7Q5D1
1953	Q556M2	2003	Q55E77	2053	Q54VZ1	2103	Q9W0A1
1954	Q9VJU9	2004	Q54LE5	2054	Q7R5E3	2104	Q76P25
1955	Q9VQI2	2005	Q54GM8	2055	Q8IRV7	2105	Q54BJ0
1956	Q9XWD6	2006	Q54DY4	2056	Q86JH0	2106	Q8T3A0
1957	Q8STG0	2007	Q86KF0	2057	Q7QJR9	2107	Q9GYK2
1958	Q9GRG4	2008	Q86AS3	2058	Q55AW5	2108	Q54VZ2
1959	O61126	2009	Q869K4	2059	Q8WRF4	2109	Q9U1T9
1960	Q9NL29	2010	Q55CD0	2060	Q7PRP5	2110	Q86GF3
1961	Q54GV4	2011	Q75JS9	2061	Q60Y28	2111	Q5K6R6
1962	Q8T5Z0	2012	Q54I84	2062	Q54HK3	2112	Q9XXZ7
1963	Q5TQ36	2013	O18375	2063	Q23JG5	2113	Q23JG3
1964	Q6NN26	2014	Q9GPM9	2064	Q7PR44	2114	Q9NAS7
1965	Q5MAQ8	2015	Q7QFS2	2065	Q60WL4	2115	Q9U1T8
1966	Q54Z87	2016	Q86KY7	2066	Q54VY9	2116	Q5BI30
1967	Q54XI0	2017	Q54N02	2067	Q54M46	2117	Q8IP58
1968	Q50WT5	2018	Q75JA5	2068	Q86GF4	2118	Q61CA8
1969	Q9NFW6	2019	Q54VD9	2069	Q22LF6	2119	Q61FT2
1970	Q6LF51	2020	Q54RJ9	2070	Q8IRV8	2120	Q55BS2
1971	Q55EK0	2021	Q54QL1	2071	Q54VZ3	2121	Q7PN80
1972	Q8MVW6	2022	Q8T314	2072	Q54VV3	2122	O44247
1973	Q9NL27	2023	Q86KY8	2073	Q61WT5	2123	Q22LF2
1974	Q7QZU9	2024	Q76NT5	2074	Q54QV6	2124	Q22BY2
1975	Q54N14	2025	Q55FB4	2075	Q23VZ7	2125	Q9W4Y4
1976	Q54GB7	2026	Q55EN6	2076	Q23BC4	2126	Q969A0
1977	Q7QEZ5	2027	Q55AW6	2077	Q26051	2127	Q7PSL4
1978	Q5TUY7	2028	Q54KN0	2078	Q69GT8	2128	Q55A32
1979	Q5TUY6	2029	Q9NL28	2079	Q55FP4	2129	Q550I2
1980	Q55GF6	2030	Q54L60	2080	O18482	2130	Q9XTS9
1981	Q54TX0	2031	Q86IA2	2081	Q23409	2131	Q54UI7
1982	Q54C09	2032	Q23G21	2082	Q23PA1	2132	Q60XP5

No.	Accession	No.	Accession	No.	Accession	No.	Accession
2133	Q54HZ6	2183	Q5C2K6	2233	Q22423	2283	Q8WTP0
2134	Q54K30	2184	Q5C1F4	2234	Q60T93	2284	Q95YG0
2135	Q54ZQ4	2185	Q5DH33	2235	O17829	2285	Q93791
2136	Q554N7	2186	Q5C0L6	2236	Q4YRN8	2286	Q9VBN1
2137	P91808	2187	Q5BXK8	2237	Q7QWR4	2287	Q93563
2138	Q9UAG0	2188	Q95P95	2238	Q54LS4	2288	Q9XYX0
2139	Q550G2	2189	Q27422	2239	Q8IGR9	2289	O45000
2140	Q22Z95	2190	O76952	2240	Q7YXD2	2290	Q6NP71
2141	Q551T2	2191	Q867Q2	2241	Q7YU36	2291	Q610T0
2142	Q86GF2	2192	O45614	2242	Q627U2	2292	Q5I5Q9
2143	Q1ZXF4	2193	Q8IFX2	2243	Q5W4Z0	2293	Q61SX3
2144	Q54W82	2194	Q86PP8	2244	Q5TWT0	2294	Q764L0
2145	Q55GF5	2195	Q967F4	2245	Q4ABE8	2295	O45602
2146	Q54VS9	2196	Q9GSA3	2246	Q2TCK8	2296	Q9GNU3
2147	Q23HW1	2197	O01335	2247	Q2TCK5	2297	Q7YU01
2148	Q23RP1	2198	Q9VC47	2248	Q9XWC4	2298	Q86K70
2149	O76809	2199	P90974	2249	Q61YC5	2299	P90891
2150	Q8IRV9	2200	Q95RU0	2250	Q9U4E4	2300	Q54JT2
2151	Q75K09	2201	Q95RQ1	2251	Q7PS28	2301	Q50JF9
2152	Q558U0	2202	Q9VVJ6	2252	Q60SY5	2302	Q4ABE9
2153	Q54M48	2203	Q764K9	2253	Q7JNV6	2303	Q17537
2154	Q54GB8	2204	Q61GZ2	2254	Q7KQX6	2304	Q23587
2155	Q86IL0	2205	Q9VJT5	2255	Q9NGV2	2305	Q68K25
2156	Q9NL26	2206	Q622G6	2256	Q8IP51	2306	Q7QS95
2157	Q54M43	2207	Q7PV65	2257	Q86G85	2307	Q54IA3
2158	Q9U3U7	2208	Q869J8	2258	Q8I6X6	2308	Q61JN3
2159	Q75S84	2209	Q86SD6	2259	Q21281	2309	Q5WN34
2160	Q95YK2	2210	O45201	2260	Q55FZ0	2310	Q9V383
2161	Q23YX9	2211	Q61KX2	2261	Q7PPF9	2311	Q19267
2162	Q7QZ44	2212	Q61GU4	2262	Q1HAY7	2312	Q23995
2163	Q7QEZ2	2213	Q964N4	2263	Q9V6Q0	2313	Q9XZC9
2164	Q22TL6	2214	Q50T79	2264	Q9GU69	2314	Q9NGV4
2165	Q961N3	2215	Q19350	2265	Q7Q0Z8	2315	Q21850
2166	O97189	2216	Q75WV8	2266	Q8MVP0	2316	Q7QUV9
2167	Q8I3A6	2217	Q8WPN0	2267	Q7JNV7	2317	Q20043
2168	Q7KWS7	2218	Q9VQ47	2268	Q8MVK6	2318	Q400N0
2169	Q9NEF9	2219	Q9W332	2269	Q8WS87	2319	Q2YI44
2170	Q55CJ5	2220	Q20852	2270	Q7Q3P0	2320	Q61XY3
2171	Q86JH3	2221	Q95Y10	2271	Q7QGV0	2321	Q962I8
2172	Q55AW2	2222	Q7QYS1	2272	Q22675	2322	Q45VP9
2173	Q869J6	2223	Q7QCT2	2273	Q21015	2323	Q9VBN0
2174	Q7YSR5	2224	Q7QR01	2274	Q21849	2324	Q7YWF4
2175	Q19482	2225	Q623I2	2275	Q61P38	2325	Q5WN95
2176	O44327	2226	O16265	2276	Q61A32	2326	Q8IQ18
2177	Q19319	2227	Q5CGS1	2277	Q7YTZ6	2327	Q09538
2178	Q9TVQ2	2228	Q18291	2278	Q5IX63	2328	Q18761
2179	Q5C3E4	2229	Q3KN41	2279	Q7KU08	2329	Q95RA3
2180	Q5DAM6	2230	Q2TCK7	2280	Q5DWF3	2330	O16004
2181	Q5C7G7	2231	Q9W343	2281	Q54YP0	2331	Q9U2D5
2182	Q5C5F4	2232	Q19617	2282	Q4H3Q7	2332	Q7QYH8

No.	Accession	No.	Accession	No.	Accession	No.	Accession
2333	Q7PQG9	2383	Q54PD6	2433	Q9VM55	2483	Q8MJN7
2334	Q614U4	2384	Q95V09	2434	O61240	2484	Q9TUN7
2335	Q60M20	2385	Q86P79	2435	Q61DQ8	2485	Q863C4
2336	Q60WS2	2386	Q7QK77	2436	Q54QE0	2486	Q6IT40
2337	O01552	2387	Q86L15	2437	Q22574	2487	Q28219
2338	Q7PH68	2388	Q86AQ9	2438	Q2TCK6	2488	O77718
2339	Q61X71	2389	Q622Z9	2439	Q86B77	2489	Q8MJN2
2340	Q60UH7	2390	Q616A5	2440	Q8T4N8	2490	Q8MJN9
2341	Q22913	2391	Q4H3Q6	2441	Q9BIM7	2491	O97702
2342	Q17657	2392	Q7QER8	2442	Q9VCZ9	2492	Q5VI41
2343	Q60TA7	2393	Q8SZX4	2443	Q9VS89	2493	Q6ECI6
2344	Q61T33	2394	Q8T9S1	2444	Q8WTJ9	2494	Q9TUN3
2345	Q61EJ2	2395	Q7QI04	2445	P91526	2495	Q8MJN1
2346	Q61AY4	2396	Q7QEK9	2446	Q7PFH4	2496	Q6H8Q4
2347	Q9XUF9	2397	Q69GU0	2447	Q61UE2	2497	Q59I58
2348	Q9VXL1	2398	Q60UG0	2448	O44565	2498	Q4R4F4
2349	Q8I497	2399	Q20971	2449	Q7R4V2	2499	Q28218
2350	Q7QEF7	2400	Q2XWP5	2450	Q7PV66	2500	Q29094
2351	Q623K4	2401	Q5TQL0	2451	Q7PS35	2501	Q59I57
2352	Q9VZ44	2402	Q9BLJ1	2452	Q8I498	2502	Q8MJN8
2353	Q8T4P0	2403	Q23410	2453	Q9GPA5	2503	Q8MJN5
2354	Q8IQG6	2404	Q7Z1P7	2454	Q7PTG9	2504	Q9GK49
2355	Q9VBN2	2405	Q7K6V2	2455	Q61H26	2505	Q6PT99
2356	Q8MP02	2406	Q61JH1	2456	Q86KE8	2506	Q5NKT5
2357	Q9GZ15	2407	Q5CJ94	2457	Q3V641	2507	Q28290
2358	Q21884	2408	Q2WBY6	2458	Q6QJC4	2508	Q5RDL9
2359	Q9V4J6	2409	Q9VZ96	2459	Q26423	2509	O18958
2360	Q25253	2410	Q60UE2	2460	Q8MQF7	2510	O19061
2361	Q20219	2411	Q61SI8	2461	Q71A42	2511	O77505
2362	Q60FX5	2412	Q4DUT3	2462	Q4W2V7	2512	P79199
2363	Q5TP37	2413	Q2LYI6	2463	Q8MP01	2513	Q07112
2364	Q550A1	2414	Q6NP66	2464	Q7YY59	2514	Q28485
2365	Q8MY77	2415	Q9BIJ2	2465	Q7QJ41	2515	Q28659
2366	Q17377	2416	Q9NHE9	2466	Q7PSV8	2516	Q28867
2367	Q967H9	2417	Q23046	2467	Q6YID6	2517	Q307K2
2368	Q95Q39	2418	Q9UB94	2468	Q5CJG0	2518	Q307K3
2369	Q7PSY6	2419	Q21756	2469	Q52V41	2519	Q307K6
2370	Q7QPM3	2420	Q9V5J7	2470	Q61K85	2520	Q30DU2
2371	Q628R5	2421	Q61PE7	2471	Q628B7	2521	Q3KS04
2372	Q60PS2	2422	Q5TU01	2472	Q61WG1	2522	Q4KTX1
2373	Q4H2P2	2423	Q5TTZ3	2473	Q8MJN6	2523	Q4R6R6
2374	Q9BI05	2424	Q3L453	2474	Q95N85	2524	Q4R728
2375	Q4W2V6	2425	Q86FJ9	2475	Q9TUN5	2525	Q5ISN4
2376	Q8MRJ7	2426	Q9TXA0	2476	Q8MJN4	2526	Q5NVF0
2377	Q19853	2427	Q95SN5	2477	Q5R8W2	2527	Q5R3Z7
2378	Q17494	2428	Q20535	2478	Q6UTY0	2528	Q5R9X4
2379	Q24132	2429	Q7KUY7	2479	Q5MAR3	2529	Q5RDB0
2380	Q9V7I4	2430	Q61K66	2480	Q5ISL2	2530	Q5RDI5
2381	Q7R2Y9	2431	Q60JW4	2481	Q9TU04	2531	Q6S4M1
2382	O44191	2432	Q60IF3	2482	Q8MJN3	2532	Q6S4M2

No.	Accession	No.	Accession	No.	Accession	No.	Accession
2533	Q6Y8C5	2583	Q2THU6	2633	Q95ND6	2683	Q93W14
2534	Q866H0	2584	Q5RDI1	2634	Q5R6R1	2684	Q93W13
2535	Q8MJK0	2585	Q9GMD9	2635	Q5R6S9	2685	Q53J34
2536	Q8SQB9	2586	Q5R7K9	2636	Q43366	2686	Q94FI0
2537	Q95JH1	2587	Q2TA09	2637	O04927	2687	Q94FI2
2538	O18977	2588	Q2Q420	2638	Q6ZK10	2688	Q94FI9
2539	O19056	2589	Q4R3X4	2639	Q7M1T5	2689	Q94FJ5
2540	O19060	2590	Q7M304	2640	Q70AH9	2690	Q94FK1
2541	O77501	2591	Q5J3Q6	2641	Q76BK3	2691	Q94FK5
2542	O77779	2592	Q9N1S4	2642	Q93Z38	2692	Q94FK6
2543	Q28484	2593	Q5RA99	2643	Q9M7L9	2693	Q94FH3
2544	Q28629	2594	Q864U4	2644	Q9SVX7	2694	Q94FG8
2545	Q28982	2595	Q28657	2645	Q9SYV1	2695	Q94FH5
2546	Q307E7	2596	Q9GM41	2646	Q6NKR6	2696	Q94FJ7
2547	Q307K4	2597	Q9BG62	2647	Q942G1	2697	Q94FI5
2548	Q307K5	2598	Q8HZR1	2648	Q942G4	2698	Q94FK8
2549	Q3MHK6	2599	Q95LG3	2649	Q94FJ1	2699	Q94FJ8
2550	Q3YAN4	2600	Q4G408	2650	Q94LI5	2700	Q94FI7
2551	Q52MQ6	2601	Q3SWW8	2651	Q9Y0F7	2701	Q94FH6
2552	Q5E9P5	2602	Q9GLF0	2652	Q94FK2	2702	Q94FG9
2553	Q5IS86	2603	Q5ISR9	2653	Q94FK3	2703	Q94FJ3
2554	Q5R9N1	2604	Q2Q422	2654	Q9Y0F6	2704	Q94FH0
2555	Q5RBP1	2605	Q9MZF7	2655	Q94FI8	2705	Q94FK7
2556	Q5RC76	2606	Q8HZ48	2656	Q9Y0F8	2706	Q94FH1
2557	Q9GL46	2607	Q38J75	2657	Q1S4J9	2707	Q94FJ4
2558	Q9N120	2608	Q28483	2658	Q1SXI6	2708	Q94FH2
2559	Q29097	2609	Q28482	2659	Q1SAA0	2709	Q94FI1
2560	Q8MJ16	2610	Q2Q426	2660	Q1SXF6	2710	Q94FH4
2561	Q2Q425	2611	Q2Q421	2661	Q1S0W6	2711	Q2PEV5
2562	Q4R8Q9	2612	Q9BDH4	2662	Q1SR23	2712	Q94FI3
2563	Q2T9U6	2613	Q5EH71	2663	Q1SF50	2713	Q94FJ2
2564	Q2PPL3	2614	Q9BDH3	2664	Q1SF52	2714	Q94FH9
2565	Q867A1	2615	Q6E0K3	2665	Q1S7H1	2715	Q94FL4
2566	Q307G7	2616	Q9BG80	2666	Q1S720	2716	Q94FJ9
2567	Q2VJ42	2617	O46652	2667	Q1T6P5	2717	Q94FI4
2568	Q2Q419	2618	Q95LG2	2668	Q6PLP7	2718	Q94FK0
2569	Q2KIT5	2619	Q3MHW2	2669	Q1SF48	2719	Q94FK4
2570	Q2Q424	2620	Q9N028	2670	Q1SF47	2720	Q94FH8
2571	Q1HK35	2621	Q866A8	2671	Q2QM75	2721	Q94FH7
2572	Q6TGK9	2622	Q5ISM4	2672	Q2R229	2722	Q94FK9
2573	Q2TBR4	2623	Q8SPR3	2673	Q2QM72	2723	Q94FL2
2574	Q867A2	2624	Q6Q144	2674	Q53J44	2724	Q94FL1
2575	Q8MKB1	2625	Q5R8J0	2675	Q6AVG4	2725	Q94FL0
2576	Q9TVB3	2626	Q6XL67	2676	Q9FE98	2726	Q94FL3
2577	Q2PFZ7	2627	Q95LN0	2677	Q9FX14	2727	Q2E1N6
2578	Q5RC26	2628	Q8SPQ9	2678	Q67ZD0	2728	Q2XAP8
2579	Q8SQ23	2629	Q4F9K9	2679	Q3E6U8	2729	Q21I97
2580	O46370	2630	Q3T0K7	2680	Q93W11	2730	Q36UM3
2581	Q28476	2631	Q8HZR0	2681	Q93VX0	2731	Q9QW15
2582	O19057	2632	Q75PQ9	2682	Q93W15	2732	Q9QW16

No.	Accession	No.	Accession	No.	Accession	No.	Accession
2733	Q8VHL6	2783	Q8R508	2833	Q70UZ7	2883	Q3UZ32
2734	Q3TA36	2784	Q99PJ3	2834	Q8BGP3	2884	Q4FJT2
2735	Q3TSW6	2785	Q9QW24	2835	Q9CYA0	2885	Q6LCD7
2736	Q3UGR3	2786	Q9WV36	2836	Q8JZM4	2886	Q810X1
2737	Q3UH52	2787	Q3TQ06	2837	Q3V0H1	2887	Q8BRP7
2738	Q3UV83	2788	Q8JZM8	2838	Q543K3	2888	Q8CG43
2739	Q4VAA3	2789	Q99ND0	2839	Q52KG2	2889	Q8VI54
2740	Q52NV1	2790	Q8CGA7	2840	Q543T1	2890	Q920Y4
2741	Q52NV2	2791	Q3UTJ7	2841	Q8CBF7	2891	Q9ES77
2742	Q5EBA7	2792	Q3UQ22	2842	Q3TYU1	2892	Q9ESA9
2743	Q5I0H1	2793	Q570Z4	2843	Q3TKX9	2893	Q9JM06
2744	Q68FG9	2794	Q8VIB7	2844	Q91V90	2894	Q810X2
2745	Q69ZY7	2795	Q4KLK5	2845	Q545E4	2895	Q6IMH8
2746	Q6LD95	2796	P97556	2846	Q542I8	2896	Q6DR99
2747	Q71SA3	2797	Q99K58	2847	Q8C8N3	2897	Q8R1U8
2748	Q810X0	2798	Q924Z9	2848	Q9DBU9	2898	P70570
2749	Q8BXY5	2799	Q66PY1	2849	Q8BGI2	2899	Q9QWQ1
2750	Q8C8E4	2800	Q3V5L4	2850	Q91V88	2900	Q8CIL6
2751	Q8CGL6	2801	Q8K3U5	2851	Q8CGB2	2901	Q8BM43
2752	Q8R2H2	2802	Q62561	2852	Q8R3D3	2902	Q80T91
2753	Q91WZ4	2803	Q8BX76	2853	Q3U6N3	2903	Q6IMH7
2754	Q924Z8	2804	Q3V364	2854	Q925U3	2904	Q68EF1
2755	Q9CVK2	2805	Q3KR76	2855	Q5FW64	2905	Q4G029
2756	Q9D4E9	2806	Q68FE0	2856	Q8VIK5	2906	Q3TLU3
2757	Q9DAU5	2807	Q8BMI5	2857	Q3U3V1	2907	Q9ESA3
2758	Q9QXG1	2808	Q543J8	2858	Q542C2	2908	Q8C6Z2
2759	Q9R1K1	2809	Q543W3	2859	Q3U1W7	2909	Q2PZL6
2760	Q9Z135	2810	Q5NBW8	2860	Q3TR66	2910	Q3UM88
2761	O35947	2811	Q3TTE0	2861	Q6DIB5	2911	Q6NV58
2762	O54796	2812	P97806	2862	Q6PE70	2912	Q8CFA3
2763	Q3TRG0	2813	Q6PDN4	2863	Q5SV72	2913	Q80WX4
2764	Q3UDX3	2814	Q3TBR2	2864	Q5SSN6	2914	Q6IR12
2765	Q3UMN9	2815	O35370	2865	Q5SV70	2915	Q3UU65
2766	Q3UQ49	2816	Q9Z0L5	2866	P70534	2916	Q3ULY0
2767	Q3UQR6	2817	Q5SSV4	2867	Q8K061	2917	Q3UEI7
2768	Q3UVX6	2818	Q3U697	2868	Q80VN5	2918	Q3U8R7
2769	Q3UYG5	2819	Q5SS56	2869	Q9JJZ5	2919	Q3U3Y2
2770	Q3V1D4	2820	Q3U1J7	2870	Q3UPV5	2920	Q3TTE2
2771	Q4L136	2821	Q5NBW7	2871	Q3U5F6	2921	O35727
2772	Q5F226	2822	Q545J3	2872	Q6V0K7	2922	Q80XH2
2773	Q5Y4N7	2823	Q8K2B8	2873	Q3TDU5	2923	Q3V029
2774	Q60815	2824	Q6PFV7	2874	Q7M763	2924	Q3UG15
2775	Q63661	2825	Q32MF1	2875	Q811Q3	2925	Q8CJA0
2776	Q6DR98	2826	Q6PAP2	2876	Q80WT7	2926	Q9WTS7
2777	Q6LBN0	2827	Q7TQF0	2877	O08743	2927	Q99L24
2778	Q7M0A9	2828	Q3UHH3	2878	O09182	2928	Q9EQC6
2779	Q810Y3	2829	Q6P7U0	2879	O70474	2929	Q5BK84
2780	Q8BUT8	2830	Q6PIP9	2880	Q336F6	2930	Q3TD57
2781	Q8BYG9	2831	Q8K002	2881	Q3U2F0	2931	Q60472
2782	Q8C3Z5	2832	Q3U5J2	2882	Q3UGZ9	2932	Q66HK9

No.	Accession	No.	Accession	No.	Accession	No.	Accession
2933	Q6P550	2983	Q3UYW9	3033	Q8C088	3083	Q3TYB4
2934	Q3TVR4	2984	Q3V1J8	3034	Q8BKJ4	3084	Q3TZE2
2935	Q8BPM8	2985	Q68EF9	3035	Q8BNH3	3085	Q3UHB0
2936	Q811K6	2986	Q70HX0	3036	Q5SRA2	3086	Q3UHK6
2937	Q7TQG7	2987	Q80UF6	3037	Q5DU39	3087	Q3V1S6
2938	Q7TMT3	2988	Q80V56	3038	Q52L98	3088	Q543X8
2939	Q923T5	2989	Q8BPP6	3039	Q3TWK8	3089	Q5ND28
2940	Q571H4	2990	Q8BVU1	3040	Q6PCS0	3090	Q5ZQU0
2941	Q571B5	2991	Q8C4U8	3041	Q69ZY6	3091	Q7TSG9
2942	Q505C9	2992	Q8CDV5	3042	Q60816	3092	Q80YC5
2943	Q61964	2993	Q8CG21	3043	Q3U5T0	3093	Q80YS4
2944	Q8CCT8	2994	Q8R4T6	3044	Q5PQQ8	3094	Q80YX0
2945	Q80Y08	2995	Q91YY0	3045	Q99K64	3095	Q8C1R8
2946	Q9QYV1	2996	Q99L19	3046	Q8C9Q4	3096	Q8K271
2947	Q9JJS0	2997	Q9ESA7	3047	Q8BX64	3097	Q8VH41
2948	Q8CGQ1	2998	Q9JJS1	3048	Q810H2	3098	Q9QYZ1
2949	Q8BTU0	2999	Q7TSB4	3049	Q66HK3	3099	Q8K0J4
2950	Q8BM06	3000	Q3TV46	3050	Q5M879	3100	Q8CA82
2951	Q8R5G5	3001	Q9R0C7	3051	Q9Z0Y6	3101	Q8BU25
2952	Q8CJG7	3002	Q6P779	3052	Q63404	3102	Q3TIW5
2953	Q3U428	3003	Q58GH6	3053	Q8C9J2	3103	Q52KG8
2954	Q3S2T6	3004	Q3URY7	3054	Q80YQ1	3104	Q99KR2
2955	Q9JJM4	3005	Q9WTS4	3055	Q80Y26	3105	Q3UE21
2956	Q6DFX1	3006	Q91ZJ1	3056	Q6A051	3106	Q3U492
2957	Q8R0Y0	3007	O08745	3057	Q68HV2	3107	Q8VDV0
2958	Q9R1K0	3008	Q8BKS4	3058	Q5Y4N8	3108	Q8BKK7
2959	Q91ZD3	3009	Q3U515	3059	Q5DTT5	3109	Q61965
2960	Q8CJG6	3010	Q3U1W3	3060	Q571L9	3110	Q60784
2961	Q8C269	3011	Q3TNZ8	3061	Q3UV74	3111	Q8CDV3
2962	Q80UM5	3012	Q3TDD1	3062	Q3UMR6	3112	Q8R226
2963	Q7TQ52	3013	Q3MID1	3063	Q3UGU1	3113	Q9CUT3
2964	Q6KAT1	3014	Q3TCH1	3064	Q3UG73	3114	Q8C6L2
2965	Q9D4F0	3015	Q2VC84	3065	Q3TS72	3115	Q6KDN2
2966	Q52R82	3016	Q80VP6	3066	Q32NZ3	3116	Q5SRA4
2967	Q4FJS7	3017	O08744	3067	Q2VWQ2	3117	Q3UI29
2968	Q3UHL7	3018	Q91XL5	3068	Q9WTS5	3118	Q3TDU8
2969	Q3U273	3019	Q8BVP9	3069	Q6ZQ25	3119	Q6NZL8
2970	Q3TQ80	3020	Q9ESB1	3070	Q4VBE4	3120	Q80WW7
2971	Q3TC49	3021	Q91WH9	3071	Q8VHL7	3121	Q9DAW5
2972	Q5U215	3022	Q5EBX5	3072	O88424	3122	Q7TMV2
2973	Q60410	3023	Q571K3	3073	Q9R1K2	3123	Q8BLZ2
2974	Q642D0	3024	Q3UZP8	3074	Q6PER0	3124	O09020
2975	Q642C9	3025	Q3UZ23	3075	Q5SNU0	3125	Q9CRX6
2976	Q5XI24	3026	Q80VA2	3076	Q5D070	3126	Q8CHF0
2977	Q501P1	3027	Q9JLC1	3077	Q3UWJ3	3127	Q8C435
2978	Q3TQE9	3028	Q99MN7	3078	Q64FW1	3128	Q60789
2979	Q3TTP6	3029	O35452	3079	Q8K0P5	3129	Q9ESA2
2980	Q3TZC6	3030	Q8R5G0	3080	O70534	3130	Q5SSN5
2981	Q3UES1	3031	Q8C8K0	3081	O88460	3131	O88459
2982	Q3UNG0	3032	Q8C4T5	3082	Q3TW70	3132	Q571J3

No.	Accession	No.	Accession	No.	Accession	No.	Accession
3133	Q52KQ5	3183	Q80V54	3233	Q63762	3283	Q68YS7
3134	Q3V0Y9	3184	Q80VQ7	3234	Q9ESA5	3284	Q5YF64
3135	Q3UTN9	3185	Q8CE01	3235	Q8C9U1	3285	Q4KSD1
3136	Q3UH67	3186	Q9ESA1	3236	Q6ZQA1	3286	Q9JF32
3137	Q3TWM3	3187	Q9ESA8	3237	Q8R4V5	3287	Q91T36
3138	Q3TVN5	3188	Q9QXA3	3238	Q5DTL5	3288	Q9DHU7
3139	Q8R465	3189	O35883	3239	Q4VAI3	3289	Q5DLW0
3140	Q8VD97	3190	Q9WU10	3240	Q3UWD7	3290	Q8B4N0
3141	Q8K428	3191	Q8VD07	3241	Q3TPN0	3291	Q71G60
3142	Q8CI01	3192	Q9CXD8	3242	Q3TL35	3292	Q7T427
3143	Q99KT4	3193	O88840	3243	Q3TNW1	3293	Q49PZ3
3144	Q8BR22	3194	Q9WUH9	3244	Q62779	3294	Q5IY12
3145	Q4G063	3195	Q8BZG2	3245	P97883	3295	Q3I7V8
3146	Q561K2	3196	Q8BNE9	3246	Q9WUH8	3296	Q8V573
3147	O88458	3197	Q8VCS4	3247	Q8K326	3297	Q2VJ96
3148	Q9JJS9	3198	Q6P6T8	3248	Q6GTJ9	3298	O41506
3149	Q8CIP1	3199	Q7TQB4	3249	Q69ZB1	3299	Q6QVZ0
3150	Q8CGQ2	3200	Q569V0	3250	Q3UI55	3300	P87605
3151	Q9ESA6	3201	Q3UVN4	3251	Q3TGL4	3301	Q7TDW3
3152	Q76LU2	3202	Q3TDB9	3252	Q336F3	3302	Q8QN40
3153	Q6IRL1	3203	Q925V4	3253	Q3TD94	3303	Q8JRQ1
3154	Q3UHN1	3204	Q8C536	3254	Q8VHF4	3304	Q7TDW4
3155	Q3UFB4	3205	Q9ESB0	3255	Q8K0H9	3305	O41504
3156	Q7M762	3206	Q8VCT0	3256	Q922H0	3306	Q9Q9F3
3157	Q7TQ50	3207	Q91ZX7	3257	Q8R542	3307	Q5RJ05
3158	Q544J9	3208	Q63124	3258	Q62287	3308	Q5RGG6
3159	Q6PCM6	3209	Q8BMS0	3259	Q8BMD9	3309	Q4SWH5
3160	Q3UQK2	3210	Q70E20	3260	Q7TN15	3310	Q4RTS1
3161	P70628	3211	Q6TYY9	3261	Q6ZQ56	3311	Q804X5
3162	Q9CWC8	3212	Q5W9H8	3262	Q6NZM2	3312	Q5SPD2
3163	Q8BSJ0	3213	Q5M7W9	3263	Q68FY8	3313	Q5RIP8
3164	Q7TQ51	3214	Q3U454	3264	Q58A84	3314	Q5RI70
3165	Q5RKM8	3215	Q3TR50	3265	Q543Q2	3315	Q5RI68
3166	Q3URX7	3216	Q61204	3266	Q3USI2	3316	Q5RGH8
3167	Q3UA33	3217	Q811T0	3267	Q3US54	3317	Q4RUN9
3168	Q3TS86	3218	Q8K2B7	3268	Q3US45	3318	Q7ZYZ9
3169	Q3TDF9	3219	Q3UWD0	3269	Q3UND5	3319	Q6DF97
3170	Q3T9K7	3220	Q3V0B1	3270	Q3UEV6	3320	Q5RI93
3171	Q6P9K9	3221	Q61291	3271	Q6PFE7	3321	Q4SUA0
3172	Q6AYF4	3222	Q9WTS6	3272	Q8BJB5	3322	Q4SNN7
3173	O70465	3223	Q80XT9	3273	Q3TWH6	3323	Q4SNM9
3174	Q3TDN7	3224	Q9R1J9	3274	Q8R417	3324	Q4SHY2
3175	Q3TP84	3225	Q810R6	3275	Q0VGR4	3325	Q4RJ05
3176	Q3TR40	3226	Q5SRA3	3276	Q77GR0	3326	Q3MKM9
3177	Q3U8S9	3227	Q569W5	3277	Q70GU8	3327	Q9PUU4
3178	Q497H5	3228	Q3UEY9	3278	Q77PC4	3328	Q804X1
3179	Q52NV3	3229	Q3U7G2	3279	Q91MZ0	3329	Q7T011
3180	Q6KAQ6	3230	Q3U1R3	3280	Q8QUT7	3330	Q5SPB5
3181	Q6MG89	3231	Q8C0M0	3281	Q77GE7	3331	Q4SHL8
3182	Q7TQ06	3232	Q9QVT6	3282	Q6VZ62	3332	Q2UZ96

No.	Accession
3333	Q5RGN7
3334	Q8AYT0
3335	Q804X2
3336	Q804W8
3337	Q4S977
3338	Q6PCH8
3339	Q6GP06
3340	Q6IRR8
3341	Q4SB52
3342	Q07012
3343	Q804X7
3344	Q804X0
3345	Q4SUA1
3346	O42347
3347	O42507
3348	P87363
3349	Q2VU96
3350	Q32NR2
3351	Q3Y6S4
3352	Q4R1B5
3353	Q4R1B6
3354	Q4RBW8
3355	Q4RCI7
3356	Q4RDX0
3357	Q4RFP1
3358	Q4RG82
3359	Q4RJE7
3360	Q4RLT5
3361	Q4RMC1
3362	Q4RN50
3363	Q4RPY1
3364	Q4RQ03
3365	Q4RQ74

Appendix C

Dataset B

No.	Accession	No.	Accession	No.	Accession	No.	Accession
1	Q4T8F6	33	Q9DDR6	65	Q4RU98	97	Q4T0K3
2	Q4T8K3	34	Q9DED0	66	Q4RUP1	98	Q4T3H9
3	Q4T8L6	35	Q9DER5	67	Q4RUS3	99	Q4T3Y2
4	Q4TBF7	36	Q9PS95	68	Q4S2G3	100	Q4T4W9
5	Q4TFA3	37	Q9PUC7	69	Q4S2S0	101	Q4T5A4
6	Q4THW6	38	Q9W7R3	70	Q4S3C4	102	Q4T5X5
7	Q4TI75	39	Q9W7R4	71	Q4S409	103	Q4T7I2
8	Q4TIJ4	40	O57659	72	Q4S488	104	Q4T8G9
9	Q4U0S1	41	P79708	73	Q4S5N7	105	Q4T963
10	Q502D2	42	Q30A07	74	Q4S5N8	106	Q4T9U6
11	Q52KT2	43	Q32N65	75	Q4S6A5	107	Q4TCQ5
12	Q5F3N3	44	Q3S2J2	76	Q4S9N6	108	Q4TGQ1
13	Q5FVX1	45	Q4RA77	77	Q4SB67	109	Q4THK4
14	Q5U4U1	46	Q4RE58	78	Q4SDH3	110	Q5NJJ5
15	Q5XLP7	47	Q4REA6	79	Q4SF34	111	Q5NJJ6
16	Q6DHG1	48	Q4RG72	80	Q4SFH7	112	Q5NJK1
17	Q6DJD9	49	Q4RG83	81	Q4SFI1	113	Q5PPZ2
18	Q6GN32	50	Q4RIP1	82	Q4SH71	114	Q5U4N0
19	Q6IR63	51	Q4RJT4	83	Q4SHC0	115	Q5XHG8
20	Q6J1M9	52	Q4RLS7	84	Q4SHU3	116	Q64EU6
21	Q6PSS9	53	Q4RQ52	85	Q4SIJ4	117	Q6AX28
22	Q6PYX2	54	Q4RQ68	86	Q4SKS4	118	Q6B4U6
23	Q7T3U2	55	Q4RQ96	87	Q4SL08	119	Q6DCQ6
24	Q7ZX63	56	Q4RSI5	88	Q4SNM5	120	Q6QH77
25	Q7ZXT0	57	Q4RT71	89	Q4SPK6	121	Q6R8J2
26	Q7ZZT0	58	Q4RT87	90	Q4SRY6	122	Q6T683
27	Q804J3	59	Q4RU98	91	Q4STQ0	123	Q6W4W6
28	Q8AXK6	60	Q4RUP1	92	Q4SU37	124	Q7LZ69
29	Q8QGG9	61	Q4RUS3	93	Q4SVG4	125	Q7T026
30	Q90WZ3	62	Q4S2G3	94	Q4SW11	126	Q7T2X3
31	Q91008	63	Q4S2S0	95	Q4SXB6	127	Q7ZXL5
32	Q9DDR5	64	Q4S3C4	96	Q4SZV7	128	Q8AVH7

No.	Accession	No.	Accession	No.	Accession	No.	Accession
129	Q8AYF0	179	Q4TGF7	229	Q4T7A2	279	Q6NWK9
130	Q8AYF1	180	Q4S7X0	230	Q32SK9	280	Q58L93
131	Q8JHD5	181	Q2UZ94	231	Q4RX37	281	Q4SVF9
132	Q90819	182	Q4TC33	232	Q5M9B3	282	Q4S473
133	Q90WM2	183	Q2PP38	233	Q4RWE3	283	Q90YA5
134	Q90WZ2	184	O13128	234	Q4S8A4	284	Q6QNF2
135	Q90Z43	185	Q4SCB6	235	Q4S9S7	285	Q6GNA2
136	Q92070	186	Q5RIP6	236	Q4TB23	286	Q5FV82
137	Q98TH6	187	Q4TC32	237	Q800Y7	287	Q4SE79
138	Q98TH8	188	Q4RLR5	238	Q4RSS0	288	Q800E4
139	Q9DDR4	189	Q4T2F9	239	O57339	289	Q8JHV6
140	Q9I9K4	190	Q4RQW2	240	Q4S2B5	290	Q6DUJ6
141	Q9PUC8	191	Q2PP40	241	Q4S226	291	Q502R1
142	Q9YHF0	192	Q4T7A1	242	Q4S4V0	292	Q4TGH7
143	Q1LXE4	193	Q4SXH1	243	Q7SYT5	293	Q4SRM9
144	Q1LVQ0	194	Q4S5A3	244	Q4SZB3	294	Q4SMT5
145	Q2PP37	195	Q4SCL0	245	Q4S2X7	295	Q4S2C4
146	Q2PP39	196	Q4S163	246	O13149	296	Q4RZK1
147	Q9PU49	197	Q4RAY3	247	Q9W737	297	Q4RQW3
148	Q6KDZ1	198	O57658	248	O42372	298	Q4RQ69
149	Q4TI74	199	Q4RKP0	249	Q2PPJ1	299	Q4RMT7
150	Q4RJU5	200	Q2TJF5	250	Q90285	300	Q8UVF1
151	Q4VA78	201	Q4S5B2	251	Q8JHC9	301	Q4TB33
152	Q4TC24	202	Q4STE9	252	Q5TZI0	302	O57587
153	Q645M5	203	Q5RFU8	253	Q8JH43	303	Q92098
154	Q5RHM2	204	Q32SK8	254	Q91590	304	Q7T3H4
155	Q4RX38	205	Q2VWH3	255	Q501R2	305	Q8AW87
156	Q4SF52	206	Q4RB71	256	Q5BKN3	306	Q6PFT3
157	Q4T686	207	Q4RJ20	257	Q3YAA1	307	Q6IT10
158	Q4SQF4	208	Q5G872	258	Q6IQW7	308	Q5NJJ2
159	Q2Q1W5	209	Q4TBF2	259	Q7T024	309	Q4T5N1
160	O12973	210	Q4RC34	260	Q90Y55	310	Q4SUA3
161	Q4SG86	211	Q5NIW0	261	O42595	311	Q4RND7
162	Q4RXP0	212	Q505M8	262	Q8JHV8	312	Q4S367
163	Q4RSD5	213	Q4T313	263	Q4RVG6	313	Q90994
164	Q4RB72	214	Q3LTM5	264	Q4RNL1	314	Q6R8J3
165	Q5SNS5	215	Q4SUV4	265	Q4RG69	315	Q90YD2
166	Q2VU93	216	Q4RJ14	266	Q4RHV2	316	Q8UVF2
167	Q4SRX0	217	Q4SEY9	267	Q4SEH2	317	Q68EW5
168	Q4RWT1	218	Q7ZTJ2	268	Q4SK30	318	Q5NJL4
169	Q4RK78	219	Q5XHI6	269	Q4T3T2	319	Q503B9
170	Q4TAR0	220	Q90ZN3	270	Q75ZI2	320	Q4SHY3
171	Q9PTB2	221	Q4SQC4	271	Q4RVC8	321	Q4RW31
172	Q4S6G8	222	Q50LG7	272	Q4RFZ0	322	O42373
173	Q2VU94	223	Q4S6A6	273	Q8AXP0	323	Q9DEQ0
174	Q4RF33	224	Q4RU04	274	Q6NS01	324	O42374
175	Q4SEE4	225	Q9W7C5	275	Q4S3T6	325	Q8JHD0
176	Q5ZJR0	226	Q59I66	276	Q90995	326	Q8AW45
177	Q6P9I9	227	O42140	277	O12960	327	Q6PAE0
178	Q58EP9	228	Q4TF05	278	Q6R8J4	328	Q4S486

No.	Accession	No.	Accession
329	Q4RND7	379	Q5NJK0
330	Q4S367	380	Q4RTY8
331	Q90994	381	Q45H72
332	Q6R8J3	382	Q4RW33
333	Q90YD2	383	Q7T3B6
334	Q8UVF2	384	Q5RI06
335	Q68EW5	385	Q5NJJ3
336	Q5NJL4	386	Q58L92
337	Q503B9	387	Q4T6G4
338	Q4SHY3	388	Q4SWI4
339	Q4RW31	389	Q4SHN1
340	O42373	390	Q4S369
341	Q9DEQ0	391	Q9IA01
342	O42374	392	Q4T2F3
343	Q8JHD0	393	Q6NUU4
344	Q8AW45	394	Q5NJJ4
345	Q6PAE0	395	Q5BLE3
346	Q4S486	396	Q4SUA2
347	Q66IL7	397	Q8AXB7
348	Q58L96	398	Q6NY79
349	Q90Y56	399	Q7ZYV5
350	Q8JHV7	400	P79787
351	Q6DE79	401	Q6IT09
352	Q4RSS2	402	Q5NJJ0
353	Q4S290	403	Q4SMT3
354	Q4SB49	404	Q4S0R8
355	Q8UWG9	405	Q75ZI3
356	Q4SU28	406	Q5RH37
357	Q4T3P3	407	Q90W12
358	Q4T8V0	408	Q9PW89
359	Q4TC37	409	Q804R1
360	Q58L95	410	Q5XH36
361	Q7ZTG7	411	Q6NTV5
362	Q4RJM2	412	Q4F879
363	Q9W6V6		
364	Q8UVQ3		
365	Q90824		
366	Q90XG2		
367	Q91925		
368	Q6DIG3		
369	Q5M980		
370	Q4RXE9		
371	Q4RE15		
372	O57484		
373	Q90656		
374	Q8AYS9		
375	Q7ZXT4		
376	Q7ZVP3		
377	Q7SXF6		
378	Q5ZKF9		

Appendix D

Dataset C

No.	Accession	No.	Accession	No.	Accession	No.	Accession
1	OMIM:104640	33	OMIM:600275	65	OMIM:603745	97	Q4RTI5
2	OMIM:109770	34	OMIM:600276	66	OMIM:603818	98	Q4RTI6
3	OMIM:118450	35	OMIM:600310	67	OMIM:603897	99	Q4RZ32
4	OMIM:118850	36	OMIM:600345	68	OMIM:604110	100	Q4RZ38
5	OMIM:121050	37	OMIM:600441	69	OMIM:604210	101	Q4S0D5
6	OMIM:125310	38	OMIM:600493	70	OMIM:604234	102	Q4S1V8
7	OMIM:126150	39	OMIM:600514	71	OMIM:604264	103	Q4S2L6
8	OMIM:131210	40	OMIM:600521	72	OMIM:604265	104	Q4S758
9	OMIM:134797	41	OMIM:600565	73	OMIM:604266	105	Q4S940
10	OMIM:135821	42	OMIM:600566	74	OMIM:604267	106	Q4S9W4
11	OMIM:142445	43	OMIM:600567	75	OMIM:604268	107	Q4SA73
12	OMIM:152780	44	OMIM:600582	76	OMIM:604269	108	Q4SC13
13	OMIM:152790	45	OMIM:600826	77	OMIM:604270	109	Q4SCB7
14	OMIM:154700	46	OMIM:601456	78	OMIM:604308	110	Q4SEB0
15	OMIM:158371	47	OMIM:601533	79	OMIM:604580	111	Q4SFW8
16	OMIM:172870	48	OMIM:601920	80	OMIM:604609	112	Q4SGX5
17	OMIM:176290	49	OMIM:602061	81	OMIM:604633	113	Q4SI87
18	OMIM:182212	50	OMIM:602281	82	OMIM:604710	114	Q4SIJ3
19	OMIM:187395	51	OMIM:602319	83	OMIM:605007	115	Q4SJN6
20	OMIM:187500	52	OMIM:602320	84	OMIM:605008	116	Q4SK80
21	OMIM:188040	53	OMIM:602570	85	OMIM:605009	117	Q4SM61
22	OMIM:188062	54	OMIM:602713	86	OMIM:605102	118	Q4SN22
23	OMIM:190198	55	OMIM:603130	87	OMIM:605185	119	Q4SP98
24	OMIM:227600	56	OMIM:603421	88	OMIM:605194	120	Q4SQ68
25	OMIM:300239	57	OMIM:603639	89	OMIM:605227	121	Q4SQA5
26	OMIM:306900	58	OMIM:603742	90	OMIM:605441	122	Q4SRW7
27	OMIM:605533	59	OMIM:606276	91	OMIM:607491	123	Q4STC1
28	OMIM:605734	60	OMIM:606582	92	OMIM:607661	124	Q4SXD4
29	OMIM:606018	61	OMIM:607114	93	OMIM:607873	125	Q4SZZ8
30	OMIM:606100	62	OMIM:607170	94	OMIM:608529	126	Q4T2D2
31	OMIM:606101	63	OMIM:607171	95	Q4RQ03	127	Q4T2J4
32	OMIM:606217	64	OMIM:607299	96	Q4RQ74	128	Q4T392

No.	Accession	No.	Accession	No.	Accession	No.	Accession
129	Q4T3X9	179	Q8AXK7	229	IPR004457	279	IPR000446
130	Q4T4X0	180	Q8AXM6	230	IPR004470	280	IPR000447
131	Q4T785	181	Q4RSA9	231	IPR005018	281	IPR000463
132	Q5NJJ1	182	Q4RXZ7	232	IPR005468	282	IPR000469
133	Q5FW21	183	Q4SNE6	233	IPR005469	283	IPR000472
134	Q4RGL7	184	Q4T3Z6	234	IPR006210	284	IPR000476
135	Q804W9	185	Q56VR3	235	IPR006586	285	IPR000479
136	O73920	186	Q58L94	236	IPR006952	286	IPR000491
137	Q5FVY5	187	Q7ZWL5	237	IPR007943	287	IPR000523
138	Q4T6A4	188	Q6P7I9	238	IPR008131	288	IPR000532
139	Q4SKD1	189	Q5ZJH4	239	IPR009030	289	IPR000539
140	Q4SDW5	190	P79941	240	IPR011170	290	IPR000562
141	Q4RSL6	191	O73809	241	IPR011359	291	IPR000571
142	Q3YAA0	192	Q766V2	242	IPR012152	292	IPR000586
143	Q6P1V9	193	Q4RQ94	243	IPR013050	293	IPR000638
144	Q90XG4	194	Q2T9I2	244	IPR013309	294	IPR000647
145	Q58EG1	195	Q4FAI8	245	IPR007951	295	IPR000654
146	Q5M7J5	196	Q91902	246	IPR000021	296	IPR000655
147	Q6RUW2	197	Q92071	247	IPR000053	297	IPR000686
148	Q6PAG2	198	O93575	248	IPR000057	298	IPR000694
149	Q7SZG1	199	Q90YK1	249	IPR000062	299	IPR000712
150	Q69GM1	200	Q7SY86	250	IPR000072	300	IPR000716
151	Q5RI69	201	Q6PPB4	251	IPR000098	301	IPR000753
152	Q5M8Y0	202	Q4SHC2	252	IPR000104	302	IPR000762
153	Q504H3	203	Q4SFQ0	253	IPR000136	303	IPR000770
154	Q4RU01	204	Q4S162	254	IPR000147	304	IPR000773
155	Q4RP66	205	IPR000020	255	IPR000148	305	IPR000779
156	Q4RLD6	206	IPR000083	256	IPR000151	306	IPR000782
157	Q8JFZ4	207	IPR000421	257	IPR000174	307	IPR000820
158	Q4SP38	208	IPR000742	258	IPR000182	308	IPR000827
159	Q4RSM9	209	IPR000800	259	IPR000186	309	IPR000837
160	Q5TZK8	210	IPR001881	260	IPR000187	310	IPR000859
161	Q4KMI9	211	IPR002049	261	IPR000190	311	IPR000867
162	Q68EY0	212	IPR006209	262	IPR000197	312	IPR000870
163	Q4T6Q3	213	IPR007803	263	IPR000222	313	IPR000883
164	Q4RV65	214	IPR010901	264	IPR000226	314	IPR000890
165	Q4RJ58	215	IPR011203	265	IPR000242	315	IPR000898
166	Q4RQ70	216	IPR013032	266	IPR000248	316	IPR000907
167	Q6P4X1	217	IPR013091	267	IPR000249	317	IPR000910
168	Q5RG03	218	IPR013111	268	IPR000269	318	IPR000932
169	Q503G2	219	IPR000033	269	IPR000301	319	IPR000962
170	Q4RVD0	220	IPR000152	270	IPR000313	320	IPR000975
171	Q4RHF2	221	IPR000494	271	IPR000327	321	IPR000976
172	Q4T0S1	222	IPR001336	272	IPR000331	322	IPR000987
173	Q9DFE9	223	IPR001438	273	IPR000367	323	IPR001019
174	Q7ZXI2	224	IPR001666	274	IPR000374	324	IPR001023
175	Q68EK6	225	IPR001740	275	IPR000381	325	IPR001089
176	Q4VBJ0	226	IPR002007	276	IPR000387	326	IPR001090
177	Q8JH44	227	IPR002172	277	IPR000403	327	IPR001106
178	Q504J5	228	IPR002610	278	IPR000405	328	IPR013548

No.	Accession	No.	Accession	No.	Accession	No.	Accession
329	IPR000493	379	IPR001245	428	IPR002022	478	IPR002112
330	IPR000523	380	IPR001246	429	IPR002059	479	IPR002121
331	IPR000532	381	IPR001251	430	IPR002072	480	IPR002153
332	IPR000539	382	IPR001308	431	IPR002074	481	IPR002154
333	IPR000562	383	IPR001318	432	IPR002083	482	IPR002160
334	IPR000571	384	IPR001321	433	IPR002087	483	IPR002161
335	IPR000586	385	IPR001323	434	IPR002100	484	IPR002183
336	IPR000638	386	IPR001337	435	IPR002101	485	IPR002188
337	IPR000647	387	IPR001343	436	IPR002112	486	IPR002209
338	IPR000654	388	IPR001355	437	IPR002121	487	IPR002212
339	IPR000655	389	IPR001368	438	IPR002153	488	IPR002259
340	IPR000686	390	IPR001377	439	IPR002154	489	IPR002277
341	IPR000694	391	IPR001400	440	IPR002160	490	IPR002285
342	IPR000712	392	IPR001402	441	IPR002161	491	IPR002348
343	IPR000716	393	IPR001408	442	IPR002183	492	IPR002353
344	IPR000753	394	IPR001422	443	IPR002188	493	IPR002354
345	IPR000762	395	IPR001426	444	IPR002209	494	IPR002393
346	IPR000770	396	IPR001476	445	IPR002212	495	IPR002400
347	IPR000773	397	IPR001477	446	IPR002259	496	IPR002405
348	IPR000779	398	IPR001506	447	IPR002277	497	IPR002418
349	IPR000782	399	IPR001512	448	IPR002285	498	IPR002423
350	IPR000820	400	IPR001545	449	IPR002348	499	IPR002446
351	IPR000827	401	IPR001555	450	IPR002353	500	IPR002473
352	IPR000837	402	IPR001606	451	IPR002354	501	IPR002475
353	IPR000859	403	IPR001627	452	IPR002393	502	IPR002491
354	IPR000867	404	IPR001632	453	IPR001824	503	IPR002554
355	IPR000870	405	IPR001672	454	IPR001839	504	IPR002557
356	IPR000883	406	IPR001678	455	IPR001844	505	IPR002587
357	IPR000890	407	IPR001690	456	IPR001846	506	IPR002633
358	IPR000898	408	IPR001770	457	IPR001852	507	IPR002634
359	IPR000907	409	IPR001806	458	IPR001856	508	IPR002643
360	IPR000910	410	IPR001811	459	IPR001858	509	IPR002644
361	IPR000932	411	IPR001824	460	IPR001877	510	IPR002649
362	IPR000962	412	IPR001824	461	IPR001885	511	IPR002661
363	IPR000975	413	IPR001839	462	IPR001893	512	IPR002666
364	IPR000976	414	IPR001844	463	IPR001904	513	IPR002714
365	IPR000987	415	IPR001846	464	IPR001929	514	IPR002770
366	IPR001019	416	IPR001852	465	IPR001932	515	IPR002836
367	IPR001023	417	IPR001856	466	IPR001955	516	IPR002856
368	IPR001089	418	IPR001858	467	IPR001983	517	IPR002869
369	IPR001090	419	IPR001877	468	IPR002003	518	IPR002880
370	IPR001106	420	IPR001885	469	IPR002011	519	IPR002963
371	IPR001111	421	IPR001893	470	IPR002022	520	IPR002975
372	IPR001116	422	IPR001904	471	IPR002059	521	IPR002976
373	IPR001132	423	IPR001929	472	IPR002072	522	IPR003012
374	IPR001181	424	IPR001932	473	IPR002074	523	IPR003014
375	IPR001184	425	IPR001955	474	IPR002083	524	IPR003064
376	IPR001192	426	IPR001983	475	IPR002087	525	IPR003085
377	IPR001214	427	IPR002003	476	IPR002100	526	IPR003087
378	IPR001217	428	IPR002011	477	IPR002101	527	IPR003093

No.	Accession	No.	Accession	No.	Accession
528	IPR003103	578	IPR003623	628	IPR003763
529	IPR003108	579	IPR003629	629	IPR003813
530	IPR003127	580	IPR003630	630	IPR003829
531	IPR003193	581	IPR003653	631	IPR003881
532	IPR003206	582	IPR003654	632	IPR003905
533	IPR003207	583	IPR003659	633	IPR003906
534	IPR003208	584	IPR003670	634	IPR003907
535	IPR003234	585	IPR003718	635	IPR003908
536	IPR003235	586	IPR003763	636	IPR003911
537	IPR003284	587	IPR003813	637	IPR003914
538	IPR003288	588	IPR003829	638	IPR003932
539	IPR003293	589	IPR003881	639	IPR003933
540	IPR003294	590	IPR003905	640	IPR003934
541	IPR003295	591	IPR003906	641	IPR003936
542	IPR003296	592	IPR003907	642	IPR003939
543	IPR003297	593	IPR003908	643	IPR003940
544	IPR003302	594	IPR003911	644	IPR003941
545	IPR003311	595	IPR003914	645	IPR003942
546	IPR003327	596	IPR003932	646	IPR003952
547	IPR003368	597	IPR003933	647	IPR003953
548	IPR003392	598	IPR003934	648	IPR003966
549	IPR003398	599	IPR003936	649	IPR003985
550	IPR003438	600	IPR003939	650	IPR004000
551	IPR003454	601	IPR003940	651	IPR004001
552	IPR003460	602	IPR003941	652	IPR004045
553	IPR003463	603	IPR003942	653	IPR004046
554	IPR003477	604	IPR003952	654	IPR004061
555	IPR003502	605	IPR003953	655	IPR004062
556	IPR003503	606	IPR003966	656	IPR004063
557	IPR003504	607	IPR003985	657	IPR004064
558	IPR003505	608	IPR004000	658	IPR004065
559	IPR003527	609	IPR004001	659	IPR004066
560	IPR003528	610	IPR004045	660	IPR008996
561	IPR003529	611	IPR004046		
562	IPR003530	612	IPR004061		
563	IPR003531	613	IPR004062		
564	IPR003532	614	IPR004063		
565	IPR003538	615	IPR004064		
566	IPR003542	616	IPR004065		
567	IPR003555	617	IPR004066		
568	IPR003560	618	IPR004074		
569	IPR003570	619	IPR004076		
570	IPR003573	620	IPR004077		
571	IPR003574	621	IPR003629		
572	IPR003577	622	IPR003630		
573	IPR003595	623	IPR003653		
574	IPR003598	624	IPR003654		
575	IPR003608	625	IPR003659		
576	IPR003619	626	IPR003670		
577	IPR003620	627	IPR003718		

No.	Accession	No.	Accession	No.	Accession	No.	Accession
1	IPR000042	51	IPR012213	101	P30443	151	P30443
2	IPR000082	52	IPR012214	102	P16209	152	P16209
3	IPR000155	53	IPR012858	103	P16211	153	P16211
4	IPR000246	54	IPR013969	104	P30515	154	P30515
5	IPR000526	55	NY-REN-27 (MTDB)	105	P30376	155	P30376
6	IPR000621	56	PDOC00284	106	P01892	156	P01892
7	IPR000715	57	PDOC00204	107	P16210	157	P16210
8	IPR000905	58	PDOC00210	108	P30377	158	P30377
9	IPR001038	59	IPR006844	109	P04439	159	P04439
10	IPR001329	60	IPR007235	110	P13748	160	P13748
11	IPR001439	61	IPR007267	111	P30378	161	P30378
12	IPR001503	62	IPR007676	112	P13749	162	P13749
13	IPR001675	63	IPR007754	113	P13746	163	P13746
14	IPR001968	64	IPR007906	114	P30447	164	P30447
15	IPR002122	65	IPR008083	115	P05534	165	P05534
16	IPR002202	66	IPR008363	116	P18462	166	P18462
17	IPR002213	67	IPR008364	117	P30450	167	P30450
18	IPR002249	68	IPR008368	118	P30512	168	P30512
19	IPR002280	69	IPR008369	119	P16188	169	P16188
20	IPR002443	70	IPR008370	120	P16189	170	P16189
21	IPR002444	71	IPR008371	121	P10314	171	P10314
22	IPR002445	72	IPR008372	122	P16190	172	P16190
23	IPR002640	73	IPR008647	123	P30453	173	P30453
24	IPR002659	74	IPR008710	124	P30455	174	P30455
25	IPR002685	75	IPR008814	125	P30456	175	P30456
26	IPR002968	76	IPR008820	126	P30457		
27	IPR003038	77	IPR008821	127	P01891		
28	IPR003342	78	IPR008853	128	P10316		
29	IPR003378	79	IPR009138	129	P30459		
30	IPR003406	80	IPR009151	130	Q09160		
31	IPR003407	81	IPR009168	131	P30379		
32	IPR003492	82	IPR009294	132	P13750		
33	IPR003674	83	IPR009448	133	P30516		
34	IPR003919	84	IPR009684	134	P30380		
35	IPR003961	85	IPR010555	135	P13751		
36	IPR003962	86	IPR010580	136	P30381		
37	IPR004139	87	IPR011143	137	P30382		
38	IPR004276	88	IPR012163	138	P01889		
39	IPR004816	89	IPR012209	139	P30460		
40	IPR004856	90	IPR012210	140	P30461		
41	IPR005013	91	IPR012211	141	P30462		
42	IPR005421	92	IPR012212	142	P30464		
43	IPR005422	93	PDOC00559	143	P30466		
44	IPR005423	94	PDOC00281	144	P03989		
45	IPR005429	95	PDOC00623	145	P30685		
46	IPR005817	96	PDOC00234	146	P18463		
47	IPR005951	97	PDOC00280	147	Q95365		
48	IPR006603	98	PDOC00209	148	P30475		
49	IPR006706	99	PDOC00248	149	Q04826		
50	IPR006813	100	P30375	150	P30375		

No.	Accession	No.	Accession
1	P39940	43	X54942
2	P40559	44	L31801
3	P50077	45	U04953
4	P39524		
5	P48510		
6	Q12518		
7	Q05785		
8	P38111		
9	P32486		
10	Q12446		
11	P40477		
12	P39105		
13	Q08108		
14	P46957		
15	P53064		
16	Q04183		
17	P38334		
18	Q03780		
19	Q05518		
20	P38809		
21	P40497		
22	P47068		
23	P40361		
24	P53947		
25	P49686		
26	P25040		
27	P53309		
28	P38856		
29	X55362		
30	M61832		
31	D13639		
32	T51288		
33	T70920		
34	U02020		
35	R61502		
36	H73758		
37	H17434		
38	M69199		
39	H55916		
40	Z49199		
41	T57468		
42	R23889		

Appendix E

O-Glucose Dataset

Table 11: EGF-Like Repeats with Known O-glucose Modifications [56]

No.	(Onco)Peptide	Accession ID	Glycosylation Site
1	Hu Factor IX	P00740	99
2	Hu Factor VII	P08709	112
3	Mouse Notch1-EGFL2	Q01705	65
4	Mouse Notch1-EGFL4	Q01705	146
5	Mouse Notch1-EGFL10	Q01705	378
6	Mouse Notch1-EGFL14	Q01705	534
7	Mouse Notch1-EGFL16	Q01705	609
8	Mouse Notch1-EGFL17	Q01705	647
9	Mouse Notch1-EGFL19	Q01705	722
10	Mouse Notch1-EGFL20	Q01705	759
11	Mouse Notch1-EGFL21	Q01705	797
12	Mouse Notch1-EGFL25	Q01705	951
13	Mouse Notch1-EGFL27	Q01705	1027
15	Mouse Notch1-EGFL28	Q01705	1065
16	Mouse Notch1-EGFL33	Q01705	1273
17	Mouse Notch1-EGFL36	Q01705	1394

Appendix F

O-Fucose Dataset

Table 12: EGF-Like Repeats with Known O-fucose Modifications [56]

No.	(Onco)Peptide	Accession ID	Glycosylation Site
1	Hu Factor IX	P00740	107
2	Hu Factor VII	P08709	120
3	Hu Factor XII	P00748	109
4	Hu uPA	P00749	38
5	Hu tPA	P00750	96
6	Hu Cripto	P13385	88
7	Mouse Notch1-EGFL2	Q01705	73
8	Mouse Notch1-EGFL3	Q01705	56
9	Mouse Notch1-EGFL5	Q01705	194
10	Mouse Notch1-EGFL12	Q01705	466
11	Mouse Notch1-EGFL20	Q01705	767
12	Mouse Notch1-EGFL21	Q01705	805
13	Mouse Notch1-EGFL23	Q01705	883
14	Mouse Notch1-EGFL24	Q01705	921
15	Mouse Notch1-EGFL26	Q01705	997
16	Mouse Notch1-EGFL27	Q01705	1035
17	Mouse Notch1-EGFL35	Q01705	1362
18	Mouse Notch1-EGFL36	Q01705	1402