TpPred: AN ONLINE TOOL FOR HIERARCHICAL PREDICTION OF TRANSPORT PROTEINS USING CLUSTER OF NEURAL NETWORKS AND SEQUENCE DERIVED FEATURES

By

Sankalp Jain-061522

THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Bachelor of Technology

IN

BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN - 173215, HIMACHAL PRADESH, INDIA

May 2010

Dr. Pradeep Kumar Naik
Assistant Professor
Biotechnology & Bioinformatics

Jaypee University of
Information Technology
Waknaghat-173215
Solan, Himachal Pradesh
Phone No.: 91-1792-239227
Fax No.: 91-1792-245362

CERTIFICATE

This is to certify that the thesis entitled "TpPred: an online tool for hierarchical prediction of transport proteins using cluster of neural networks and sequence derived features" submitted by **Sankalp Jain** to the Jaypee University of Information Technology, Waknaghat in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Bioinformatics** is a record of bona fide research work carried out by him under my supervision and guidance and no part of this work has been submitted for any other degree or diploma.

(Dr. P.K. Naik)

DECLARATION

I hereby declare that the work presented in this thesis has been carried out by me under the supervision of Dr. Pradeep Kumar Naik, Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan-173215, Himachal Pradesh, and has not been submitted for a degree or diploma of any other university. All assistance and help received during the course of the investigation has been duly acknowledged.

Sankalp Jain

ACKNOWLEDGEMENT

"To speak gratitude is courteous and pleasant, to enact gratitude is generous and noble, but to live gratitude is to touch Heaven"

I would like to express my heartfelt gratitude to all those who have contributed directly or indirectly towards obtaining my bachelor degree. I am highly indebted to my esteemed supervisor, Dr. Pradeep Kumar Naik, who has guided me through thick and thin. Many people have contributed to this project in a variety of ways over the past few months. To the individuals who have helped me, I again express my appreciation. I express my special thanks to Prasad Bajaj & Piyush Ranjan for guiding me in development of the server edition of the tool. I also acknowledge the many helpful comments Received from my Teachers of the bioinformatics department. I am indebted to all those who provided reviews & suggestion for improving the result and the topics covered in my project, and I extended my apologies to anyone I may have failed to mention.

Sankalp Jain

Contents:

List of Figures:

List of Tables:

List of abbreviations

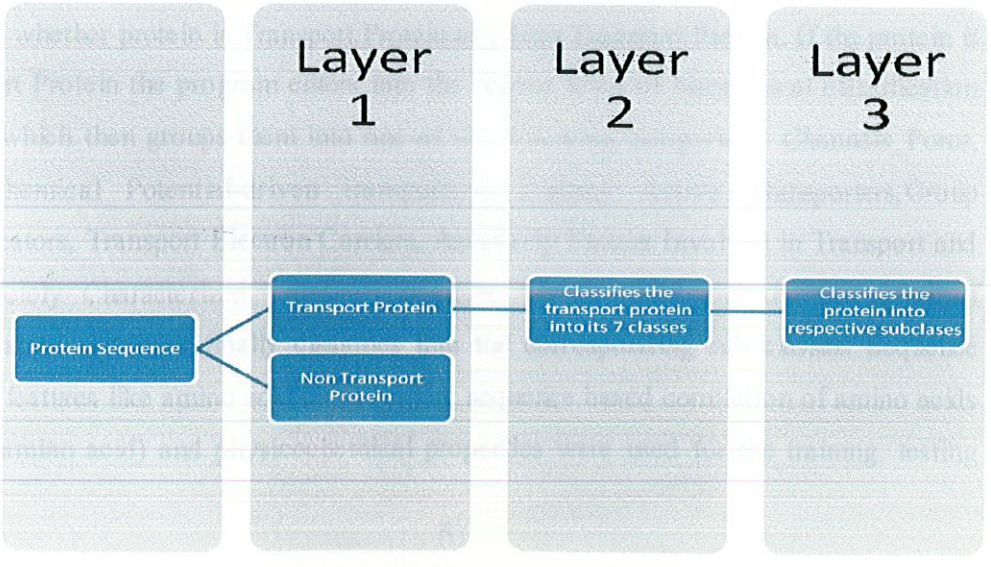
ANN	Artificial Neural Network
pI	Iso-electric Point
TCDB	Transport Classification Database
PEPSTAT	Protein Statistics (which calculates physicochemical properties)
SANN	Statistica Automated Neural Network
ANS	Automated Network Search
CNS	Custom Network Search
AA comp	Amino Acid Composition
PseAA	Pseudo Amino Acid Composition

TpPred : A tool for the prediction of Transport Proteins

.....ACDEGVGCMAAAG.....
(Protein sequence)

Sequence derived features:
1. Amino acid composition
2. Pseudo amino acid composition
3. Physicochemical properties
4. Amino acid composition + Pseudo amino acid composition
5. Amino acid composition + Physicochemical properties
6. Pseudo amino acid composition + Physicochemical properties
7. Amino acid composition + Pseudo amino acid composition + Physicochemical properties

Cluster of Neural Network (CNN)



ABSTRACT

Motivation: Transport proteins are difficult to understand by biological experiments due to the difficulty in obtaining crystals suitable for X-ray diffraction. Therefore, the use of computational techniques is a powerful approach to annotate the function of proteins. Thus a faster means of annotation would be to match them with the already annotated sequences using sequence based similarity search method like BLAST. It is a discrete method of calculating the similarity between protein sequences simply by measuring the number of matches and mismatches. However, the function of a protein is not only depends on its primary sequence but also very much depends on how the protein folds into 3D structure which in turn also depends on the hydrophobicity and hydrophilicity properties of the proteins. Therefore it is needed to capture sequence order information, short term and long term interactions between amino acids in a protein sequence as well as to capture proportion of hydrophobicity and hydrophilicity properties of the proteins in order to correctly annotate the raw protein sequence. Therefore, we are motivated to develop an online prediction server for predicting the transport proteins from sequence derived features. These methods achieved good prediction accuracies and could nicely complement experimental approaches for identification of transport proteins. The prediction methods are unique in the sense that they do not require homologous protein sequences.

Result: The tool would be consisting of 3 level of Classification. First layer classification includes whether protein is Transport Protein or a Non Transport Protein. If the protein is Transport Protein the program enters into the second level of hierarchical classification system which then groups them into one of the following categories - Channels_Pores, Electrochemical Potential-driven transporters, Primary Active Transporters, Group Translocators, Transport Electron Carriers, Accessory Factors Involved in Transport and Incompletely Characterized Transport Systems. In the third and the last level of classification system it finally classifies into the corresponding sub classes. Sequence derived features like amino acid composition; sequence based correlation of amino acids (pseudoamino acid) and physicochemical properties were used for the training, testing

and validation of the tool. Our tool is robust and successfully classifies the novel protein sequence into transport protein, then into its major class and finally predicts specific function. The performance accuracy of our tool at 1st layer is 88.00%, at 2nd layer is 84.21% and at 3rd layer is 82.09%. Using Jackknifing validation technique the performance accuracy of our tool is 81.44% at 1st layer, 68.46% at 2nd layer and 70.04% at the 3rd layer.

Availability: The TpPred tool is available for free use to non commercial users and can be downloaded to be used in-house as a standalone server from following links.
http://www.juit.ac.in/

CHAPTER 1

INTRODUCTION

Transport proteins are proteins within the membranes of cells that transport and help in the movement of substances such as ions, small molecules, or macromolecules (such as another protein) across the membrane or within the cell. These transport proteins are often globular proteins. They are generally tightly packed with polar side groups on the outside to enhance their solubility in water. They typically have non-polar side groups folded to the inside to keep water from getting in and unfolding them. Serum albumin is one example. It transports water-insoluble lipids in the bloodstream. According to function, there are two different types of transport proteins:

1. Those that carry molecules to "distant" locations (within a cell or an organism),
2. Those that serve as gateways, carrying molecules across otherwise impermeable membranes.

1.1 About Transport Proteins

Transport protein can refer to:

- Membrane transport protein
- Vesicular transport protein
- Water-soluble carriers of small molecules

Membrane Transport Protein

A membrane transport protein (or simply transporter) is a protein involved in the movement of ions, small molecules, or macromolecules, such as another protein across a biological membrane. Transport proteins are integral membrane proteins; that is they exist within and span the membrane across which they transport substances. The proteins may assist in the movement of substances by facilitated diffusion or active transport.

The mechanism of action

The mechanism of action of these proteins is known as carrier-mediated transport. There are two forms of carrier- mediated transport, active transport and facilitated diffusion.

Facilitated diffusion

Facilitated diffusion (or facilitated transport) is a process of diffusion, a form of passive transport facilitated by transport proteins. Facilitated diffusion is the spontaneous passage of molecules or ions across a biological membrane passing through specific transmembrane transport proteins. The facilitated diffusion may occur either across biological membranes or through aqueous compartments of an organism (Figure 1).

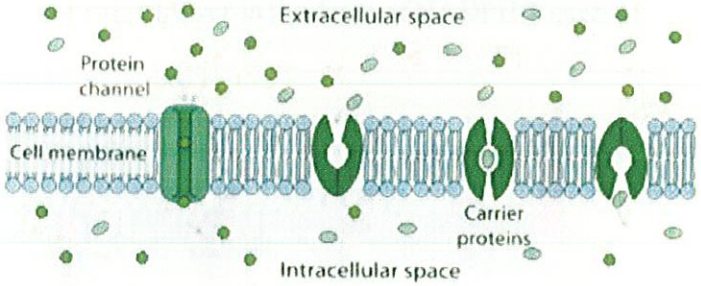


Figure 1 Facilitated diffusion.

Polar molecules and charged ions are dissolved in water but they cannot diffuse freely across cell membranes due to the hydrophobic nature of the phospholipids that make up the lipid bilayers. Only small nonpolar molecules, such as oxygen can diffuse easily across the membrane. All polar molecules are transported across membranes by proteins that form transmembrane channels. These channels are gated so they can open and close, thus regulating the flow of ions or small polar molecules. Larger molecules are transported by transmembrane carrier proteins, such as permeases that change their conformation as the molecules are carried through, for example glucose or amino acids. Non-polar molecules, such as retinol or fatty acids are poorly soluble in water. They are transported through aqueous compartments of cells or through extracellular space by water-soluble carriers as retinol binding protein. The metabolites are not changed because no energy is required for facilitated diffusion. Only permease changes its shape in order to transport the metabolites. The form of transport through cell membrane which modifies its metabolites is the group translocation transportation.

Active transport

Transport proteins are also used in active transport, which by definition does require an energy output. Most of the enzymes that perform this type of transport are transmembrane ATPases. A primary ATPase universal to all cellular life is the sodium-potassium pump, which helps maintain the cell potential. Other sources of energy for Primary active transport are redox energy and photon energy (light). An example of primary active transport using Redox energy is the mitochondrial electron transport chain that uses the reduction energy of NADH to move protons across the inner mitochondrial membrane against their concentration gradient. An example of primary active transport using light energy are the proteins involved in photosynthesis that use the energy of photons to create a proton gradient across the thylakoid membrane and also to create reduction power in the form of NADPH(Figure 2).

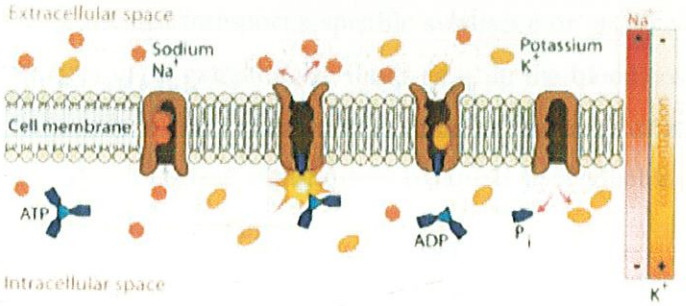


Figure 2 Active transport.

Chemiosmotic transport utilizes electrochemical gradients to drive transport. As the creation and maintenance of chemiosrnotic gradients require energy input from the cell, this is a form of active transport. Prokaryotes typically use hydrogen ions as the driving force for chemiosmotic transport, while eukaryotes typically use sodium ions. A symporter coporter transports a chemical in the same direction as the electrochemical gradient, while an antiporter moves the target chemical in a direction opposite to the gradient.

Vesicular transport protein

A vesicular transport protein is a membrane protein which use vesicles to move the contents of the cell(Figure 3). Examples include: Archain and Clathrin.

Figure 3 Visicular transport protein.

Water soluble carrier of small molecules

Carrier proteins are proteins that transport a specific substance or group of substances through intracellular compartments or in extracellular fluids (e.g. in the blood) or else across the cell membrane. Some of the carriers are water-soluble proteins that may or may not interact with biological membranes, such as some transporters of small hydrophobic molecules, whereas others are integral transmembrane proteins.

Carrier proteins transport substances out of or into the cell by facilitated diffusion and active transport. Each carrier protein is designed to recognize only one substance or one group of very similar substances.

The molecule or ion to be transported (the substrate) must first bind at a binding site at the carrier molecule, with a certain binding affinity. Following binding, and while the binding site is facing, say, outwards, the carrier will capture or occlude (take in and retain) the substrate within its molecular structure and cause an internal translocation, so that it now faces the other side of the membrane. The substrate is finally released at that site, according to its binding affinity there. All steps are reversible (Figure 4).

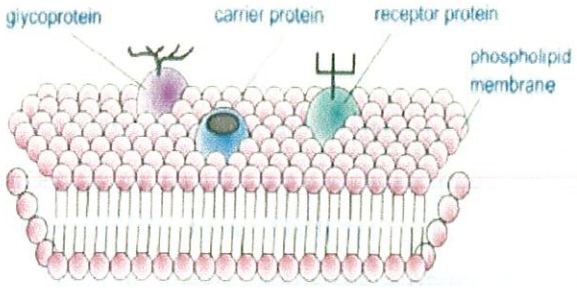


Figure 4 carried molecules embedded in phopholipid bilayer, helps in transportation of molecules across the membrane.

For example:

1. Diffusion of sugars, amino acid, nucleoside.
2. Uptake of glucose.
3. Transportation of salts, glucose, and amino acids

1.2 Importance of Transport Proteins

The transport proteins have an important function of transporting or storing some chemical compounds and ions. In order for vast numbers of multicellular organisms to exist, they must have a system for delivering oxygen to all their cells, especially those cells that are not in direct contact with the organism's external environment Some of well known examples include that of : cytochrome C - electron transport; hemoglobin and myoglobin - oxygen transport; albumin - fatty acid transport in the blood stream etc.

Hemoglobin is an example of an oxygen-transport protein and is a part of these oxygen delivery systems. A single human hemoglobin molecule consists of four polypeptide chains. Each of these chains contains a tightly bound prosthetic group called heme. A prosthetic group is a small organic molecule (non-amino acid) that is bound tightly to a protein. At the heart of each **heme group** is a tightly bound iron atom, to which oxygen binds. The function of hemoglobin is to bind oxygen in the oxygen-rich environment of the lungs, then to release that oxygen to oxygen-poor tissues elsewhere. The polypeptide chains are wrapped around the heme groups in such a way that the affinity between the iron and oxygen is strong enough for hemoglobin to bind oxygen in the lungs, but the resulting bond is weak enough such that hemoglobin will release the oxygen when it encounters organs or tissues that need oxygen.

Hemoglobin also performs the complementary function of accepting carbon dioxide from the peripheral tissues and releasing it in the lungs. Cell membranes are impermeable to charged and polar molecules, meaning that these molecules cannot cross them spontaneously. Some transport proteins are intrinsic to cell membranes and facilitate the transport of polar molecules across the membranes. Each cell of the human body needs **glucose** , a very polar molecule, and human beings have five different glucose transport proteins (known as GLUT1 through GLUT5) that all serve a similar function: They carry glucose molecules across membranes and into cells. Without these transport proteins, the rate of glucose entry into cells would be very low indeed. Other membrane-linked transport proteins carry other molecules across membranes, including amino acids, ions, and vitamins.

1.3 Classification of Transport Proteins

There is a comprehensive classification system fur membrane transport proteins known as the Transport Classification (TC) system The TC system is analogous to the Enzyme Commission (EC) system for classification of enzymes, except that it incorporates both functional and phylogenetic information Descriptions, TC numbers, and examples of over
500 families of transport proteins are provided. Transport systems are classified on the basis of five criteria, and each of these criteria corresponds to one of the five numbers or letters within the TC# for a particular type of transporter. Thus TC # normally has five components as follows: V.W.X.Y.Z. V (a number) corresponds to the transporter class (i.e., channel, carrier (porter), primary active transporter or group translocator); W (a lettter) corresponds to the transporter subclass which in the case of primary active transporters refers to the energy source used to drive transport; X (a number) corresponds to the transporter family (sometimes actually a superfarnily); Y (a number) corresponds to the subfamily in which a transporter is found, and Z corresponds to the substrate or range of substrates transported. Any two transport systems in the same subfamily ofa transporter family that transport the same substrate(s) are given the same TC#, regardless ofwhether they are orthologues (e.g., arose in distinct organisms by speciation) or paralogues (e.g., arose within a single organism by gene duplication). Sequenced horrnlogues of unknown function are not normally assigned a TC# unless they represent a unique (sub)furnily or are from an unrepresented organismal kingdom. If multiple dissimilar

subunits are present, they are numbered SI, S2, S3 Sn, Classification categories 8 and 9 are reserved for accessory transport proteins and incompletely characterized (families of) transporters, respectively.

According to TCDB we obtained the following classification system at level 1 (Figure 5):

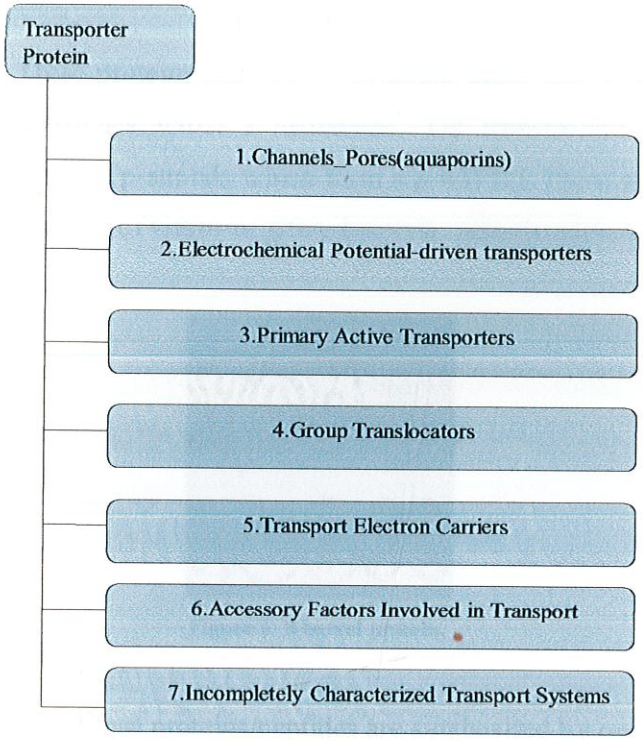


Figure 5 Classificaton chart.

1. Channel/IPores

Channel-type facilitators. Proteins in this category have transmembrane channels which consist largely of a-helical or b-strand-type spanners. Transport systems of this type catalyze facilitated diffusion (by an energy-independent process) by passage through a transmembrane aqueous pore or channel without evidence for a carrier-mediated mechanism. They do not exhibit stereospecificity but may be specific for a particular molecular species or class of molecules.

These include:

I.A. a-Type channels. Transmembrane channel proteins of this class are ubiquitously found in the membranes of all types of organisms from bacteria to higher eukaryotes. These transporters

usually catalyze movement of solutes by an energy-independent process by passage through a transmembrane aqueous pore or channel without evidence for a carriermediated mechanism. These channel proteins usually consist largely ofa-helical spanners, although p-strands may be present and may even contribute to the channel. Outer membrane porin-type channel proteins are exc luded from this class and are instead included in class I. B.

I.B. β-Barrel porins. These proteins form transmembrane pores that usually allow the energy independent passage of solutes across a membrane. The transmembrane portions of these proteins consist exclusively of p-strands which form a p-b31Tel. These porin-type proteins are found in the outer membranes of Gram-negative bacteria, mitochondria and plastids(Figure 6).

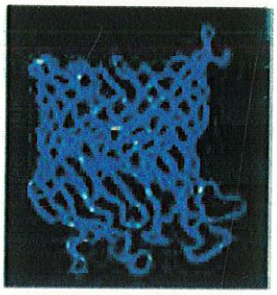


Figure 6 -B barrel protein.

1.C. Pore-forming toxins. These proteins/peptides are synthesized by one cell and secreted for insertion into the membrane of another cell where they form transmembrane pores. They may exert their toxic effects by allowing the free flow of electrolytes and other small molecules across the membrane, or they may allow entry into the target cell cytoplasm of a toxin protein that ultimately kills the cell. Both protein (large) and ribosomally synthesized peptide (small) toxins are included in this category.

1.D. Non-ribosomally synthesized channels. These molecules, often chains of **L-** and D-amino acids as well as other small molecular building blocks such as hydroxy acids (ie., lactate), form oligomeric transmembrane ion channels. Voltage may induce channel formation by promoting assembly of the oligomeric transmembrane pore-forming structure.
These depsipeptides are often made by bacteria and fungi as agents of biological warfare.

Other substances, completely lacking amino acids, are also capable of channel-formation.

1.E. Holins. Holins consist of about forty distinct families of proteins that exhibit common structural and functional characteristics but which do not exhibit statistically significant sequence similarity between members of distinct families. They are encoded within the genomes of Gram-positive and Gram-negative bacteria as well as those of the bacteriophage of these organisms. Their primary function appears to be transport of murein hydrolases across the cytoplasmic membrane to the cell wall where these enzymes hydrolyze the cell wall polymer as a prelude to cell lysis. When chromosomally encoded, these enzymes are therefore autolysins. Holins may also facilitate leakage of electrolytes and nutrients from the cellcytoplasm, thereby promoting cell death.

2. Electrochemical Potential-driven transporters

Secondary carrier-type facilitators. Transport systems are included in this category if they utilize a carrier-mediated process to catalyze uniport (a single species is transported by facilitated diffusion in a process not coupled to the utilization of a primary source of energy), antiport (two or more species are transported in opposite directions in a tightly coupled process not directly linked to a form of energy other than chemiosmotic energy) and or symport (two or more species are transported together in the same direction in a tightly coupled process not directly linked to a form of energy other than chemiosmotic energy) (Figure 7).

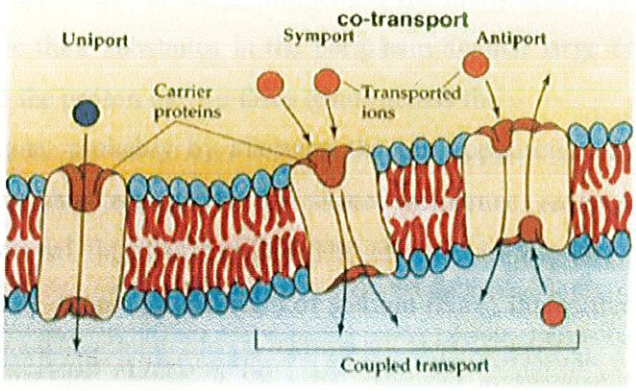


Figure 7 Electrochemical Potential-driven transporters.

These include:

2.A Porters (uniporters, sympo rters, antiporters). Transport systems are included in this subclass if they utilize a carrier-mediated process to catalyze uniport (a single species is transported either by facilitated diffusion or in a membrane potential-dependent process if the solute is charged), antiport (two or more species are transported in opposite directions in a tightly coupled process, not coupled to a direct form of energy other than chemiosmotic energy) and/or symport (two or more species are transported together in the same direction in a tightly coupled process, not coupled to a direct form of energy other than chemiosmotic energy).

2.B Non-ribosomally synthesized porters. These substances, like non-ribosomally synthesized channels, may be depsipeptides or non-peptide-like substances. They complex a cation in their hydrophilic interior and facilitate translocation of the complex across the membrane, exposing their hydrophobic exterior, by moving from one side of the bilayer to the other. If the free porter can cross the membrane in the uncomplexed form, the transport process can be electrophoretic, but if only the complex crosses the membrane, transport is electroneutral.

2.C Ion gradient-driven energizers. Normally, outer membrane porins (1.B) of Gramnegative bacteria catalyze passive transport of solutes across the membrane, but coupled to energizers, they may accumulate their substrates in the periplasm against large concentration gradients. These energizers use the proton motive force (pmf) across the
cytoplasmic membrane, probably by allowing the electrophoretic transport of protons, and conveying conformational change to the outer membrane receptor/porins. Homologous energizers drive bacterial flagellar motility. The mechanism is poorly understood, but these energizers undoubtedly couple proton (H+) or sodium (Na+) fluxes through themselves to the energized process.

3 Primary Active Transporters

The transporters use a primary source of energy to drive active transport of a solute against a concentration gradient. A secondary ion gradient is not considered a primary energy source because it is created by the expenditure of a primary energy source. Primary energy sources known to be coupled to transport are chemical, electrical and solar. These include:

3.A. P-P-bond hydrolysis-driven transporters. Transport systems are included in this subclass if they hydrolyze the diphosphate bond of inorganic pyrophosphate, ATP, or another nucleoside triphosphate, to drive the active uptake and/or extrusion ofa solute or solutes. The transport protein mayor may not be transiently phosphorylated, but the substrate is not phosphorylated.

3.B. Decarboxylation-driven transporters. Transport systems that drive solute (e.g., ion) uptake or extrusion by decarboxylation of a cytoplasmic substrate are included in this subclass.These transporters are currently thought to be restricted to prokaryotes.

3.C. Methyltransfer-driven transporters. A single characterized multisubunit protein family currently falls into this subclass, the Na+-transporting methyltetrahydromethanopterin:coenzyme M methyltransferase. These transporter complexes arecurrently thought to be restricted to archaea.

3.D. Oxidoreduction-driven transporters. Transport systems that drive transport ofa solute (e.g., an ion) energized by the exothermic flow of electrons from a reduced substrate to an oxidized substrate are included in this subclass.

3.E. Light absorption-driven transporters. Transport systems that utilize light energy to drive transport ofa solute (e.g., an ion) are included in this subclass.

4 Group Translocators

PEP-dependent, phosphoryl transfer-driven group translocators of the bacterial phosphoenolpyruvate:sugar phosphotransferase system are the best characterized group translocators included in TC category 4. The product of the reaction, derived from extracellular sugar, is a cytoplasmic sugar-phosphate. The enzymatic constituents, catalyzing sugar phosphorylation, are superimposed on the transport process in a tightly coupled process.

The second type of putative group translocators use nicotinamide ribonucleoside uptakeporters (pnuC; 4.B.l) of bacteria. They are ATP and phosphorylate external nicotinamide ribonucleoide to give cytoplasmic nicotinamide mono nucleotide (NMN) plus ADP.

The third type of group translocators are the putative acyl-CoA ligase-coupled transporters.(4.C.l, 2 and 3) They use the energy of ATP to thioesterify fatty acids and other acids such as carnitine in a process believed to be coupled to transport. A role in group translocation is not fully accepted, and many acyl-CoA ligases clearly do not function directly in transport.

These include:

4.A Phosphotransfer-driven group translocators.

4.B The Nicotinamide Ribonucleoside Uptake Transporters.

4.C The Acyl CoA Ligase-Coupled Transporters.

5 Transport Electron Carriers

Transmembrane electron flow systems. Systems that catalyze electron flow across a biological membrane, from donors localized to one side of the membrane to acceptors localized on the other side, are grouped into TC category 5. These systems contribute to or subtractfi·om the membrane potential, depending on the direction of electron flow. They are therefore important to cellular energetic(Figure 8).

These include:

5.A Transmembrane 2-Electron Transfer Carriers

5.B Transmembrane I-Electron Transfer Carriers

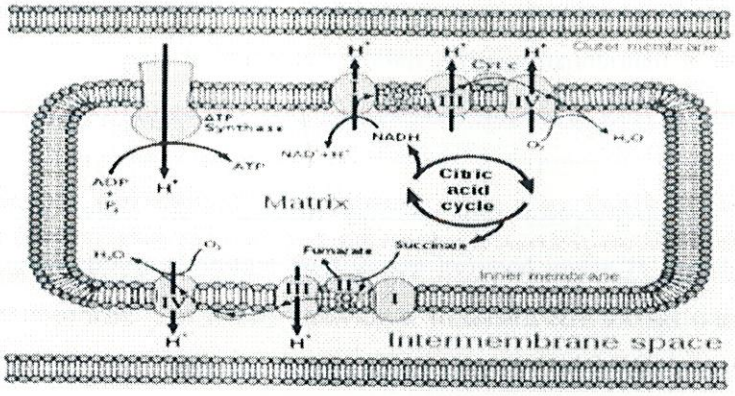


Figure 8 Electron transport chain.

8. Accessory Factors Involved in Transport

Auxiliary transport proteins. Proteins that function with or are complexed to known transport proteins are included in this category. An example would be the membrane fusion proteins that facilitate transport across the two membranes of the Gram-negative bacterial cell envelope in a single step driven by the energy source (AJP) utilized by a cytoplasmic membrane transporter. Energy coupling and regulatory proteins that do not actually participate in transport represent other possible examples. In some cases auxiliary proteins are considered to be part of the transport system with which they function, and in such cases nodistinct entry in category 8 is provided.

These include:

8.A Auxiliary transport proteins. Proteins that in some way facilitate transport across one or more biological membranes but do not themselves participate directly in transport are included in this class. These proteins always function in conjunction with one or more established transport systems. They may provide a function connected with energy coupling transport, play a structural role in complex formation, serve a biogenic or stability function or function in regulation.

8.B. Ribosomally synthesized proteinpeptide toxins that target channels and carriers.

9 Incompletelv Characterized Transport Systems

Transporters of unknown classification. Transport protein families of unknown classification are grouped under TC category 9. Permeases within families maintained in the 9A class are of unknown mode of transport or energy-coupling mechanism, but at least one member of each of these families has clearly been shown to function as a transporter. These families will be classified elsewhere when the transport process and energy-coupling mechanism are characterized. Putative transport protein families are grouped under TC number 9B if they are putative transporters in which no family member is an established transporter. The family will either be classified elsewhere when the transport function of a member becomes established or will be eliminated from the TC classification system if the proposed transport function is disproven. These families include a member or members fur which a transport function has been suggested, but evidence for such a function is not yet compelling. Functionally characterized transporters for which sequences and/or family association are not available are grouped in class 9C.

These include:

9.A Transporters of unknown biochemical mechanism. Transport protein families of unknown classification are grouped in this subclass and will be classified elsewhere when the transport mode and/or energy coupling mechanism are characterized. These families include at least one member fur which a transport function has been established, but either the mode of transport or the energy coupling mechanism is not known.

9.B Putative uncharacterized transport proteins. Putative transport protein families are grouped in this subclass and will either be classified elsewhere when the transport function of a member becomes established, or will be eliminated from the TC classification system if the proposed transport function is disproven.

9.C Functionally characterized transport proteins with unidentified sequences.

Transporters of particular physiological significance will be included in this category even though a family assignment cannot be made. When their sequences are identified, they will be

assigned to an established family. This is the only protein subclass which includes individual proteins rather than protein families.

1.4 Need of Prediction and classification of Transport proteins

Membrane proteins perform a diverse variety of functions, including the transport of ions and molecules across the membrane, bind to small molecules at the extra cellular space, recognize the immune system and energy transducers. The functional annotation of membrane proteins in genomic sequences is an important problem in bioinformatics and computational biology. Membrane transporters are a large group of proteins that span the cell membrane and form an intricate system of pumps and channels through which they deliver essential nutrients, eject waste products and assist the cell to sense environmental conditions. They play indispensable roles in the fundamental cellular processes of all organisms.

Using the protein engineering techniques, new transport proteins are been created. The large international genome sequence projects are gaining a great amount of pub lie attention and huge sequence data bases are created. It becomes more and more obvious that we are very limited in our ability to access functional data for the gene products - the proteins. It seems quite improbable to experimentally determine function and structure of each candidate protein. So a revolutionary method is needed to solve this computation catastrophe. Primary sequence of these proteins are readily available, therefore a method using the sequence derived features will prove a much valuable and a cost effective process of determining and classifying these proteins into broader transporter/non-transporter and specifically into? major classes as defined by Transport Classification (TC) system.

CHAPTER 2

NEURAL NETWORKS FOR PROTEIN CLASSIFICATION

Molecular biology is a field that has experienced dramatic developments in recent years. A large number of data are constantly being generated thanks to several genomes –sequencing projects throughout the world. However, little information can readily extracted from these data and, therefore, data analysis has becomes a central issue in molecular biology. The analysis includes methods and algorithms fur preprocessing visualization, knowledge discovery and data- mining of genomic and proteomic data. A vertiginous increase in the rate at which new protein structures are discovered has taken place as a by-product of ongoing sequencing projects. The functional annotation of membrane proteins in genomic sequences is an important problem in bioinformatics and computational biology.

2.1 What is an artificial neural network?

ANN is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. Why would be necessary the implementation of artificial neural networks? Although computing these days is truly advanced, there are certain tasks that a program made for a common microprocessor is unable to perform; even so a software implementation of a neural network can be made with their advantages and disadvantages.

Advantages:

- A neural network can perform tasks that a linear program cannot.
- When an element of the neural network fails, it can continue without any problem by their parallel nature.
- A neural network learns and does not need to be reprogrammed.
- It can be implemented in any application.
- It can be implemented without any problem.

Disadvantages:

- The neural network needs training to operate.
- The architecture of a neural network is different from the architecture of microprocessors therefore needs to be emulated.
- Requires high processing time for large neural networks.

Another aspect of the artificial neural networks is that there are different architectures, which consequently requires different types of algorithms, but despite to be an apparently complex system, a neural network is relatively simple.

Artificial neural networks (ANN) are among the newest signal-processing technologies in the engineer's toolbox. The field is highly interdisciplinary, but our approach will restrict the view to the engineering perspective. In engineering, neural networks serve two important functions: as pattern classifiers and as nonlinear adaptive filters. We will provide a brief overview of the theory, learning rules, and applications of the most important neural network models. Definitions and Style of Computation an Artificial Neural Network is an adaptive, most often nonlinear system that learns to perform a function (an input/output map) from data. Adaptive means that the system parameters are changed during operation, normally called the training phase. After the training phase the Artificial Neural Network parameters are fixed and the system is deployed to solve the problem at hand (the testing phase). The Artificial Neural Network is built with a systematic step-by-step procedure to optimize a performance criterion or to follow some implicit internal constraint, which is commonly referred to as the learning rule. The input/output training data are fundamental in neural network technology, because they convey the necessary information to "discover" the optimal operating point. The nonlinear nature of the neural network processing elements (PEs) provides the system with lots of flexibility to achieve practically any desired input/output map, i.e., some Artificial Neural Networks are universal mappers. There is a style in neural computation that is worth describing.

In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase (Figure 9).

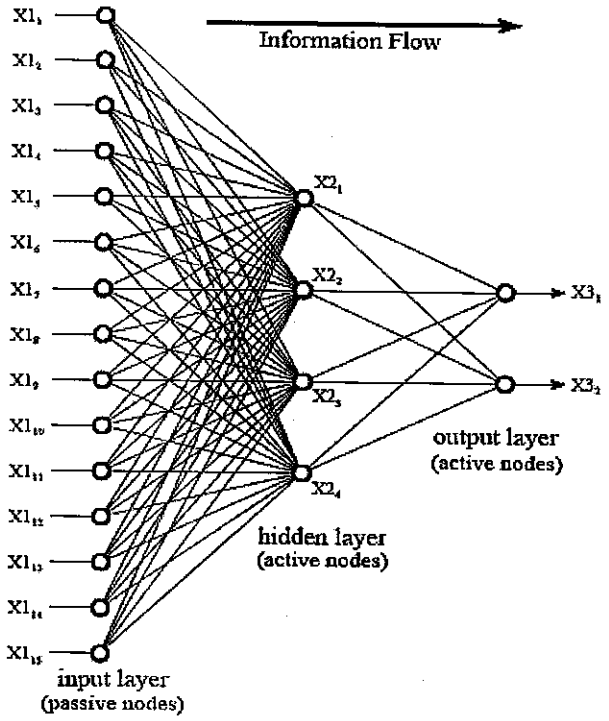


Figure 9 Architecture of Neural Network.

An input is presented to the neural network and a corresponding desired or target response set at the output (when this is the case the training is called supervised). An error is composed from the difference between the desired response and the system output. This error information is fed back to the system and adjusts the system parameters in a systematic fashion (the learning rule). The process is repeated until the performance is acceptable. It is clear from this description that the performance hinges heavily on the data. If one does not have data that cover a significant portion of the operating conditions or if they are noisy, then neural network technology is probably not the right solution. On the other hand, if there is plenty of data and the problem is poorly understood to derive an approximate model, then neural network technology is a good choice. This operating procedure should be contrasted with the traditional engineering design, made of exhaustive subsystem specifications and intercommunication protocols. In artificial neural networks, the designer chooses the network topology, the performance function, the learning rule, and the criterion to stop the training phase, but the system automatically adjusts the parameters. So, it is difficult to bring a priori information into

the design, and when the system does not work properly it is also hard to incrementally refine the solution. But ANN-based solutions are extremely efficient in terms of development time and resources, and in many difficult problems artificial neural networks provide performance that is difficult to match with other technologies. Denker 10 years ago said that "artificial neural networks are the second best way to implement a solution" motivated by the simplicity of their design and because of their universality, only shadowed by the traditional design obtained by studying the physics of the problem. At present, artificial neural networks are emerging as the technology of choice for many applications, such as pattern recognition, prediction, system identification, and control.

2.2 The Biological Model

Artificial neural networks emerged after the introduction of simplified neurons by McCulloch and Pitts in 1943 (McCulloch & Pitts, 1943). These neurons were presented as models of biological neurons and as conceptual components for circuits that could perform computational tasks. The basic model of the neuron is founded upon the functionality of a biological neuron. "Neurons are the basic signaling units of the nervous system" and "each neuron is a discrete cell whose several processes arise from its cell body".

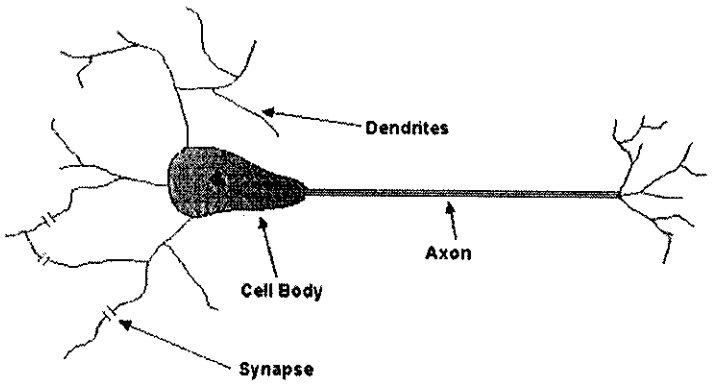


Figure 10 Biological neuron.

The neuron has four main regions to its structure. The cell body, or soma, has two offshoots from it, the dendrites, and the axon, which end in presynaptic terminals (Figure 10). The cell body is the heart of the cell, containing the nucleus and maintaining protein synthesis. A neuron may have many dendrites, which branch out in a treelike structure, and receive signals from other neurons. A neuron usually only has one axon which grows out from a part of the

cell body called the axon hillock. The axon conducts electric signals generated at the axon hillock down its length. These electric signals are called action potentials. The other end of the axon may split into several branches, which end in a presynaptic terminal. Action potentials are the electric signals that neurons use to convey information to the brain. All these signals are identical. Therefore, the brain determines what type of information is being received based on the path that the signal took. The brain analyzes the patterns of signals being sent and from that information it can interpret the type of information being received. Myelin is the fatty tissue that surrounds and insulates the axon. Often short axons do not need this insulation. There are uninsulated parts of the axon. These areas are called Nodes of Ranvier. At these nodes, the signal traveling down the axon is regenerated. This ensures that the signal traveling down the axon travels fast and remains constant (i.e. very short propagation delay and no weakening of the signal). The synapse is the area of contact between two neurons. The neurons do not actually physically touch. They are separated by the synaptic cleft, and electric signals are sent through chemical 13 interaction. The neuron sending the signal is called the presynaptic cell and the neuron receiving the signal is called the postsynaptic cell. The signals are generated by the membrane potential, which is based on the differences in concentration of sodium and potassium ions inside and outside the cell membrane.

2.3 The Mathematical Model

When creating a functional model of the biological neuron, there are three basic components of importance. First, the synapses of the neuron are modeled as weights. The strength of the connection between an input and a neuron is noted by the value of the weight. Negative weight values reflect inhibitory connections, while positive values designate excitatory connections (Figure 11).

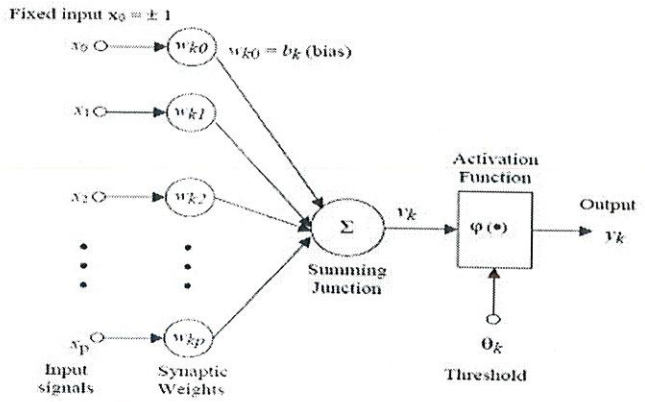


Figure 11 Mathematical description of Neural Network.

The next two components model the actual activity within the neuron cell. An adder sums up all the inputs modified by their respective weights. This activity is referred to as linear combination. Finally, an activation function controls the amplitude of the output of the neuron. An acceptable range of output is usually between 0 and 1, or -1 and 1.

From this model the interval activity of the neuron can be shown to be:

$$v_k = \sum_{j=1}^{p} w_{kj} x_j$$

The output of the neuron, v_k, would therefore be the outcome of some activation function on the value of v_k.

2.4 Activation functions

As mentioned previously, the activation function acts as a squashing function, such that the output of a neuron in a neural network is between certain values (usually 0 and 1, or -1 and 1) (Figure 12). In general, there are three types of activation functions, denoted by Φ(.) . First, there is the Threshold Function which takes on a value of 0 if the summed input is less than a certain threshold value (v), and the value 1 if the summed input is greater than or equal to the threshold value.

$$\varphi(v) = \begin{cases} 1 & \text{if } v \geq 0 \\ 0 & \text{if } v < 0 \end{cases}$$

Secondly, there is the Piecewise-Linear function. This function again can take on the values of 0 or 1, but can also take on values between that depending on the amplification factor in a certain region of linear operation.

$$\varphi(v)=\begin{cases}1 & v\geq\frac{1}{2}\\ v & -\frac{1}{2}>v>\frac{1}{2}\\ 0 & v\leq-\frac{1}{2}\end{cases}$$

Thirdly, there is the sigmoid function. This function can range between 0 and 1, but it is also sometimes useful to use the -1 to 1 range. An example of the sigmoid function is the hyperbolic tangent function.

$$\varphi(v)=\tanh\left(\frac{v}{2}\right)=\frac{1-\exp(-v)}{1+\exp(-v)}$$

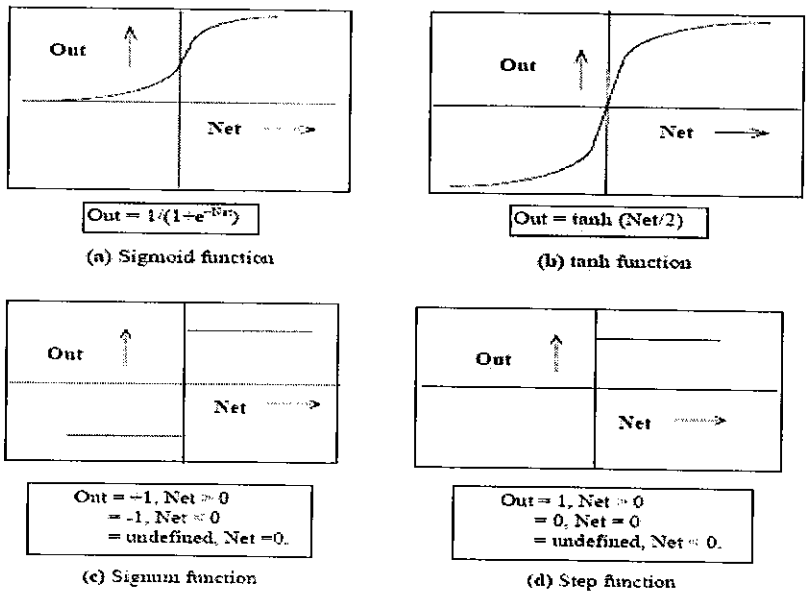


Figure 12 Common nonlinear function used for synaptic inhibition. Soft non-linearity - (a) Sigmoid (b) tanh function. Hard non-linearity - (c) Signum (d) Step function.

The artificial neural networks which we describe are all variations on the parallel distributed processing (PDP) idea. The architecture of each neural network is based on very similar

building blocks which perform the processing. In this chapter we first discuss these processing units and discuss different neural network topologies.

2.5 The Multilayer Perceptron Neural Network Model

This network has an input layer (on the left) with three neurons, one hidden layer (in the middle) with three neurons and an output layer (on the right) with three neurons(Figure 13). There is one neuron in the input layer fur each predictor variable. In the case of categorical variables-*l* neurons are used to represent the *N* categories of the variable. The following diagram illustrates a perceptron network with three layers:

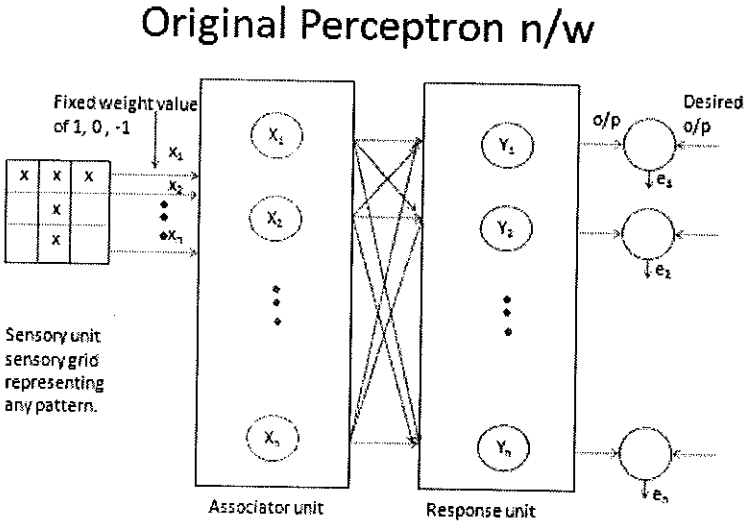


Figure 13 Multilayer perceptron neural network model

Input Layer- A vector of predictor variable values *(x, ...x_p)* is presented to the input layer. The input layer (or processing before the input layer) standardizes these values so that the range of each variable is -1 to 1. The input layer distributes the values to each of the neurons in the hidden layer. In addition to the predictor variables, there is a constant input of 1.0, called the *bias* that is fed to each of the hidden layers; the bias is multiplied by a weight and added to the sum going into the neuron.

Hidden Layer – The arriving at a neuron in the hidden layer, the value from each input neuron is multiplied by a weight (W_{ji}), and the resulting weighted values are added together producing a combined value u_i The weighted sum (u_j) is fed into a transfer function, σ, which outputs a value h_j. The outputs from the hidden layer are distributed to the output layer.

Output Layer- Arriving at a neuron in the output layer, the value from each hidden layer neuron is multiplied by a weight (W_{kj}), and the resulting weighted values are added together producing a combined value vi The weighted sum (v) is fed into a transfer function, σ , which outputs a value Y_k. The y values are the outputs of the network.

If a regression analysis is being performed with a continuous target variable, then there is a single neuron in the output layer, and it generates a single y value.

2.6 Multilayer Perceptron Architecture

The network diagram shown above is a full-connected; three layer, feed-forward, perceptron neural network. "Fully connected" means that the output from each input and hidden neuron is distributed to all of the neurons in the following layer. "Feed forward" means that the values only move from input to hidden to output layers; no values are fed back to earlier layers (a Recurrent Network allows values to be fed backward). All neural networks have an input layer and an output layer, but the number of hidden layers may vary. When there is more than one hidden layer, the output from one hidden layer is fed into the next hidden layer and separate weights are applied to the sum going into each layer.

2.7 Training Multilayer Perceptron Networks

The goal of the training process is to find the set of weight values that will cause the output from the neural network to match the actual target values as closely as possible. There are several issues involved in designing and training a multi layer perceptron network:

- Selecting how many hidden layers to use in the network.
- Deciding how many neurons to use in each hidden layer.
- Finding a globally optimal solution that avoids local minima.
- Converging to an optimal solution in a reasonable period of time.
- Validating the neural network to test for over fitting.

2.8 Selecting the Number of Hidden Layers

For nearly all problems, one hidden layer is sufficient. Two hidden layers are required for modeling data with discontinuities such as a saw tooth wave pattern. Using two hidden layers rarely improves the model, and it may introduce a greater risk of converging to a local minima. There is no theoretical reason for using more than two hidden layers. SANN Statistica can build models with one or two hidden layers. Three layer models with one hidden layer are recommended.

2.9 Deciding how many neurons to use in the hidden layers

One of the most important characteristics of a perceptron network is the number of neurons in the hidden layer(s). If an inadequate number of neurons are used, the network will be unable to model complex data, and the resulting fit will be poor.

SANN Statistica includes an automated feature to find the optimal number of neurons in the hidden layer.You specify the minimum and maximum number of neurons you want it to test, and it will build models using varying numbers of neurons and measure the quality using either cross validation or hold-out data not used for training. This is a highly effective method for finding the optimal number of neurons, but it is computationally expensive, because many models must be built, and each model has to be validated. If you have a multiprocessor computer, you can configure SANN Statistica to use multiple CPU's during the process.

The automated search for the optimal number of neurons only searches the first hidden layer. If you select a model with two hidden layers, you must manually specify the number of neurons in the second hidden layer.

2.10 Finding a globally optimal solution

A typical neural network might have a couple of hundred weighs whose values must be found to produce an optimal solution If neural networks were linear models like linear regression, it would be a breeze to find the optimal set of weights. But the output of a neural network as a function of the inputs is often highly nonlinear; this makes the optimization process complex.

2.11 Converging to the Optimal Solution – BFGS

Given a set of randomly-selected starting weight values, SANN Statistica uses the BFGS algorithm to optimize the weight values. Most training algorithms follow this cycle to refine the weight values: (1) run a set of predictor variable values through the network using a tentative set of weights, (2) compute the difference between the predicted target value and the actual target value for this case, (3) average the error information over the entire set of training cases, (4) propagate the error backward through the network and compute the gradient (vector of derivatives) of the change in error with respect to changes in weight values, (5) make adjustments to the weights to reduce the error. Each cycle is called an *epoch*.

OBJECTIVE:

With the explosion of protein sequences entering into databanks, it is highly desirable to explore the feasibility of selectively classifying newly found protein sequences into their respective transport protein classes by means of an automated method. This is indeed important because knowing which protein belongs to which particular class may help to deduce its catalytic mechanism and specificity, giving clues to the relevant biological function. With the availaibility of huge amount of genome sequencing data generated each day and for their functional annotation, sequence derived features are useful approaches.

Here in this study an attempt has been taken for distinguishing protein sequences into transport protein classes using ANN for annotation of protein sequence with following objectives:

1. To extract sequence derived features and selection of important features from protein sequence to be used for prediction and classification of transport proteins.
2. To develop and optimize the 1st layer for classifying the user input protein sequence into transport and non-transport based on sequence derived features.
3. To develop and optimized the 2nd layer for classifying the predicted transport protein into seven major classes based on sequence derived features.
4. To develop a 3rd layer for classifying the predicted class of transport protein into their corresponding sub-classes and thus their specific functions.

In the study along with usage of machine learning approach like ANN automated as wells as customized, we have also used three types of parameters like amino acid composition, pseudo

amino acid composition and physicochemical properties. Using this combination of we have seven neural network models- ANN $_{AA\ comp}$, ANN$_{PseAA}$ and ANN $_{pepstat}$, as the individual ANNs and rest all are their combinations. For building up the neural network models we have used *STATISTICA* v.9.1 by Statsoft. SANN is *STATISTICA* Enterprise-Wide Data Mining System (Data Miner) that offers a comprehensive selection of Neural Network solutions. By joining these models we have developed 7 neural network clusters. These neural networks cluster takes the sequences one by one for the prediction.

Chapter 3

Materials and Methods

Overview of the work

In this study we have developed a cluster of neural networks consisting of three layers with usage of machine learning approach-ANN. The sequence derived features that were used are amino acid composition, pseudo amino acid composition and physicochemical properties. Using these parameters and their combination we have developed in total seven neural network clusters-ANN_{AAcomp}, ANN_{PseAA}, $ANN_{pepstat}$, $ANN_{AA+PseAA}$, $ANN_{AA+pepstat}$, $ANN_{PseAA+pepstat}$, $ANN_{AA+PseAA+pepstat}$. The overall protocol used in this study is described below (Figure 14).

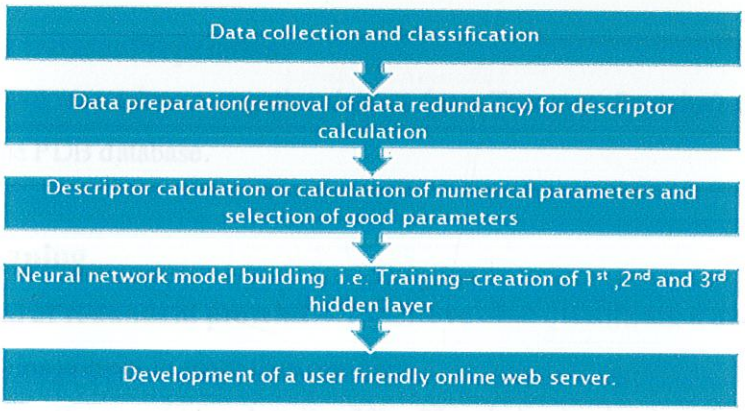


Figure 14 Protocol used in this study.

3.1 Data sources

The data is taken from Transport Classification Database (TCDB) in which the proteins are classified on the basis of functions (Table 1).

Table 1 Initial Dataset.

Class	Total no. of sequence
1.Channels_Pores Transport protein	1139
2.Electrochemical Potential-driven transporters	1456
3.Electrochemical Potential-driven transporters	2045
4.Group Translocators	107
5.Transport Electron Carriers	106
6.Accessory Factors Involved in Transport	129
7.Incompletely Characterized Transport Systems	377

Non-Transport Dataset Preparation:

Similarly we have also taken a negative dataset (non-Transport Proteins) consisting of 2907 proteins from the PDB database.

3.2. Data cleaning

I. **Removal of redundant proteins** – To achieve clear classification division we removed the proteins present in more than one class.

II. **Sequences Removed (seq. length <20)** - The chains having length less than 20 amino acids are not believed to go under protein folding process and thus, would not form a proper domain and may not function properly. We therefore, removed such protein chains from each class.

III. **Removed Incorrect Sequences (B, U, Z, X, *)** – Removal of all protein chains containing non-standard amino acids, nucleotide sequences (sequences containing only A,T,G,C) and un-annotated amino acids (X or *).

IV. **Data Scaling** –

- Similar sequences removed (using blastclust with similarity <30%).
- Those sub-classes having no. of sequences <10(or very less as compared other sub-classes of the same class) are not considered for layer3 while the same are considered for layer2.

- To remove the biasness in layer 2, the data set(at class level) is further reduced(sequences are removed randomly as similarity among the sequences was very less) (Table 2,3).

Table 2 Data set used for training , testing and validation of 2nd layer of ANN.

Class	Total no. of sequence	Training set	Test set	Validation set
1.Channels_Pores Transport protein	1139	545	157	164
2.Electrochemical Potential-driven transporters	1456	558	148	73
3.Electrochemical Potential-driven transporters	2045	896	210	134
4.Group Translocators	107	90	17	20
5.Transport Electron Carriers	106	81	25	21
6.Accessory Factors Involved in Transport	129	109	20	26
7.Incompletely Characterized Transport Systems	377	268	75	49

Table 3 Data set used for training , testing and validation of 3rd layer of ANN. 1, 2, ... represent sub-classes and are described in Appendix I

Class & Subclass	Total no. of sequence	Training set	Test set	Validation set
1				
1.A	481	386	95	50
1.B	269	212	57	52
1.C	309	246	63	57
1.E	38	34	4	5
3				
3.A	1612	1280	332	67
3.B	22	20	2	3
3.D	370	301	69	61
3.E	27	24	3	3
4				
4.A	91	73	18	17
4.C	12	10	2	3
5				
5.A	61	50	11	11
5.B	45	35	10	10
8				
8.A	94	78	16	17
8.B	35	26	9	9
9				
9.A	211	168	43	26
9.B	164	132	32	23

3.3 Descriptor Calculation

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, and health researches, as well as in quality control, being the way molecules, thought of as real bodies, are transformed into numbers, allowing some mathematical treatment of the chemical information contained in the molecule. This was defined by Todeschini and Consonni as:

"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment."

The descriptors used are:

1. **Amino Acid composition:** This descriptor consist of 20 factors each representing composition of 20 standard amino acids in the protein sequences that include A,C,D,E,F,G,H,I,K,L,M,P,Q,R,S,T,V,W,X,andY.The formula to calculate this composition is:

$$AA\ comp(i) = \frac{Freq.of\ AA(i)}{\sum Freq.of\ AA\ in\ seq.}$$

2. **Physicochemical Properties**: This descriptor consists of 12 properties calculated using EMBOSS (EBI) package. The parameters include Molecular weight, Average residue weight, Charge, Isoelectric point., A280 Molar Extinction Coefficient, A280 Extinction Coefficient 1mg/ml, Mole percentages of Tiny, Small, Aliphatic, Aromatic, Non-polar, Polar, Charged, Acidic, Basic amino acids .The different categories include different sets of amino acids like Tiny (A+C+G+S+T), Small (A+B+C+D+G+N+P+S+T+V), Aliphatic (I+L+V), Aromatic (F+H+W+Y), Non-polar (A+C+F+G+I+L+M+P+V+W+Y), Polar (D+E+H+K+N+Q+R+S+T+Z), Charged (B+D+E+H+K+R+Z), Basic (H+K+R) and Acidic (B+D+E+Z).

3. **Pseudo AA composition**: This descriptor is a collection of 37 factors, 20 of which are simple amino acid compositions and rest 17 are correlation factors calculated among amino acids of the given sequences. It was introduced by Kuo-Chen Chou in 2001 to represent protein samples for statistical prediction.

The simplest discrete model is using the AA composition to represent protein samples, as formulated as follows. Given a protein sequence **P** with L amino acid resides, i.e.,

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \qquad (1)$$

where R_1 represents the 1st residue of the protein **P**, R_2 the 2nd residue, and so forth, according to the AA composition model, the protein **P** of Eq.1 can be expressed by

$$\mathbf{P} = \begin{bmatrix} f_1 & f_2 & \cdots & f_{20} \end{bmatrix}^{\mathbf{T}} \qquad (2)$$

where $f_u\,(u = 1, 2, \cdots, 20)$ are the normalized occurrence frequencies of the 20 native amino acids in **P**, and **T** the transposing operator. The additional factors are a series of rank-different correlation factors along a protein chain, but they can also be any combinations of other factors so long as they can reflect some sorts of sequence-order effects one way or the other. The algorithm for this is as follows:

According to the PseAA composition model, the protein **P** of Eq.1 can be formulated as

$$\mathbf{P} = \begin{bmatrix} p_1, & p_2, & \cdots, & p_{20}, & p_{20+1}, & \cdots, & p_{20+\lambda} \end{bmatrix}^{\mathbf{T}}, \quad (\lambda < L) \qquad (3)$$

where $20 + \lambda$ the components are given by

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda} \tau_k}, & (1 \le u \le 20) \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda} \tau_k}, & (20+1 \le u \le 20+\lambda) \end{cases} \qquad (4)$$

Where w is the weight factor, and τ_k the k-th tier correlation factor that reflects the sequence order correlation between all the k-th most contiguous residues as formulated by

$$\tau_k = \frac{1}{L-k}\sum_{i=1}^{L-k} \mathrm{J}_{i,i+k}, \quad (k < L) \qquad (5)$$

with

$$\mathrm{J}_{i,i+k} = \frac{1}{\Gamma}\sum_{g=1}^{\Gamma}\left[\Phi_\xi\left(\mathrm{R}_{i+k}\right) - \Phi_\xi\left(\mathrm{R}_i\right)\right]^2 \qquad (6)$$

Where $\Phi_\xi(R_i)$ is the ξ-th function of the amino acid R_i, and Γ the total number of the functions considered. $\Phi_1(R_i)$, $\Phi_2(R_i)$ and $\Phi_3(R_i)$ are respectively the hydrophobicity value, hydrophilicity value, and side chain mass of amino acid R_i(Table 4); while $\Phi_1(R_{i+k})$, $\Phi_2(R_{i+k})$ and $\Phi_3(R_{i+k})$ the corresponding values for the amino acid R_{i+k}. Therefore, the total number of functions considered there is Γ=3.

Table 4 Scales used in PseAA (a)The hydrophobicity values are from JACS, 1962, 84: 4240-4246. (C. Tanford),(b)The hydrophilicity values are from PNAS, 1981, 78:3824-3828 (T.P.Hopp & K.R.Woods) and(c)The side-chain mass for each of the 20 amino acids.

Amino acid	Hydrophobicity a	Hydrophilicity b	Side chain mass c
A	0.62	-0.5	15
C	0.29	-1	47
D	-0.9	3	59
E	-0.74	3	73
F	1.19	-2.5	91
G	0.48	0	1
H	-0.4	-0.5	82
I	1.38	-1.8	57
K	-1.5	3	73
L	1.06	-1.8	57
M	0.64	-1.3	75
N	-0.78	0.2	58
P	0.12	0	42
Q	-0.85	0.2	72
R	-2.53	3	101
S	-0.18	0.3	31
T	-0.05	-0.4	45
V	1.08	-1.5	43
W	0.81	-3.4	130
Y	0.26	-2.3	107

It can be seen from Eq.3 that the first 20 components, i.e. $p_1, p_2, \cdots, p_{20}$ are associated with the conventional AA composition of protein , while the remaining components $p_{20+1}, \cdots, p_{20+\lambda}$ are the correlation factors that reflect the 1st tier, 2nd tier, ..., and the λ-th tier sequence order correlation patterns. It is through these additional λ factors that some important sequence-order effects are incorporated

3.4 Binary classification of proteins

In this classification level for a sequence calculation of composition of 20 amino acids is given as input to the binary level and on the basis of this we predict whether it is Transport protein or a non-Transport protein. If it is transport then it is further send to second level of classification system. Similarly we can classify the sequence on the basis of physicochemical properties, pseudo amino acid composition, or any combination of the three.

3.5 Classifications of transport protein into major classes

In this classification level for a sequence calculation of composition of 20 amino acids is given as input to the level and on the basis of this composition we prediction whether the particular sequence belong to Channels_Pores, Electrochemical Potential-driven transporters, Primary Active Transporters, Group Translocators, Transport Electron Carriers, Accessory Factors Involved in Transport and Incompletely Characterized Transport Systems . Belonging to any class it will further be sent to third level of classification system. Similarly we can classify the sequence on the basis of physicochemical properties, pseudo amino acid composition, or any combination of the three.

3.6 Classification of transport protein into subclasses

Taking the input parameters as the composition of 20 amino acids calculated from a sequence and then going for sub class classification. In this particular layer the transport protein is classified into its respective sub-classes. If we classify sequence as Channels_Pores in the second layer then it can be classified as 1.A. a-Type channels or 1.B. ß-Barrel porins or 1.C. Pore-forming toxins (proteins and peptides) or 1.E. Holins. If in layer 2 it was classified Electrochemical Potential-driven transporters then in this layer it will be classified P-P-bond-

hydrolysis-driven transporters or Decarboxylation-driven transporters or Oxidoreduction-driven transporters or Light absorption-driven transporters. If in layer 2 it was classified Group Translocators , then in this layer it will be classified Phosphotransfer-driven group translocators or Acyl CoA ligase-coupled transporters. If in layer 2 it was classified Transport Electron Carriers, then in this layer it will be classified Transmembrane 2-electron transfer carriers or Transmembrane 1-electron transfer carriers. If in layer 2 it was classified Accessory Factors Involved in Transport, then in this layer it will be classified Auxiliary transport proteins or Ribosomally synthesized proteinpeptide toxins that target channels and carriers and if previously it was categorized as Incompletely Characterized Transport Systems then it will further be classified Recognized transporters of unknown biochemical mechanism or Putative transport proteins (Figure 15).

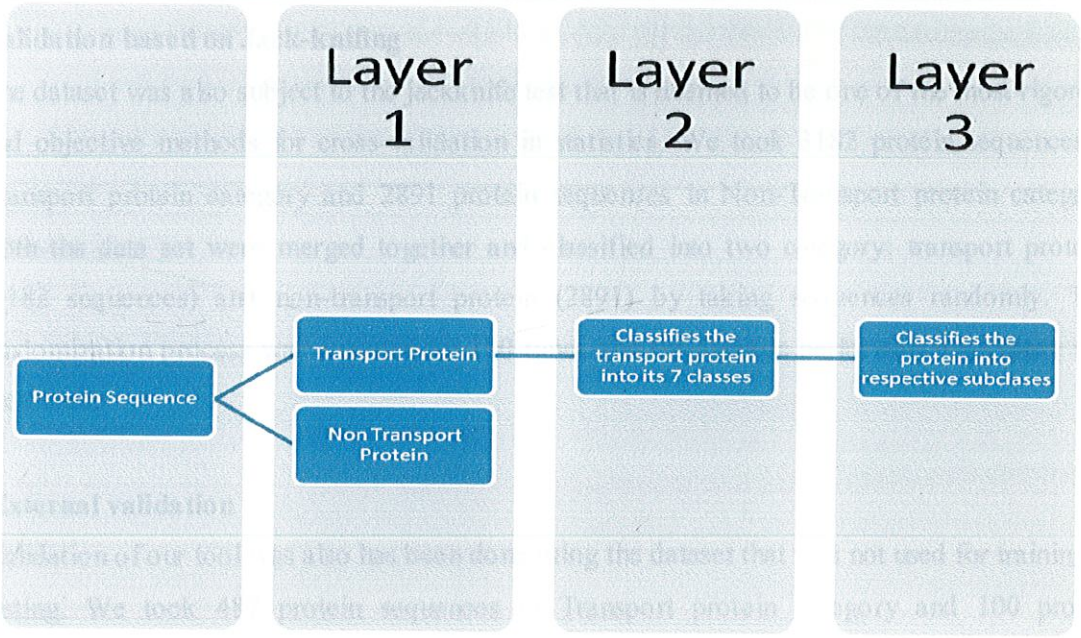


Figure 15 Classification heirarchy of TpPred.

3.7 Validation of hierarchical classification model

The validation is the way to confirm the validity of data, information, processes or a model. We have used three different approaches to validate our tool as follows.

Validation based on self consistency

The performance of our online tool TpPred was validated using self consistency method. In this approach the data set of transport proteins were given as input to the tool and the predicted result was observed. The predicted accuracy at each level of classification was calculated based on the predicted output. We took 3182 protein sequences in Transport protein category and 2891 protein sequences in Non-Transport protein category.

Validation based on Jack-knifing

The dataset was also subject to the jackknife test that is deemed to be one of the most rigorous and objective methods for cross-validation in statistics. We took 3182 protein sequences in Transport protein category and 2891 protein sequences in Non-Transport protein category. Both the data set were merged together and classified into two category: transport proteins (3182 sequences) and non-transport protein (2891) by taking sequences randomly. This randomization process was repeated for 100 times and the average performance accuracy was measured.

External validation

Validation of our tool was also has been done using the dataset that was not used for training or testing. We took 487 protein sequences in Transport protein category and 100 protein sequences in Non-Transport protein category as independent data set.

3.8 Standalone and Server development of TpPred

A standalone as well as online version of our tool (TpPred) has been developed and uploaded into our University web server. The following steps have been used for the development of the server:

- Deployed and modified the C codes for each Neural Network model generated.
- Converted the C codes to C library references as Header files.

- Generated a main parser code, which can take in the descriptors from a file and can send them to the particular network models in their corresponding header files and retrieves the output of the model. Based on output, it takes the decision to go which way in the hierarchy or to which particular model to feed the descriptor and retrieve the output. This step is repeated until a case gets predicted to the terminal node in the hierarchy i.e. reaching to a particular sub-class if it is predicted as a Transport protein.
- Generated Perl parsers to link the webpage with the prediction cluster codes, in case of online tool and to link prediction clusters directly in case of standalone. These perl codes can retrieve the sequence and the choice of sequence based feature from user and then can convert the protein sequence to features and present them to prediction clusters for prediction.

  We have generated a Perl parser which can take the constraints of the descriptor from the user and take in the sequences from the file, then it can pipe in the sequences to the other codes for calculation of the desired descriptor and thereby the prediction cluster is fired, which has been developed in the previous case and this retrieve the output and presents it to the user.

Figure 16 illustrates the front end of the online version with sequence to be submitted as query and options for the sequence derived features to be used and its output for an example sequence.

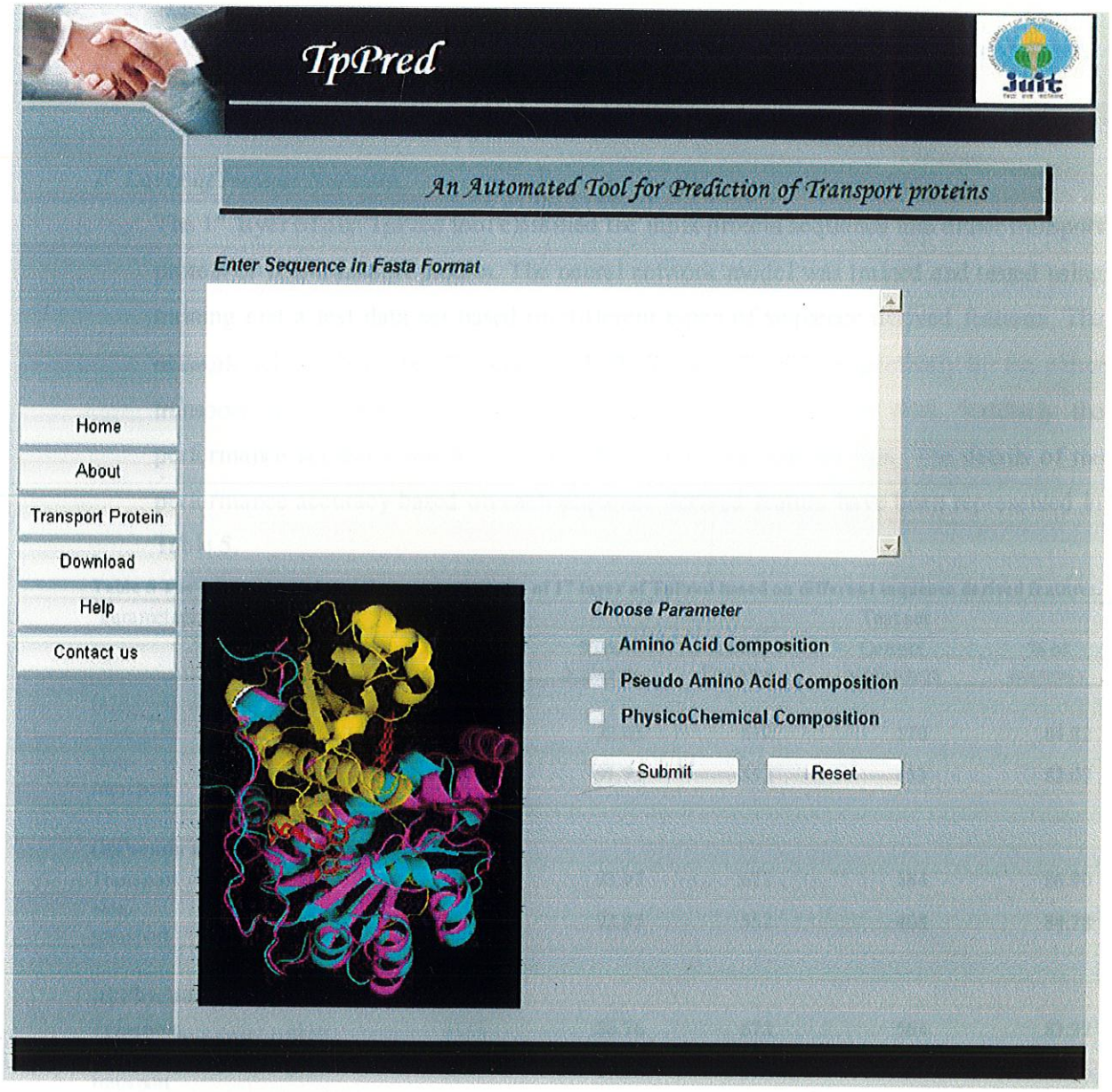


Figure 16 The frontend of the online version of TpPred

Chapter 4

Results and Discussion

1st Layer of Neural Network

The 1st layer of our TpPred tool classified the input protein sequence into either transport protein or non-transport protein. The neural network model was trained and tested using training and a test data set based on different types of sequence derived features. The network achieved an overall accuracy of 93.27% and 93.20% respectively for the either transport protein and non-transport protein for the training set data. Similarly the performance accuracy was 84.92% and 84.24% for the test set data. The details of the performance accuracy based on each sequence derived feature have been represented in Table 5.

Table 5 The summary of the performance accuracy of 1st layer of TpPred based on different sequence derived features.

Parameters	Training set			Test set		
	Total Proteins	Correct predictions	% of accuracy	Total Proteins	Correct predictions	% of accuracy
(a)Amino acid composition						
Transport	2510	2310	92.03	672	570	84.82
Non-transport	2339	2147	91.79	552	453	82.07
(b)Pseudo amino acid composition						
Transport	2510	2346	93.47	672	584	86.90
Non-transport	2339	2174	92.95	552	468	84.78
(c) Physicochemical Properties						
Transport	2510	2228	88.76	672	564	83.92
Non-transport	2339	2050	87.64	552	448	81.16
(d) Amino acid+ Pseudo amino acid						
Transport	2510	2405	95.81	672	571	84.97
Non-transport	2339	2230	95.34	552	464	84.06
(e) Amino acid+ PhysicoChemical Properties						
Transport	2510	2380	94.82	672	554	82.44
Non-transport	2339	2187	93.50	552	457	82.79

(f) Pseudo amino acid + PhysicoChemical Properties						
Transport	2510	2255	89.84	672	576	85.71
Non-transport	2339	2219	94.87	552	490	88.77
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties						
Transport	2510	2465	98.20	672	576	85.71
Non-transport	2339	2254	96.37	552	475	86.05
Average						
Transport			93.27			84.92
Non-Transport			93.20			84.24

The performance accuracy was further validated using self consistency test and jackknife test. The overall accuracy of the 1st layer of TpPred is 77.73% and 89.18% for the transport and non-transport protein classes based on self consistency test. Similarly using Jackknife test the accuracy was found to be 71.54% and 81.01% for the transport protein and non-transport protein classes (Table 6). Moreover the results were robust and hence the TpPred could successfully predict the novel protein sequence into either of transport protein and non-transport protein as evident from the independent data set used for validation Table 6.

Table 6 The performance accuracy of the 1st layer of TpPred based on validation techniques (self consistency test, Jackknife and independent set validation).

Parameters	Total Proteins	Self Consistency	Jackknife Validation	Independent set	
				Total	Correct%
(a)Amino acid composition					
Transport	3182	75.78	71.97	487	59.75
Non-transport	2891	88.79	81.27	100	96
(b)Pseudo amino acid composition					
Transport	3182	78.89	71.39	487	60.36
Non-transport	2891	90.76	85.29	100	96
(c) PhysicoChemical Properties					
Transport	3182	68.57	62.96	487	50.10
Non-transport	2891	85.02	78.19	100	97
(d) Amino acid+ Pseudo amino acid					
Transport	3182	81.45	75.28	487	62.21
Non-transport	2891	90.73	81.17	100	98
(e) Amino acid+ PhysicoChemical Properties					
Transport	3182	75.69	66.92	487	59.34

Non-transport	2891	83.91	78.25	100	99
(f) Pseudo amino acid+PhysicoChemical Properties					
Transport	3182	82.16	76.38	487	66.11
Non-transport	2891	90.71	82.95	100	100
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties					
Transport	3182	81.61	75.93	487	70.43
Non-transport	2891	94.39	86.95	100	100
Average					
Transport		77.73	71.54		61.18
Non-Transport		89.18	81.01		98

By comparing the performance accuracy of the 1st layer of TpPred for the individual sequence derived feature; it has been observed that the accuracy was better by fusing amino acid, pseudo amino acid and physicochemial properties.

2nd Layer of Neural Network

The 2nd layer of our TpPred tool classified the input protein sequence into either Channels_Pores, Electrochemical Potential-driven transporters, Primary Active Transporters, Group Translocators, Transport Electron Carriers, Accessory Factors Involved in Transport and Incompletely Characterized Transport Systems. The neural network model was trained and tested using training and a test data set based on different types of sequence derived features. The network achieved an overall accuracy of 90.29%, 87.60%, 90.31%, 95.39%, 93.82%, 88.85% and 90.29% respectively for the Channels_Pores, Electrochemical Potential-driven transporters, and Primary Active Transporters, Group Translocators, Transport Electron Carriers, Accessory Factors Involved in Transport and Incompletely Characterized Transport Systems proteins for the training set data. Similarly the performance accuracy was 70.60%, 72.68%, 72.92%, 79.82%, 70.85%, 71.42% and 83.99% for the test set data. The details of the performance accuracy based on each sequence derived feature have been represented in Table 7.

Table 7 The summary of the performance accuracy of 2nd layer of TpPred based on different sequence derived features.

Parameters	Training set			Test set		
	Total Proteins	Correct predictions	% of accuracy	Total Proteins	Correct predictions	% of accuracy
(a)Amino acid composition						
1	545	457	83.85	157	118	75.15
2	558	460	82.43	148	111	75
3	896	755	84.26	210	134	63.80
4	90	85	94.44	17	14	82.35
5	81	70	86.41	25	19	76
8	109	81	74.31	20	15	75
9	268	240	89.55	75	60	80
(b)Pseudo amino acid composition						
1	545	493	90.45	157	112	71.33
2	558	503	90.14	148	110	74.32
3	896	830	92.63	210	152	72.38
4	90	90	100	17	12	70.58
5	81	81	100	25	16	64
8	109	109	100	20	16	80
9	268	268	100	75	62	82.66
(c) PhysicoChemical Properties						
1	545	468	85.87	157	109	69.42
2	558	457	81.89	148	113	76.35
3	896	780	87.05	210	172	81.90
4	90	80	88.88	17	13	76.47
5	81	68	83.95	25	19	76
8	109	80	73.39	20	11	55
9	268	232	86.56	75	63	84
(d) Amino acid+ Pseudo amino acid						
1	545	530	97.24	157	111	70.70
2	558	497	89.06	148	110	74.32
3	896	830	92.63	210	154	73.33
4	90	90	100	17	15	88.23
5	81	81	100	25	16	64
8	109	109	100	20	15	75
9	268	268	100	75	64	85.33
(e) Amino acid+ PhysicoChemical Properties						
1	545	470	86.23	157	105	66.87
2	558	480	86.02	148	104	70.27
3	896	772	86.16	210	157	74.76
4	90	80	88.88	17	15	88.23

5	81	72	88.88	25	18	72
8	109	84	77.06	20	13	65
9	268	255	95.14	75	64	85.33
(f) Pseudo amino acid+PhysicoChemical Properties						
1	545	505	92.66	157	116	73.88
2	558	501	89.78	148	96	64.86
3	896	864	96.42	210	154	73.33
4	90	86	95.55	17	13	76.47
5	81	79	97.53	25	19	76
8	109	106	97.24	20	13	65
9	268	266	99.25	75	65	86.66
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties						
1	545	522	95.77	157	105	66.87
2	558	524	93.90	148	109	73.64
3	896	834	93.08	210	149	70.95
4	90	90	100	17	13	76.47
5	81	81	100	25	17	68
8	109	109	100	20	17	85
9	268	268	100	75	63	84

Average		
1	90.29	70.60
2	87.60	72.68
3	90.31	72.92
4	95.39	79.82
5	93.82	70.85
8	88.85	71.42
9	90.29	83.99

The performance accuracy was further validated using self consistency test and jackknife test. The overall accuracy of the 2nd layer of TpPred is 60.91%, 75.96%, 51.93%, 78.14%, 87.95%, 84.23% and 77.16% respectively for the Channels_Pores, Electrochemical Potential-driven transporters, and Primary Active Transporters, Group Translocators, Transport Electron Carriers, Accessory Factors Involved in Transport and Incompletely Characterized Transport Systems classes based on self consistency test. Similarly, using jackknife test, the accuracy was found to be 49.99%, 62.57%, 39.69%, 65.90%, 73.41%, 70.24% and 63.33% respectively for the Channels_Pores, Electrochemical Potential-driven transporters, and Primary Active Transporters, Group

Translocators, Transport Electron Carriers, Accessory Factors Involved in Transport and Incompletely Characterized Transport Systems proteins classes (Table 8). Moreover, the results were robust and hence, the TpPred could successfully predict the novel protein sequence into either of Channels_Pores, Electrochemical Potential-driven transporters, and Primary Active Transporters, Group Translocators, Transport Electron Carriers, Accessory Factors Involved in Transport or Incompletely Characterized Transport Systems proteins class as evident from the independent data set used for validation Table 8.

Table 8 The performance accuracy of the 2nd layer of TpPred based on validation techniques (self consistency test, jackknife test and independent set validation).

Parameters	Total Proteins	Self Consistency	Jackknife validation	Independent set	
				Total	Correct%
(a)Amino acid composition					
1	702	64.58	51.27	164	61.58
2	706	81.08	63.86	73	78.08
3	1106	46.53	34.94	134	42.53
4	107	84	71.57	20	80
5	106	85.95	69.58	21	80.95
8	129	79.07	65.27	26	73.07
9	343	52.97	41.95	49	48.97
(b)Pseudo amino acid composition					
1	702	61.31	51.74	164	57.31
2	706	75.60	61.76	73	72.60
3	1106	43.05	39.34	134	38.05
4	107	77	63.97	20	75
5	106	94.47	75.35	21	90.47
8	129	90.46	74.99	26	88.46
9	343	85.59	71.94	49	79.59
(c) PhysicoChemical Properties					
1	702	45.463	39.18	164	41.463
2	706	72.49	58.37	73	68.49
3	1106	47.04	33.74	134	41.04
4	107	55	41.83	20	50
5	106	68.90	59.63	21	61.90
8	129	60.69	48.69	26	57.69
9	343	69.34	45.49	49	67.34
(d) Amino acid+ Pseudo amino acid					
1	702	61.53	46.78	164	58.53
2	706	84.82	71.58	73	80.82
3	1106	49.02	34.28	134	44.02
4	107	75	61.84	20	70
5	106	94.47	79.98	21	90.47
8	129	86.76	71.25	26	80.76
9	343	77.42	68.37	49	71.42
(e) Amino acid+ PhysicoChemical Properties					

1	702	63.75	59.64	164	59.75
2	706	55.68	48.94	73	50.68
3	1106	54.50	41.35	134	48.50
4	107	74	61.97	20	70
5	106	79.19	69.38	21	76.19
8	129	86.76	75.59	26	80.76
9	343	82.55	69.49	49	77.55
(f) Pseudo amino acid+PhysicoChemical Properties					
1	702	69.85	55.38	164	65.85
2	706	72.12	61.92	73	67.12
3	1106	55.25	41.48	134	49.25
4	107	86	78.25	20	80
5	106	95.47	81.37	21	90.47
8	129	96.30	82.64	26	92.30
9	343	83.63	70.99	49	81.63
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties					
1	702	59.92	45.98	164	57.92
2	706	89.93	71.58	73	84.93
3	1106	68.17	52.74	134	64.17
4	107	96	81.93	20	90
5	106	97.23	78.64	21	95.23
8	129	89.61	73.28	26	84.61
9	343	88.63	75.09	49	81.63

Average			
1	60.91	49.99	57.48
2	75.96	62.57	71.81
3	51.93	39.69	46.79
4	78.14	65.90	73.57
5	87.95	73.41	83.66
8	84.23	70.24	79.66
9	77.16	63.33	72.59

By comparing the performance accuracy of the 2nd layer of TpPred between the individual sequence derived features; it has been observed that the accuracy was better by combining amino acid, pseudo amino acid and physiochemical properties.

3rd Layer of Neural Network for Channels_Pores class

The 3rd layer of our TpPred tool developed for Channels_Pores class classified the input protein sequence to be either a-Type channels, ß-Barrel porins, Pore-forming toxins (proteins and peptides) or Holins. The neural network model was trained and tested using training and a test data set based on different types of sequence derived features. The network achieved an overall accuracy of 96.44%, 91.97%, 89.21% and 89.49%

respectively for the a-Type channels, ß-Barrel porins, Pore-forming toxins (proteins and peptides) and Holins proteins for the training set data. Similarly the performance accuracy was 83.30%, 82.70%, 82.98% and 75.00% for the test set data. The details of the performance accuracy based on each sequence derived feature have been represented in Table 9.

Table 9 The summary of the performance accuracy of 3rd layer of TpPred developed for Channels_Pores class based on different sequence derived features.

Parameters	Training set			Test set		
	Total Proteins	Correct predictions	% of accuracy	Total Proteins	Correct predictions	% of accuracy
(a)Amino acid composition						
1						
1.A	386	368	95.33	95	78	82.10
1.B	212	188	88.67	57	46	80.70
1.C	246	193	78.45	63	52	82.53
1.E	34	25	73.52	4	3	75
(b)Pseudo amino acid composition						
1						
1.A	386	373	96.63	95	82	86.31
1.B	212	191	90.09	57	49	85.96
1.C	246	229	93.08	63	54	85.71
1.E	34	30	88.23	4	3	75
(c) PhysicoChemical Properties						
1						
1.A	386	349	90.41	95	74	77.89
1.B	212	169	79.71	57	44	77.19
1.C	264	201	76.13	63	49	77.77
1.E	34	28	82.35	4	3	75
(d) Amino acid+ Pseudo amino acid						
1						
1.A	386	378	97.92	95	80	84.21
1.B	212	206	97.16	57	48	84.21
1.C	246	234	95.12	63	53	84.12
1.E	34	34	100	4	3	75
(e) Amino acid+ PhysicoChemical Properties						
1						
1.A	386	372	96.37	95	79	83.15
1.B	212	195	91.981	57	47	82.45
1.C	246	218	88.61	63	52	82.53
1.E	34	32	94.11	4	3	75

(f) Pseudo amino acid+PhysicoChemical Properties						
1						
1.A	386	385	99.74	95	79	83.15
1.B	212	212	100	57	47	82.45
1.C	246	243	98.78	63	52	82.53
1.E	34	34	100	4	3	75
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties						
1						
1.A	386	381	98.70	95	82	86.31
1.B	212	204	96.22	57	49	85.96
1.C	246	232	94.30	63	54	85.71
1.E	34	30	88.23	4	3	75

Average		
1.A	96.44	83.30
1.B	91.97	82.70
1.C	89.21	82.98
1.E	89.49	75

The performance accuracy was further validated using self consistency test and jackknife test. The overall accuracy of the 3rd layer of TpPred for Channels_Pores class is 41.95%, 79.43%, 64.10% and 72.36% respectively for the a-Type channels, ß-Barrel porins, Pore-forming toxins (proteins and peptides) and Holins classes based on self consistency test. Similarly, using jackknife test, the accuracy was found to be 33.71%, 65.37%, 54.26% and 61.15% respectively for the a-Type channels, ß-Barrel porins, Pore-forming toxins (proteins and peptides) and Holins classes (Table 10). Moreover, the results were robust and hence, the TpPred could successfully predict the novel protein sequence into either of a-Type channels, ß-Barrel porins, Pore-forming toxins (proteins and peptides) or Holins class as evident from the independent data set used for validation Table 10.

Table 10 The performance accuracy of the 3rd layer of TpPred developed for Channels_Pores based on validation techniques (self consistency test, jackknife test and independent set validation).

Parameters	Total Proteins	Self Consistency	Jackknief validation	Independent Set	
				Total	Correct%
(a)Amino acid composition					
1					
1.A	481	83.54	61.22	50	82
1.B	269	77.92	60.34	52	76.92
1.C	309	39.08	39.55	57	35.08
1.E	38	45.87	34.53	5	40
(b)Pseudo amino acid composition					
1					
1.A	481	41	35.19	50	36
1.B	269	75.16	60.35	52	71.153
1.C	309	68.43	51.37	57	61.40
1.E	38	82.85	69.17	5	80
(c) PhysicoChemical Properties					
1					
1.A	481	42.85	31.87	50	38
1.B	269	52.73	45.16	52	50
1.C	309	40.49	31.97	57	36.84
1.E	38	44.85	33.08	5	40
(d) Amino acid+ Pseudo amino acid					
1					
1.A	481	21.67	18.97	50	14
1.B	269	98.79	75.13	52	94.23
1.C	309	65.17	56.98	57	63.15
1.E	38	82	75.19	5	80
(e) Amino acid+ PhysicoChemical Properties					
1					
1.A	481	43.76	31.28	50	40
1.B	269	75.25	61.96	52	69.23
1.C	309	71.72	57.84	57	66.66
1.E	38	83	66.94	5	80
(f) Pseudo amino acid+PhysicoChemical Properties					
1					
1.A	481	37	36.18	50	32
1.B	269	86.27	77.84	52	80.76
1.C	309	82.49	69.85	57	80.70
1.E	38	83	71.25	5	80
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties					
1					
1.A	481	23.87	21.28	50	10
1.B	269	89.92	76.86	52	82.69

1.C	309	81.36	72.26	57	75.43
1.E	38	85	77.89	5	80
Average					
1.A		41.95	33.71		36
1.B		79.43	65.37		74.99
1.C		64.10	54.26		59.89
1.E		72.36	61.15		68.57

By comparing the performance accuracy of the 3rd layer of TpPred for Channels_Pores class between the individual sequence derived features; it has been observed that the accuracy was better by combining amino acid, pseudo amino acid and physiochemical properties.

3rd Layer of Neural Network for Electrochemical Potential-driven transporters class

The 3rd layer of our TpPred tool developed for Electrochemical Potential-driven transporters class classified the input protein sequence to be either P-P-bond-hydrolysis-driven transporters, Decarboxylation-driven transporters, Oxidoreduction-driven transporters or Light absorption-driven transporters. The neural network model was trained and tested using training and a test data set based on different types of sequence derived features. The network achieved an overall accuracy of 98.32%, 73.57%, 85.94% and 76.18% respectively for the P-P-bond-hydrolysis-driven transporters, Decarboxylation-driven transporters, Oxidoreduction-driven transporters and Light absorption-driven transporters proteins for the training set data. Similarly the performance accuracy was 89.71 %, 100%, 89.43% and 80.94% for the test set data. The details of the performance accuracy based on each sequence derived feature have been represented in Table 11.

Table 11 The summary of the performance accuracy of 3rd layer of TpPred developed for Electrochemical Potential-driven transporters class based on different sequence derived features.

Parameters	Training set			Test set		
	Total Proteins	Correct predictions	% of accuracy	Total Proteins	Correct predictions	% of accuracy
(a)Amino acid composition						
3						
3.A	1280	1253	97.89	332	295	88.85
3.B	20	6	30	2	2	100
3.D	301	216	71.76	69	61	88.40
3.E	24	13	54.16	3	2	66.66
(b)Pseudo amino acid composition						
3						
3.A	1280	1260	98.43	332	297	89.45
3.B	20	18	90	2	2	100
3.D	301	270	89.70	69	62	89.85
3.E	24	20	83.33	3	2	66.66
(c) PhysicoChemical Properties						
3						
3.A	1280	1230	96.09	332	298	89.75
3.B	20	10	50	2	2	100
3.D	301	230	76.41	69	62	89.85
3.E	24	15	62.5	3	2	66.66
(d) Amino acid+ Pseudo amino acid						
3						
3.A	1280	1273	99.45	332	302	90.96
3.B	20	18	90	2	2	100
3.D	301	280	93.02	69	63	91.30
3.E	24	23	95.83	3	2	66.66
(e) Amino acid+ PhysicoChemical Properties						
3						
3.A	1280	1257	98.20	332	299	90.06
3.B	20	18	90	2	2	100
3.D	301	255	84.71	69	61	88.40
3.E	24	15	62.5	3	3	100
(f) Pseudo amino acid+PhysicoChemical Properties						
3						
3.A	1280	1264	98.75	332	295	88.85
3.B	20	15	75	2	2	100
3.D	301	274	91.02	69	61	88.40
3.E	24	19	79.16	3	3	100
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties						
3						
3.A	1280	1273	99.45	332	299	90.06
3.B	20	18	90	2	2	100

3.D	301	286	95.01	69	62	89.85
3.E	24	23	95.83	3	3	100
Average						
3.A			98.32			89.71
3.B			73.57			100
3.D			85.94			89.43
3.E			76.18			80.94

The performance accuracy was further validated using self consistency test and jackknife test. The overall accuracy of the 3rd layer of TpPred for Electrochemical Potential-driven transporters class is 52.29%, 57.37%, 50.84% and 78.98% respectively for the a P-P-bond-hydrolysis-driven transporters, Decarboxylation-driven transporters, Oxidoreduction-driven transporters and Light absorption-driven transporters classes based on self consistency test. Similarly, using jackknife test, the accuracy was found to be 41.14%, 45.86%, 38.71% and 63.52% respectively for the P-P-bond-hydrolysis-driven transporters, Decarboxylation-driven transporters, Oxidoreduction-driven transporters and Light absorption-driven transporters classes (Table 12). Moreover, the results were robust and hence, the TpPred could successfully predict the novel protein sequence into either of P-P-bond-hydrolysis-driven transporters, Decarboxylation-driven transporters, Oxidoreduction-driven transporters or Light absorption-driven transporters class as evident from the independent data set used for validation Table 12.

Table 12 The performance accuracy of the 3rd layer of TpPred developed for Electrochemical Potential-driven transporters based on validation techniques (self consistency test, jackknife test and independent set validation).

Parameters	Total Proteins	Self Consistency	Jackknife validation	Independent Set	
				Total	Correct%
(a)Amino acid composition					
3					
3.A	1612	50.06	35.32	67	47.76
3.B	22	39.38	29.75	3	33.33
3.D	370	41.05	33.33	61	37.70
3.E	27	38.31	31.74	3	33.33
(b)Pseudo amino acid composition					
3					
3.A	1612	47.27	35.15	67	40.29
3.B	22	37.48	32.94	3	33.33
3.D	370	39.46	29.19	61	34.42
3.E	27	70.65	58.14	3	66.66

(c) PhysicoChemical Properties					
3					
3.A	1612	43.86	35.96	67	38.80
3.B	22	34.43	27.06	3	33.33
3.D	370	46.67	31.94	61	42.62
3.E	27	100	80.29	3	100
(d) Amino acid+ Pseudo amino acid					
3					
3.A	1612	46.73	33.87	67	41.79
3.B	22	69.99	58.97	3	66.66
3.D	370	48.21	35.14	61	44.26
3.E	27	71.63	53.97	3	66.66
(e) Amino acid+ PhysicoChemical Properties					
3					
3.A	1612	57.74	43.75	67	47.76
3.B	22	71.71	59.38	3	66.66
3.D	370	52.49	39.28	61	47.54
3.E	27	72.29	59.19	3	66.66
(f) Pseudo amino acid+PhysicoChemical Properties					
3					
3.A	1612	49.48	36.97	67	36.97
3.B	22	72.39	51.28	3	51.28
3.D	370	57.92	42.25	61	42.25
3.E	27	100	75.79	3	75.79
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties					
3					
3.A	1612	70.92	66.96	67	64.17
3.B	22	76.21	61.69	3	66.66
3.D	370	70.14	59.86	61	62.29
3.E	27	100	85.58	3	100

Average			
3.A	52.29	41.14	45.36
3.B	57.37	45.86	50.17
3.D	50.84	38.71	44.44
3.E	78.98	63.52	72.72

By comparing the performance accuracy of the 3rd layer of TpPred for Electrochemical Potential-driven transporters class between the individual sequence derived features; it has been observed that the accuracy was better by combining amino acid, pseudo amino acid and physiochemical properties.

3rd Layer of Neural Network for Group Translocators class

The 3rd layer of our TpPred tool developed for Group Translocators class classified the input protein sequence to be either Phosphotransfer-driven group translocators or Acyl CoA ligase-coupled transporters. The neural network model was trained and tested using training and a test data set based on different types of sequence derived features. The network achieved an overall accuracy of 100% and 98.57% respectively for the Phosphotransfer-driven group translocators and Acyl CoA ligase-coupled transporters proteins for the training set data. Similarly the performance accuracy was 99.20% and 100% for the test set data. The details of the performance accuracy based on each sequence derived feature have been represented in Table 13.

Table 13 The summary of the performance accuracy of 3rd layer of TpPred developed for Group Translocators class based on different sequence derived features.

Parameters	Training set			Test set		
	Total Proteins	Correct predictions	% of accuracy	Total Proteins	Correct predictions	% of accuracy
(a)Amino acid composition						
4						
4.A	73	73	100	18	17	94.44
4.C	10	9	90	2	2	100
(b)Pseudo amino acid composition						
4						
4.A	73	73	100	18	18	100
4.C	10	10	100	2	2	100
(c) PhysicoChemical Properties						
4						
4.A	73	73	100	18	18	100
4.C	10	10	100	2	2	100
(d) Amino acid+ Pseudo amino acid						
4						
4.A	73	73	100	18	18	100
4.C	10	10	100	2	2	100
(e) Amino acid+ PhysicoChemical Properties						
4						
4.A	73	73	100	18	18	100
4.C	10	10	100	2	2	100
(f) Pseudo amino acid+PhysicoChemical Properties						
4						
4.A	73	73	100	18	18	100
4.C	10	10	100	2	2	100
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties						

4						
4.A	73	73	100	18	18	100
4.C	10	10	100	2	2	100
Average						
4.A			100			99.20
4.C			98.57			100

The performance accuracy was further validated using self consistency test and jackknife test. The overall accuracy of the 3rd layer of TpPred for Group Translocators class is 85.77% and 43.91% respectively for the Phosphotransfer-driven group translocators and Acyl CoA ligase-coupled transporters classes based on self consistency test. Similarly, using jackknife test, the accuracy was found to be 70.71% and 32.50% respectively for the Phosphotransfer-driven group translocators and Acyl CoA ligase-coupled transporters classes (Table 14). Moreover, the results were robust and hence, the TpPred could successfully predict the novel protein sequence into either of Phosphotransfer-driven group translocators or Acyl CoA ligase-coupled transporters class as evident from the independent data set used for validation Table 14.

Table 14 The performance accuracy of the 3rd layer of TpPred developed for Group Translocators based on validation techniques (self consistency test, jackknife test and independent set validation).

Parameters	Total Proteins	Self Consistency	Jackknife validation	Independent Set	
				Total	Correct%
(a)Amino acid composition					
4					
4.A	91	90.03	69.37	17	88.23
4.C	12	43.38	31.09	3	33.33
(b)Pseudo amino acid composition					
4					
4.A	91	85.57	75.03	17	82.35
4.C	12	40.37	27.36	3	33.33
(c) PhysicoChemical Properties					
4					
4.A	91	62.39	51.01	17	58.82
4.C	12	35.30	25.99	3	33.33
(d) Amino acid+ Pseudo amino acid					
4					
4.A	91	84.32	70.98	17	82.35
4.C	12	35.30	26.98	3	33.33

(e) Amino acid+ PhysicoChemical Properties					
4					
4.A	91	88.39	77.78	17	82.35
4.C	12	37	27.35	3	33.33
(f) Pseudo amino acid+PhysicoChemical Properties					
4					
4.A	91	92.46	71.39	17	88.23
4.C	12	39.78	22.27	3	33.33
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties					
4					
4.A	91	97.28	79.47	17	94.11
4.C	12	76.25	66.52	3	66.66

Average			
4.A	85.77	70.71	82.34
4.C	43.91	32.50	38.09

By comparing the performance accuracy of the 3rd layer of TpPred for Group Translocators class between the individual sequence derived features; it has been observed that the accuracy was better by combining amino acid, pseudo amino acid and physiochemical properties.

3rd Layer of Neural Network for Transport Electron Carriers class

The 3rd layer of our TpPred tool developed for Transport Electron Carriers class classified the input protein sequence to be either Transmembrane 2-electron transfer carriers or Transmembrane 1-electron transfer carriers. The neural network model was trained and tested using training and a test data set based on different types of sequence derived features. The network achieved an overall accuracy of 100% and 100% respectively for the Transmembrane 2-electron transfer carriers and Transmembrane 1-electron transfer carriers proteins for the training set data. Similarly the performance accuracy was 98.70% and 98.57% for the test set data. The details of the performance accuracy based on each sequence derived feature have been represented in Table 15.

Table 15 The summary of the performance accuracy of 3rd layer of TpPred developed for Transport Electron Carriers class based on different sequence derived features.

Parameters	Training set			Test set		
	Total Proteins	Correct predictions	% of accuracy	Total Proteins	Correct predictions	% of accuracy
(a)Amino acid composition						

5						
5.A	50	50	100	11	11	100
5.B	35	35	100	10	10	100
(b)Pseudo amino acid composition						
5						
5.A	50	50	100	11	11	100
5.B	35	35	100	10	10	100
(c) PhysicoChemical Properties						
5						
5.A	50	50	100	11	10	90.90
5.B	35	35	100	10	9	90
(d) Amino acid+ Pseudo amino acid						
5						
5.A	50	50	100	11	11	100
5.B	35	35	100	10	10	100
(e) Amino acid+ PhysicoChemical Properties						
5						
5.A	50	50	100	11	11	100
5.B	35	35	100	10	10	100
(f) Pseudo amino acid+PhysicoChemical Properties						
5						
5.A	50	50	100	11	11	100
5.B	35	35	100	10	10	100
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties						
5						
5.A	50	50	100	11	11	100
5.B	35	35	100	10	10	100

Average		
5.A	100	98.7
5.B	100	98.57

The performance accuracy was further validated using self consistency test and jackknife test. The overall accuracy of the 3rd layer of TpPred for Transport Electron Carriers class is 79.91% and 93.27% respectively for the Transmembrane 2-electron transfer carriers and Transmembrane 1-electron transfer carriers classes based on self consistency test. Similarly, using jackknife test, the accuracy was found to be 64.35% and 77.70% respectively for the Transmembrane 2-electron transfer carriers and Transmembrane 1-electron transfer carriers classes (Table 16). Moreover, the results were robust and hence, the TpPred could successfully predict the novel protein sequence into either of Transmembrane 2-electron transfer carriers or Transmembrane

1-electron transfer carriers class as evident form the independent data set used for validation Table 16.

Table 16 The performance accuracy of the 3rd layer of TpPred developed for Transport Electron Carriers based on validation techniques (self consistency test, jackknife test and independent set validation).

Parameters	Total Proteins	Self Consistency	Jackknief validation	Independent Set	
				Total	Correct%
(a)Amino acid composition					
5					
5.A	61	82.91	64.75	11	81.81
5.B	45	81	71.98	10	80
(b)Pseudo amino acid composition					
5					
5.A	61	85.89	76.09	11	81.81
5.B	45	100	81.30	10	100
(c) PhysicoChemical Properties					
5					
5.A	61	49.91	38.19	11	45.45
5.B	45	83	61.29	10	80
(d) Amino acid+ Pseudo amino acid					
5					
5.A	61	84.17	63.26	11	81.81
5.B	45	100	81.26	10	100
(e) Amino acid+ PhysicoChemical Properties					
5					
5.A	61	69.92	51.26	11	63.63
5.B	45	94	79.92	10	90
(f) Pseudo amino acid+PhysicoChemical Properties					
5					
5.A	61	93	80.58	11	90.90
5.B	45	94.93	79.25	10	90
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties					
5					
5.A	61	93.59	76.38	11	90.90
5.B	45	100	88.96	10	100
Average					
5.A		79.91	64.35		76.61
5.B		93.27	77.70		91.42

By comparing the performance accuracy of the 3rd layer of TpPred for Transport Electron Carriers class between the individual sequence derived features; it has been observed that the

accuracy was better by combining amino acid, pseudo amino acid and physiochemical properties.

3rd Layer of Neural Network for Accessory Factors Involved in Transport class

The 3rd layer of our TpPred tool developed for Accessory Factors Involved in Transport class classified the input protein sequence to be either Auxiliary transport proteins or Ribosomally synthesized proteinpeptide toxins that target channels and carriers. The neural network model was trained and tested using training and a test data set based on different types of sequence derived features. The network achieved an overall accuracy of 99.63% and 98.35% respectively for the Auxiliary transport proteins and Ribosomally synthesized proteinpeptide toxins that target channels and carriers proteins for the training set data. Similarly the performance accuracy was 100% and 100% for the test set data. The details of the performance accuracy based on each sequence derived feature have been represented in Table 17.

Table 17 The summary of the performance accuracy of 3rd layer of TpPred developed for Accessory Factors Involved in Transport class based on different sequence derived features.

Parameters	Training set			Test set		
	Total Proteins	Correct predictions	% of accuracy	Total Proteins	Correct predictions	% of accuracy
(a)Amino acid composition						
8						
8.A	78	78	100	16	16	100
8.B	26	26	100	9	9	100
(b)Pseudo amino acid composition						
8						
8.A	78	78	100	16	16	100
8.B	26	26	100	9	9	100
(c) PhysicoChemical Properties						
8						
8.A	78	78	100	16	16	100
8.B	26	26	100	9	9	100
(d) Amino acid+ Pseudo amino acid						
8						
8.A	78	78	100	16	16	100
8.B	26	26	100	9	9	100
(e) Amino acid+ PhysicoChemical Properties						
8						
8.A	78	77	98.71	16	16	100
8.B	26	24	92.30	9	9	100

(f) Pseudo amino acid+PhysicoChemical Properties						
8						
8.A	78	78	100	16	16	100
8.B	26	26	100	9	9	100
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties						
8						
8.A	78	77	98.71	16	16	100
8.B	26	25	96.15	9	9	100
Average						
8.A			99.63			100
8.B			98.35			100

The performance accuracy was further validated using self consistency test and jackknife test. The overall accuracy of the 3rd layer of TpPred for Accessory Factors Involved in Transport class is 79.77% and 87.83% respectively for the Auxiliary transport proteins and Ribosomally synthesized proteinpeptide toxins that target channels and carriers classes based on self consistency test. Similarly, using jackknife test, the accuracy was found to be 67.91% and 73.37% respectively for the Auxiliary transport proteins and Ribosomally synthesized proteinpeptide toxins that target channels and carriers classes (Table 18). Moreover, the results were robust and hence, the TpPred could successfully predict the novel protein sequence into either of Auxiliary transport proteins or Ribosomally synthesized proteinpeptide toxins that target channels and carriers class as evident from the independent data set used for validation Table 18.

Table 18 The performance accuracy of the 3rd layer of TpPred developed for Accessory Factors Involved in Transport based on validation techniques (self consistency test, jackknife test and independent set validation).

Parameters	Total Proteins	Self Consistency	Jackknife validation	Independent Set	
				Total	Correct%
(a)Amino acid composition					
8					
8.A	94	78.42	69.53	17	76.47
8.B	35	69.96	51.86	9	66.66
(b)Pseudo amino acid composition					
8					
8.A	94	95.21	82.07	17	94.11
8.B	35	78.87	63.19	9	77.77
(c) PhysicoChemical Properties					
8					
8.A	94	46.12	36.79	17	41.17
8.B	35	92.83	71.39	9	88.88

(d) Amino acid+ Pseudo amino acid					
8					
8.A	94	79.42	72.12	17	76.47
8.B	35	90.98	81.47	9	88.88
(e) Amino acid+ PhysicoChemical Properties					
8					
8.A	94	76.72	61.27	17	70.58
8.B	35	100	86.52	9	100
(f) Pseudo amino acid+PhysicoChemical Properties					
8					
8.A	94	91.37	78.29	17	88.23
8.B	35	100	81.94	9	100
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties					
8					
8.A	94	91.17	75.34	17	88.23
8.B	35	82.18	77.24	9	77.77

Average			
8.A	79.77	67.91	76.46
8.B	87.83	73.37	85.70

By comparing the performance accuracy of the 3rd layer of TpPred for Accessory Factors Involved in Transport class between the individual sequence derived features; it has been observed that the accuracy was better by combining amino acid, pseudo amino acid and physiochemical properties.

3rd Layer of Neural Network for Incompletely Characterized Transport Systems class

The 3rd layer of our TpPred tool developed for Incompletely Characterized Transport Systems class classified the input protein sequence to be either Recognized transporters of unknown biochemical mechanism or Putative transport proteins. The neural network model was trained and tested using training and a test data set based on different types of sequence derived features. The network achieved an overall accuracy of 98.29% and 97.50% respectively for the Recognized transporters of unknown biochemical mechanism and Putative transport proteins proteins for the training set data. Similarly the performance accuracy was 76.07% and 76.78% for the test set data. The details of the performance accuracy based on each sequence derived feature have been represented in Table 19

Table 19 The summary of the performance accuracy of 3rd layer of TpPred developed for Incompletely Characterized Transport Systems class based on different sequence derived features.

Parameters	Training set			Test set		
	Total Proteins	Correct predictions	% of accuracy	Total Proteins	Correct predictions	% of accuracy
(a)Amino acid composition						
9						
9.A	168	162	96.42	43	33	76.74
9.B	132	124	93.93	32	24	75
(b)Pseudo amino acid composition						
9						
9.A	168	168	100	43	33	76.74
9.B	132	132	100	32	25	78.12
(c) PhysicoChemical Properties						
9						
9.A	168	154	91.66	43	30	69.76
9.B	132	117	88.63	32	22	68.75
(d) Amino acid+ Pseudo amino acid						
9						
9.A	168	168	100	43	34	79.06
9.B	132	132	100	32	26	81.25
(e) Amino acid+ PhysicoChemical Properties						
9						
9.A	168	168	100	43	33	76.74
9.B	132	132	100	32	25	78.12
(f) Pseudo amino acid+PhysicoChemical Properties						
9						
9.A	168	168	100	43	33	76.74
9.B	132	132	100	32	25	78.12
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties						
9						
9.A	168	168	100	43	33	76.74
9.B	132	132	100	32	25	78.12
Average						
9.A			98.29			76.07
9.B			97.508			76.78

The performance accuracy was further validated using self consistency test and jackknife test. The overall accuracy of the 3rd layer of TpPred for Incompletely Characterized Transport Systems class is 76.19% and 77.26% respectively for the Recognized transporters of unknown biochemical mechanism and Putative transport proteins classes based on self consistency test. Similarly, using jackknife test, the accuracy was found to be 63.88% and 60.63% respectively for the Recognized transporters of unknown biochemical mechanism and Putative transport

proteins classes (Table 20). Moreover, the results were robust and hence, the TpPred could successfully predict the novel protein sequence into either of Recognized transporters of unknown biochemical mechanism or Putative transport proteins class as evident from the independent data set used for validation Table 20.

Table 20 The performance accuracy of the 3rd layer of TpPred developed for Incompletely Characterized Transport Systems based on validation techniques (self consistency test, jackknife test and independent set validation).

Parameters	Total Proteins	Self Consistency	Jackknief validation	Independent Set	
				Total	Correct%
(a)Amino acid composition					
9					
9.A	211	40.43	31.87	26	38.46
9.B	164	66.81	50.12	23	60.86
(b)Pseudo amino acid composition					
9					
9.A	211	75.78	60.69	26	73.07
9.B	164	90.91	71.64	23	86.95
(c) PhysicoChemical Properties					
9					
9.A	211	73.25	65.97	26	69.23
9.B	164	68.68	51.13	23	65.21
(d) Amino acid+ Pseudo amino acid					
9					
9.A	211	80.14	69.06	26	76.92
9.B	164	69.25	58.92	23	65.21
(e) Amino acid+ PhysicoChemical Properties					
9					
9.A	211	83.92	71.38	26	80.76
9.B	164	78.57	62.14	23	73.91
(f) Pseudo amino acid+PhysicoChemical Properties					
9					
9.A	211	89.42	75.38	26	84.61
9.B	164	81.48	68.84	23	78.26
(g) Amino acid+ Pseudo amino acid+PhysicoChemical Properties					
9					
9.A	211	90.39	72.85	26	84.61
9.B	164	85.17	61.64	23	78.26
Average					
9.A		76.19	63.88		72.52
9.B		77.26	60.63		72.66

By comparing the performance accuracy of the 3rd layer of TpPred for Incompletely Characterized Transport Systems class between the individual sequence derived features; it has been observed that the accuracy was better by combining amino acid, pseudo amino acid and physiochemical properties.

Chapter 5

CONCLUSION

From a practical point of view, the most important aspect of a prediction model is its ability to make correct predictions. Till date most of the available methods use the 3-D structure of the protein to predict and classify transport protein. This is a very tedious job and requires much costlier endeavors. The sequence of a protein is an important determinant for the detailed molecular function of proteins, and would consequently also be useful for prediction of transport protein and classes. Additionally much encouraging results have been predicted using the sequence derived parameters technique. Therefore, a much accurate and reliable method is that predicts the transport proteins and transport protein classes based on both strategies.

This thesis contains detailed work on transport protein prediction and classification. We achieved an accuracy of ~ 78% for the prediction of the Transport proteins and its classification into major class and sub-classes using three layer artificial neural networks. The first level of network imitates the binary model, the second level of network classify the predicted transport protein into 7 major classes and the third level of network uses the predicted results of the former to provide a much detailed and useful classification. The neural network architecture used for the prediction was optimized for maximum accuracy. This was achieved by gradually testing networks with variable hidden nodes and retaining the one with highest true predictions. This is the only best prediction tool available till date, but to the contrary, uses a much simpler and efficient prediction method based on sequence features. This application not only gives optimal results with the dataset used but also predicts transport proteins from complex genomes to a very high satisfactory level. A much elaborate analysis has been done, which is evident from the extracted data, figures and tables compiled.

PUBLICATIONS

1. TpPred: An online tool for hierarchical prediction of Transport proteins using cluster of neural networks and sequence derived features.

   Sankalp Jain, Piyush Ranjan, Pooja Kesari and Pradeep Kumar Naik

   (Communicated to: Journal of Computational Biology)

2. MetalloPred: An online tool for hierarchical prediction of Metal Ion Binding proteins using cluster of neural networks and sequence derived features.

   Piyush Ranjan, Pooja Kesari, Sankalp Jain and Pradeep Kumar Naik

   (Communicated to: Journal of Computational Biology)

REFERENCES

1. Lodish et al., Molecular Cell Biology, 6, W. H. Freeman 2008.
2. Gunther Winkehnann, Microbial Transport System.Wiley-VCH, 2001
3. Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer ISBN 0-387-31073-8 (2007)
4. Neural Computing and Applications, Springer- Verlag
5. Altschul sf Gish W, Miller W, Myers EW, Lipman DJ: basic Local Alignment Search tool. J Mol Boil 1990, 215:415-41 O.
6. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: Hidden markov roodels in computational biology. Applications to protein modeling. J Mol Biol 1994, 235: 1501-1531.
7. Vinga S, Ahneida J: Alignment free sequence comparison -a review, Bioinformatics 2003, 19:513-523.
8. Abascal F, Valencia A: Aotomatic annotation of protein funcyion based on family identification. Proteins 2003, 53: 683-692.
9. Krebs WG, Bourne PE; Stastically rigorous automated protein annotation. Bioinformatics 2004, 20:1066-1073.
10. Devos D, Valencia A: Intrinsic errors in genome annotation. Trends Genet 2001,17: 429-431.
11. Holm L, Sander C: Mapping the protein universe. Science 1996,273: 595-603.
12. Anfisen, C.B. 1973 Principles that govern the folding of protein chains. Science 181,223-230.
13. Chou, K.C. and Zhang, C.T 1995. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. BioI. 30, 275-349.
14. Klein, P. 1986. Prediction of protein structural class by disriminant analysis. Biochem. Biophys. Acta. 874, 205-215.

Appendix I

Abbreviations used in tables

Classes	Abbreviations
1.Channels_Pores	**1**
a-Type channels	1.A
ß-Barrel porins	1.B
Pore-forming toxins (proteins and peptides)	1.C
Holins	1.E
3.Electrochemical Potential-driven transporters	**3**
P-P-bond-hydrolysis-driven transporters	3.A
Decarboxylation-driven transporters	3.B
Oxidoreduction-driven transporters	3.D
Light absorption-driven transporters	3.E
4.Group Translocators	**4**
Phosphotransfer-driven group translocators	4.A
Acyl CoA ligase-coupled transporters	4.C
5.Transport Electron Carriers	**5**
Transmembrane 2-electron transfer carriers	5.A
Transmembrane 1-electron transfer carriers	5.B
8.Accessory Factors Involved in Transport	**8**
Auxiliary transport proteins	8.A
Ribosomally synthesized proteinpeptide toxins that target channels and carriers	8.B
9.Incompletely Characterized Transport Systems	**9**
Recognized transporters of unknown biochemical mechanism	9.A
Putative transport proteins	9.B

Appendix II

Cluster c code:

An example c parser code of ANN_{PseAA} cluster.

```
#include<stdio.h>
#include<math.h>
#include<stdlib.h>
#include<conio.h>
#include<string.h>
#include "L1.h"
#include "L2.h"
#include "L3A.h"
#include "L3AE.h"
#include "L3T.h"
#include "L4TG1.h"
#include "L4TG2.h"
#include "L4TG3.h"

int main()
{

//making Outfile
FILE *OUT;
OUT=fopen("metallopred_out.txt","w");
fclose(OUT);

//Inputting Descriptors
FILE *PAR;
PAR=fopen("par.xls","r");
double desc[37];
char r;
int i;
if(PAR == NULL)
        {
        printf("cannot open file");
        }
for(i=0;i<37;i++)
        {
        fscanf(PAR,"%lg",&desc[i]);
        }
fclose(PAR);
```

```
//sending to Layer 1
r=L1::pseaaL1(desc);
if(r=='M')
        {
        //sending to Layer 2
        r=L2::pseaaL2(desc);
        if(r=='A')
                {
                //sending to Layer 3 Alkali
                r=L3A::pseaaL3A(desc);
                }
        else if(r=='E')
                {
                //sending to Layer 3 Alkali Earth
                r=L3AE::pseaaL3AE(desc);
                if(r=='C')
                        {

                        }
                else if(r=='M')
                        {

                        }
                }
        else if(r=='T')
                {
                //sending to Layer 3 Transition
                r=L3T::pseaaL3T(desc);
                if(r=='1')
                        {
                        r=L4TG1::pseaaL4TG1(desc);
                        }
                else if(r=='2')
                        {
                        r=L4TG2::pseaaL4TG2(desc);
                        }
                else if(r=='3')
                        {
                        r=L4TG3::pseaaL4TG3(desc);
                        }
                }
        }

return 0;
}
```

Appendix III

Parser perl code:

Perl parser which links frontend with the descriptor calculation codes and prediction clusters

```
#!"C:/xampp/perl/bin/perl.exe"
#!"C:/xampp/perl/lib"

#prediction starter and output web page compiler

print "Content-type: text/html; charset=iso-8859-1\n\n";
print "<html>";
use CGI qw(:standard);
$pred=new CGI;
use FileHandle;
#taking in the sequence from the html page
$sequence=$pred->param("sequence");
#taking choice of parameters form html page
$pseaa=$pred->param("pseaa");
$aa=$pred->param("aa");
$pep=$pred->param("pepstat");
#checking for errors
#error type - no parameter
if(($pseaa ne "y") && ($aa ne "y") && ($pep ne "y"))
        {
        print "Error!!<br>No Parameter type selected... Go Back Again";
        goto end;
        }
#error type - no sequence
if(!$sequence)
        {
        print "Error!!<br>Sequence field empty... Go Back Again";
        goto end;
        }
#preparing input sequence file
print "<br>Input Sequence:<br>";
open(INP,"+>par.txt");
@seq=split(/[\n]/,$sequence);
$sequence="";
#removing Fasta comment line
if($seq[0] =~ /^>/)
        {
        print splice(@seq,0,1);
        }
```

```
#formatting sequence to be in a single line
$sequence=join("",@seq);
$sequence =~ tr/a-z/A-Z/;
@seq=();
@seq=split(//,$sequence);
$sequence="";
#removing any other exception (errors/non-standard aa) in the sequence
foreach $y(@seq)
        {
        if($y =~ /[ACDEFGHIKLMNPQRSTVWY]/)
                {
                $sequence=$sequence.$y;
                }
        }
print INP ">Query|PDBID|CHAIN|SEQUENCE\n$sequence";
print "<br>$sequence<br>";
close(INP);
#STARTING PREDICTION based on the choice of parameters(user given)
#firing predictor executers accordingly
if(($pseaa eq "y") && ($aa ne "y") && ($pep ne "y"))
        {
        system "modified_pseaa_desc_calc.exe";
        system "1_PSEAA_ANN.exe";
        }
elsif(($pseaa ne "y") && ($aa eq "y") && ($pep ne "y"))
        {
        `2_AA_aa_comp_desc_calc.pl`;
        system "2_AA_ANN.exe";
        }
elsif(($pseaa ne "y") && ($aa ne "y") && ($pep eq "y"))
        {
        `pepstats_calc.pl`;
        `3_PEP_pep_parser.pl`;
        system "3_PEP_ANN.exe";
        }
elsif(($pseaa eq "y") && ($aa eq "y") && ($pep ne "y"))
        {
        system "modified_pseaa_desc_calc.exe";
        `4_PSEAA_AA_aa_comp_desc_calc.pl`;
        system "4_PSEAA_AA_ANN.exe";
        }
elsif(($pseaa eq "y") && ($aa ne "y") && ($pep eq "y"))
        {
        system "modified_pseaa_desc_calc.exe";
        `pepstats_calc.pl`;
        `5_PSEAA_PEP_pep_parser.pl`;
```

```
        system "5_PSEAA_PEP_ANN.exe";
        }
elsif(($pseaa ne "y") && ($aa eq "y") && ($pep eq "y"))
        {
        `6_AA_PEP_aa_comp_desc_calc.pl`;
        `pepstats_calc.pl`;
        `6_AA_PEP_pep_parser.pl`;
        system "6_AA_PEP_ANN.exe";
        }
elsif(($pseaa eq "y") && ($aa eq "y") && ($pep eq "y"))
        {
        system "modified_pseaa_desc_calc.exe";
        `7_PSEAA_AA_PEP_aa_comp_desc_calc.pl`;
        `pepstats_calc.pl`;
        `7_PSEAA_AA_PEP_pep_parser.pl`;
        system "7_PSEAA_AA_PEP_ANN.exe";
        }
#printing the output
open(OUT,"metallopred_out.txt");
@output=<OUT>;
close(OUT);
`del par.xls`;
`del par.txt`;
if(glob("pepstat.xls")) {`del pepstat.xls`;}
`del metallopred_out.txt`;
print "<p align=\"center\"><h3>Output of MetalloPred</h3></p><br>";
foreach $y(@output)
        {
        print "$y<br>";
        }
end:
print "</html>";
```

Appendix IV

Pseudo amino acid c code:

C code for calculation of sequence derived features which preserves sequence order information

```
/* Pseudo Amino Acid Composition */

#include<stdio.h>
#include<string.h>
#include<stdlib.h>
#include<conio.h>
#include<fstream.h>
#include<iostream.h>
#include<math.h>

int pcount=0;
void getseq();
int aacheck(char);
float H1(int);
float H2(int);
float M(int);
float SD(float A[20]);
float avg(float A[20]);
float J(int,int);
void main()
{
        clrscr();
        getseq();
        cout<<"No of proteins in the file :"<<pcount;
        getch();
}

void getseq()
{
        char ch,file[15],file1[15]={0};
        cout<<"Enter the file containing the sequenecs :";
        cin>>file;
        ifstream infile(file);
        int v=0;
        while(file)
        {
                file1[v]=file[v];
                if(file[v]=='.')
                {
```

```
                file1[v+1]='x';
                file1[v+2]='f';
                file1[v+3]='s';
                break;
        }
        v++;
}
ofstream outfile(file1);
while(infile)
{
        infile.get(ch);
        if(ch=='>')
        {
                char pname[15]={0};
                int plength=0;
                int i=0;
                while(ch)
                {
                        infile.get(ch);
                        if(ch=='\n')
                        {
                                break;
                        }
                        if(ch=='|')
                                i++;
                        int j=0;
                        while(i==0)
                        {
                                        infile.get(ch);
                                        pname[j]=ch;
                                        j++;
                                        if(ch=='|')
                                                i++;
                        }
                }
                cout<<pname<<"\n";
                char seq[1800];
                int n=0;
                while(infile)
                {
                                infile.get(ch);
                                if(ch=='\n')
                                {
                                        infile.get(ch);
                                        if(ch=='\n')
                                                break;
```

```
                }
                seq[n]=ch;
                n++;
        }
        plength=n;
        int count[21]={0},f;
        for(i=0;i<plength;i++)
        {
                f=aacheck(seq[i]);
                count[f]=count[f]+1;
        }
        float arr[20],P[37];
        for(int j=0;j<20;j++)
        {
                arr[j]=(float)count[j]/plength;
        }
        float T[17];
        for(int r=1;r<18;r++)
        {
                float k=0.0;
                for(i=0;i<plength-r;i++)
                {
                        int e,f;
                        e=aacheck(seq[i]);
                        f=aacheck(seq[i+r]);
                        if(e!=20 && f!=20)
                                k=k+J(e,f);
                }
                T[r-1]=(k/(plength-r));
        }
        float t=0.0;
        for(i=0;i<17;i++)
                t=t+T[i];

        float g=0.0;
        for(i=0;i<20;i++)
        {
                g=g+arr[i];
        }
        float tmp=0.0;
        tmp=g+(0.5*t);
        for(i=0;i<20;i++)
        {
                P[i]=(arr[i]*100)/tmp;
        }
        for(i=0;i<17;i++)
```

```
                        {
                                P[20+i]=(0.5*T[i]*100)/tmp;
                        }
                        for(i=0;i<37;i++)
                        {
                                outfile<<P[i];
                                outfile<<'\t';
                        }
                        outfile<<'\n';
        pcount++;
                }
        }
}
int aacheck(char h)
{
        int a;
        if(h=='A')
                a=0;
        else if(h=='C')
                a=1;
        else if(h=='D')
                a=2;
        else if(h=='E')
                a=3;
        else if(h=='F')
                a=4;
        else if(h=='G')
                a=5;
        else if(h=='H')
                a=6;
        else if(h=='I')
                a=7;
        else if(h=='K')
                a=8;
        else if(h=='L')
                a=9;
        else if(h=='M')
                a=10;
        else if(h=='N')
                a=11;
        else if(h=='P')
                a=12;
        else if(h=='Q')
                a=13;
        else if(h=='R')
                a=14;
```

```
        else if(h=='S')
                a=15;
        else if(h=='T')
                a=16;
        else if(h=='V')
                a=17;
        else if(h=='W')
                a=18;
        else if(h=='Y')
                a=19;
        else
                a=20;
        return a;
}
float J(int x1,int x2)
{
        float j,k;
        k=(pow((H1(x2)-H1(x1)),2)+pow((H2(x2)-H2(x1)),2)+pow((M(x2)-M(x1)),2));
        j=k/3;
        return j;
}
float H1(int s)
{
        float H1[20]={0.62,0.29,-0.90,-0.74,1.19,0.48,-0.40,1.38,-1.50,1.06,0.64,-0.78,0.12,-
0.85,-2.53,-0.18,-0.05,1.08,0.81,0.26};
        float H;
        H=(H1[s]-avg(H1))/SD(H1);
        return H;
}
float H2(int s)
{
        float H2[20]={-0.5,-1.0,3.0,3.0,-2.5,0.0,-0.5,-1.8,3.0,-1.8,-1.3,0.2,0.0,0.2,3.0,0.3,-0.4,-
1.5,-3.4,-2.3};
        float H;
        H=(H2[s]-avg(H2))/SD(H2);
        return H;
}
float M(int s)
{
        float
M[20]={15.0,47.0,59.0,73.0,91.0,1.0,82.0,57.0,73.0,57.0,75.0,58.0,42.0,72.0,101.0,31.0,45.0,4
3.0,130.0,107.0};
        float m;
        m=(M[s]-avg(M))/SD(M);
        return m;
}
```

```
float SD(float A[20])
{
        float sd,a,s=0.0;
        a=avg(A);
        for(int i=0;i<20;i++)
                s=s+pow((A[i]-a),2);
        sd=sqrt(s/20);
        return sd;
}
float avg(float A[20])
{
        float avg,a=0.0;
        for(int i=0;i<20;i++)
                a=a+A[i];
        avg=a/20;
        return avg;
}
```

Appendix V

Amino acid perl code:

Perl code for calculation of sequence derived features based on amino acid composition

```
#Amino Acid Compostion Based Descriptors
#inputting file
print "\nInput filename (.txt):\t";
$filename=<>;
open (file,$filename)
        or print "cannot open sequence file";

#reading file into array
$i=0;
while(<file>)
        {
        if(/^>/)
                {
                $i++;
                $name[$i]=$_;
                }
        else
                {
                chomp($_);
                $seq[$i]=$seq[$i].$_;
                }
        }


#Reference array
$ref=(ACDEFGHIKLMNPQRSTVWY);
@ref=split(",$ref);

#output file open
print "\nenter output filename: ";
$out=<>;
open (desc,"+>$out");

#opening sequence and calculating frequency of amino acids
for($i=1;$i<$#name+1;$i++)
        {
        @pro=();
        @pro=split(",$seq[$i]);
        for($y=0;$y<$#ref+1;$y++)
                {
```

```
                $freq[$y]=0;
                }
        foreach $aa(@pro)
                {
                for($j=0;$j<$#ref+1;$j++)
                        {
                        if ($aa eq $ref[$j])
                                {
                                $freq[$j]+=1;
                                }
                        }
                }
        $proname=(split /[|]/,$name[$i])[0];
        print "protein: $proname\t@freq\n";
        print desc "$proname\t";
        for($k=0;$k<$#ref+1;$k++)
                {
                $probab=$freq[$k]/($#pro+1);
                if($freq[$k] eq 0)
                        {
                        $probab=0;
                        }
                print desc "$probab\t";
                }
print desc "\n";
}
```