International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol.4 Issue 9, September-2020, Pages: 43-50

ANN Model for Predicting Protein Localization Sites in Cells

Mohammed Nafez Abu Samra, Bilal Ezz El-Din Abed, Hossam Abdel Nasser Zaqout, Samy S. Abu-Naser

Department of Information Technology,
Faculty of Engineering and Information Technology,
Al-Azhar University, Gaza, Palestine

Abstract: *To automate examination of massive amounts of sequence data for biological function, it is important to computerize interpretation based on empirical knowledge of sequence-function relationships. For this purpose, we have been constructing an Artificial Neural Network (ANN) by organizing various experimental and computational observations as a collection ANN models. Here we propose an ANN model which utilizes the Dataset for UCI Machine Learning Repository, for predicting localization sites of proteins. We collected data for 336 proteins with known localization sites and divided them into training data and validating data. It was found that the accuracy rate for predicting Protein Localization Sites in Cells is 92.11%. This Indicates that Artificial Neural Network approach is powerful and flexible enough to be used in Protein Localization Sites prediction.*

Keyword: Prediction, Protein Localization Sites, Cells

INTRODUCTION

Computational approaches are becoming indispensable components of molecular and cellular biology, especially in the analyses of human and other complex genomes for which massive amounts of sequence data must be examined for biological function. Functional information can be obtained from sequence information not by solving equations of first principles, but by inference based on empirical knowledge. Although the sequences data are now collected and organized in publicly available databases, functional data are not well organized, except, perhaps, in the brain of a human expert.

The aim of this paper is to build a model that can accurately predict the Protein Localization Sites in Cells using Neural Network. Knowing a protein's localization helps elucidate its function, its role in both healthy processes and in the onset of disease and its potential use as a drug target. Experimental characterization of protein localization is accurate but slow and labor-intensive. However, the amino acid sequence of a protein usually provides crucial indication to its cellular localization sites. On the other hand, sequenced genomic data is experiencing an exponential increase in recent years due to maturation of High-Throughput sequencing techniques [1]. Thus, many computational methods have been developed to try to set up the link between a protein sequence and its cellular location. These include McGeoch's method for signal sequence recognition, discriminant analysis of the amino acid content of outer membrane and periplasmic proteins, etc. However, each of these methods can only deal with one protein category, i.e. giving the probability of a sequence being a membrane protein, or deciding whether it is a nucleus protein or not. Thus, for a new protein sequence on which people have no pre-knowledge, the only way to decide its localization site is to check all available methods to get a sense [2]. However, people still need to judge between these results to decide which method is more reliable, what is the cutoff probability for it to be safe to say a protein is in a certain cellular localization site but not in other sites. Thus, it is in a great need to develop a comprehensive system, integrating protein sequence-derived data and prediction results from the methods described above. It has been showed that a variety of machine learning methods can be used for this purpose[3,4].

In this paper, we are proposing an Artificial Neural Network for predicting a protein's subcellular localization site. The input should be vectors, each of which corresponds to a protein. Each atom in a vector is the result (score) obtained by running a certain computational method on this protein sequence. The output should be the predicted localization site. We collected the data from UCI Machine Learning Repository with 336 instances and 8 features, representing the kingdom of prokaryote. The actual localization sites of the proteins are already known.

ARTIFICIAL NEURAL NETWORK

Artificial Neural Networks are the computational models that are inspired by the human brain. Many of the recent advancements have been made in the field of Artificial Intelligence, including Voice Recognition, Image Recognition, Robotics using Artificial Neural Networks. Artificial Neural Networks are the biologically inspired simulations performed on the computer to perform certain specific tasks like – Clustering, Classification, and Pattern Recognition[5].

Artificial Neural Networks, in general – is a biologically inspired network of artificial neurons configured to perform specific tasks. These biological methods of computing are considered to be the next major advancement in the Computing Industry.

The term 'Neural' is derived from the human (animal) nervous system's basic functional unit 'neuron' or nerve cells that are present in the brain and other parts of the human (animal) body. A neural network is a group of algorithms that certify the underlying relationship in a set of data similar to the human brain. The neural network helps to change the input so that the network gives the best result without redesigning the output procedure.

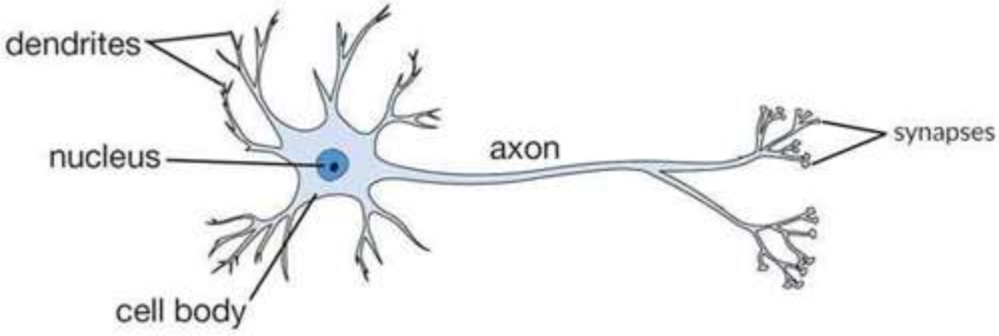


Figure 1: Biological Cell

A typical Artificial Neural Network contains a large number of artificial neurons called units arranged in a series of layers. In typical Artificial Neural Network, comprise different layers :

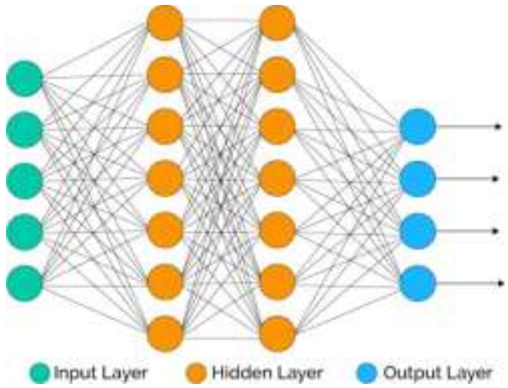


Figure 2: Typical architecture of ANN

- **Input layer** – It contains those units (Artificial Neurons) which receive input from the outside world on which the network will learn, recognize about or otherwise process.
- **Output layer –** It contains units that respond to the information about how it's learned any task.
- **Hidden layer –** These units are in between input and output layers. The job of the hidden layer is to transform the input into something that the output unit can use in some way.

Most Neural Networks are fully connected which means to say each hidden neuron is fully linked to every neuron in its previous layer(input) and to the next layer (output) layer.

Learning Techniques in Neural Networks includes:

- **Supervised Learning**

  In supervised learning, the training data is input to the network, and the desired output is known weights are adjusted until production yields desired value.

- **Unsupervised Learning**

  The input data is used to train the network whose output is known. The network classifies the input data and adjusts the weight by feature extraction in input data.

- **Reinforcement Learning**

  Here the value of the output is unknown, but the network provides feedback on whether the output is right or wrong. It is Semi-Supervised Learning.

- **Offline Learning**

  The adjustment of the weight vector and threshold is made only after all the training set is presented to the network. It is also called Batch Learning.

- **Online Learning**

  The adjustment of the weight and threshold is made after presenting each training sample to the network.

LITERATURE REVIEW

Artificial Neural Networks have been used many fields. In Education such as: Predicting Student Performance in the Faculty of Engineering and Information Technology using ANN, Prediction of the Academic Warning of Students in the Faculty of Engineering and Information Technology in Al-Azhar University-Gaza using ANN, Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach[5].

In the field of Health such as: Parkinson’s Disease Prediction, Classification Prediction of SBRCTs Cancers Using ANN, Predicting Medical Expenses Using ANN, Predicting Antibiotic Susceptibility Using Artificial Neural Network, Predicting Liver Patients using Artificial Neural Network, Blood Donation Prediction using Artificial Neural Network, Predicting DNA Lung Cancer using Artificial Neural Network, Diagnosis of Hepatitis Virus Using Artificial Neural Network, COVID-19 Detection using Artificial Intelligence[5].

In the field of Agriculture: Plant Seedlings Classification Using Deep Learning , Prediction of Whether Mushroom is Edible or Poisonous Using Back-propagation Neural Network, Analyzing Types of Cherry Using Deep Learning, Banana Classification Using Deep Learning, Mango Classification Using Deep Learning, Type of Grapefruit Classification Using Deep Learning, Grape Type Classification Using Deep Learning, Classifying Nuts Types Using Convolutional Neural Network, Potato Classification Using Deep Learning, Age and Gender Prediction and Validation Through Single User Images Using CNN[6].

In other fields such as : Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis: Statistical Methods, Predicting Overall Car Performance Using Artificial Neural Network, Glass Classification Using Artificial Neural Network, Tic-Tac-Toe Learning Using Artificial Neural Networks, Energy Efficiency Predicting using Artificial Neural Network, Predicting Titanic Survivors using Artificial Neural Network, Classification of Software Risks with Discriminant Analysis Techniques in Software planning Development Process, Handwritten Signature Verification using Deep Learning, Email Classification Using Artificial Neural Network, Predicting Temperature and Humidity in the Surrounding Environment Using Artificial Neural Network, English Alphabet Prediction Using Artificial Neural Networks[6].

Furthermore, in Protein Localization Sites in Cells such as: authors in [5] presented NNPS, an approach using artificial neural networks (ANNs) for predicting four eukaryotic (cytoplasmic, extracellular, mitochondrial, and nuclear) and three prokaryotic (cytoplasmic, extracellular, and periplasmic) subcellular localizations. Several alternative algorithms have been applied to the data set presented by Reinhardt and Hubbard, including Kohonen’s self-organizing maps, Support Vector Machines (SVMs), and Markov chain models.

The authors in [6] outlined the most comprehensive method based on Nterminal targeting sequences is TargetP, which allows for prediction of chloroplast, mitochondrial, secretory pathway, and other proteins. TargetP can be seen as an integration of the SignalP and the ChloroP methods.

Authors in [7] presented a method that assigns the subcellular localization by constructing phylogenetic profiles of the proteins. Authors in [8] discussed domains for predicting cytoplasmic, secreted, and nuclear proteins. The method PredictNLS is a method specialized on recognizing nuclear proteins, based on a collection of nuclear localization sequences (NLSs). A nearest neighbour approach using the composition of functional domains has also been presented and tested on the Reinhardt and Hubbard data set.

Authors in [9] outlined PSORT which was the first published program to predict subcellular localization. Subsequent tools and websites have been released using techniques such as artificial neural networks, support vector machine and protein motifs. Predictors can be specialized for proteins in different organisms. Some are specialized for eukaryotic proteins, some for human proteins and some for plant proteins. Methods for the prediction of bacterial localization predictors, and their accuracy, have been reviewed [10].

METHODOLOGY

Dataset of the Protein Localization Sites

We collected the dataset form UCI Machine learning repository [11]. The dataset consists of 336 samples with 9 attributes as can be seen in Table 1 and Table 2.

Table 1: Input Attributes of the dataset

#	Attribute	Meaning
1	Sequence Name	Accession number for the SWISS-PROT database
2	mcg	McGeoch's method for signal sequence recognition
3	gvh	von Heijne's method for signal sequence recognition
4	lip	von Heijne's Signal Peptidase II consensus sequence score. Binary attribute
5	chg	Presence of charge on N-terminus of predicted lipoproteins. Binary attribute
6	aac	score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins.
7	alm1	score of the ALOM membrane spanning region prediction program.
8	alm2	score of ALOM program after excluding putative cleavable signal regions from the sequence

Table 2: Output Attribute of the dataset

#	Class Abbreviation	Meaning	Class Distribution
1	cp	Cytoplasm	143 samples
2	im	Inner membrane without signal sequence	77
3	pp	Perisplasm	52
4	imU	Inner membrane, uncleavable signal sequence	35
5	om	Outer membrane	20
6	omL	Outer membrane lipoprotein	5
7	imL	Inner membrane lipoprotein	2
8	imS	Inner membrane, cleavable signal sequence	2

Data Preprocessing

Neural Network Prefer to work with numeric small numbers. The Output Class was transformed to numeric values ranging for 0 to 7. After that this category was normalized to be between 0 to 1.

Building the ANN Model

We have used Just Neural Network tool to build a multilayer ANN model. The proposed model consists of 4 Layers: Input Layer with 8 nodes, First Hidden Layer with 5 nodes, Second Hidden Layer with 3 nodes, and Output Layer with one node as can be seen in Figure 3.

We have sat the parameters of the proposed model as follows: Learning Rate 0.151 and the Momentum to be 0.136, and Average Error rate to be 0.01 (as shown in Figure 4).

Evaluating the ANN model

The Protein Localization Sites dataset consists of 336 samples with 9 attributes as Table 1 and Table 2. We imported the CSV file of the Protein Localization Sites dataset into the JNN environment (as seen in Figure 5). We divided the imported dataset into two groups (Training and Validation) randomly using the JNN tool. The Training consists of approximately 67% (222 samples) and the validation set consists of 33% of the dataset (114 samples). After making sure that the parameter control was sat properly, we started training the ANN model and keeping eye on the learning curve, loss error and validation accuracy. We kept training the ANN model for 79524 cycles. The best accuracy we got was 92.11% (as seen in Figure 6). We determined the most influential factors in the Protein Localization Sites Dataset as in Figure 7. Figure 8 shows the summary of the proposed model.

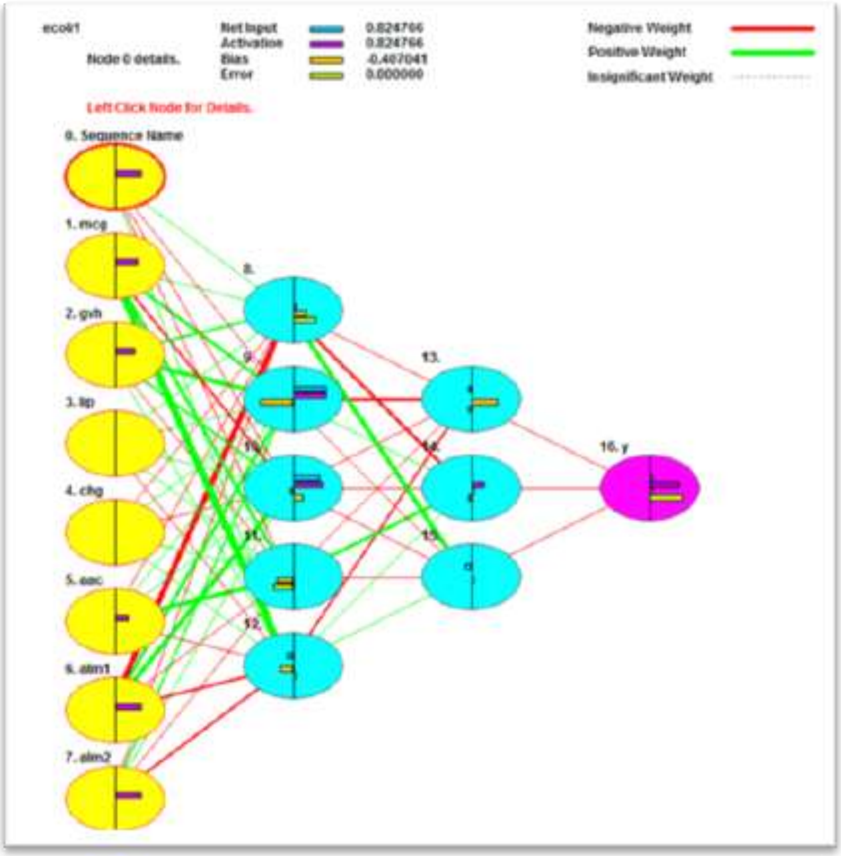
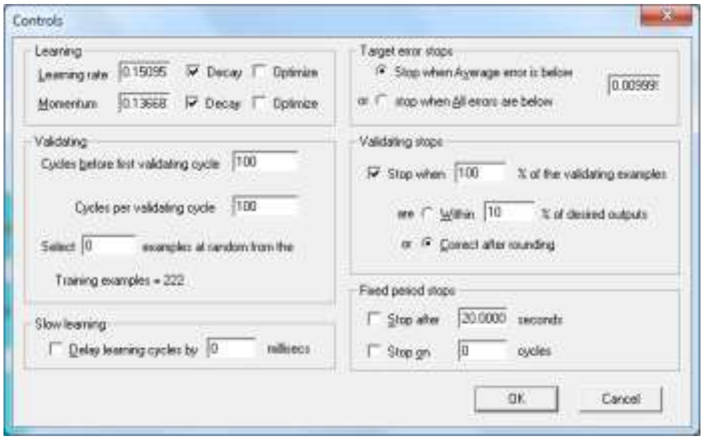


Figure 3: Architecture of Proposed ANN Model



Learning

Learning rate 0.15095 ☑ Decay □ Optimize

Momentum 0.13668 ☑ Decay □ Optimize

Target error stops

● Stop when Average error is below 0.00999

or ○ stop when All errors are below

Validating

Cycles before first validating cycle 100

Cycles per validating cycle 100

Select 0 examples at random from the

Training examples = 222

Validating stops

☑ Stop when 100 % of the validating examples

are ○ Within 10 % of desired outputs

or ● Correct after rounding

Fixed period stops

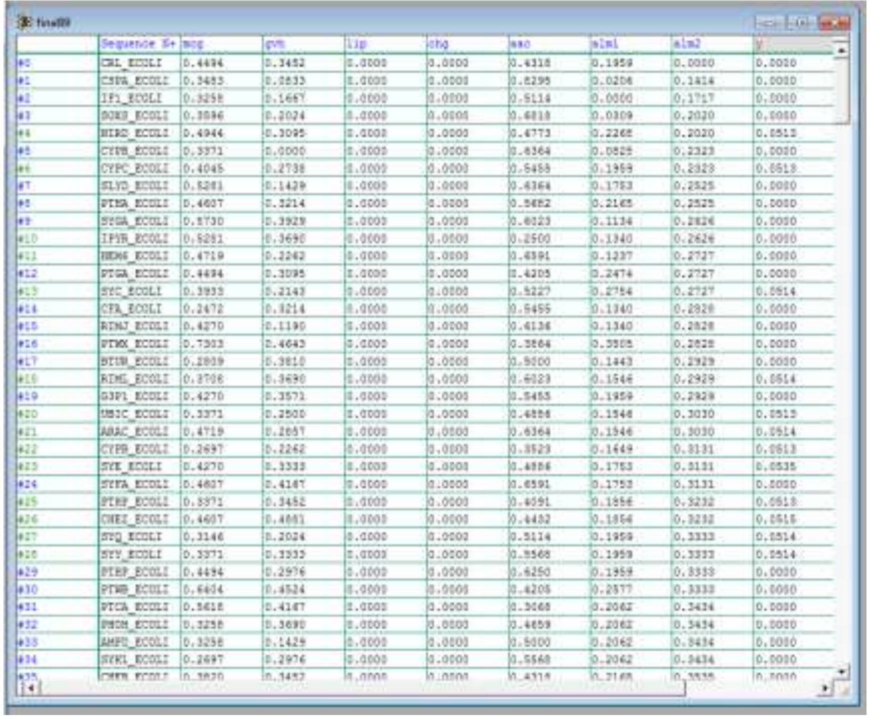
□ Stop after 20.0000 seconds

□ Stop on 0 cycles

Slow learning

□ Delay learning cycles by 0 millisecs

Figure 4: Control parameters of Proposed ANN model



	Sequence N+	mcg	gvh	lip	chg	aac	alm1	alm2	y
#0	CRL_ECOLI	0.4494	0.3452	0.0000	0.0000	0.4318	0.1959	0.0000	0.0000
#1	CSPA_ECOLI	0.3483	0.2833	0.0000	0.0000	0.6295	0.0206	0.1414	0.0000
#2	IF1_ECOLI	0.3258	0.1667	0.0000	0.0000	0.5114	0.0000	0.1717	0.0000
#3	SOXS_ECOLI	0.3596	0.2024	0.0000	0.0000	0.6818	0.0309	0.2020	0.0000
#4	HIRD_ECOLI	0.4944	0.3095	0.0000	0.0000	0.4773	0.2268	0.2020	0.0513
#5	CYPB_ECOLI	0.3371	0.0000	0.0000	0.0000	0.6364	0.0825	0.2323	0.0000
#6	CYPC_ECOLI	0.4045	0.2738	0.0000	0.0000	0.5455	0.1959	0.2323	0.0513
#7	SLYD_ECOLI	0.5281	0.1429	0.0000	0.0000	0.6364	0.1753	0.2525	0.0000
#8	PTHA_ECOLI	0.4607	0.3214	0.0000	0.0000	0.5682	0.2165	0.2525	0.0000
#9	SYGA_ECOLI	0.5730	0.3929	0.0000	0.0000	0.6023	0.1134	0.2626	0.0000
#10	IPYR_ECOLI	0.5281	0.3690	0.0000	0.0000	0.2500	0.1340	0.2626	0.0000
#11	HEM6_ECOLI	0.4719	0.2262	0.0000	0.0000	0.6591	0.1237	0.2727	0.0000
#12	PTGA_ECOLI	0.4494	0.3095	0.0000	0.0000	0.4205	0.2474	0.2727	0.0000
#13	SYC_ECOLI	0.3933	0.2143	0.0000	0.0000	0.5227	0.2784	0.2727	0.0514
#14	CFA_ECOLI	0.2472	0.3214	0.0000	0.0000	0.5455	0.1340	0.2828	0.0000
#15	RIMJ_ECOLI	0.4270	0.1190	0.0000	0.0000	0.6136	0.1340	0.2828	0.0000
#16	PTMX_ECOLI	0.7303	0.4643	0.0000	0.0000	0.3864	0.3505	0.2828	0.0000
#17	BTUR_ECOLI	0.2809	0.3810	0.0000	0.0000	0.5000	0.1443	0.2929	0.0000
#18	RIML_ECOLI	0.3708	0.3690	0.0000	0.0000	0.6023	0.1546	0.2929	0.0514
#19	G3P1_ECOLI	0.4270	0.3571	0.0000	0.0000	0.5455	0.1959	0.2929	0.0000
#20	UBIC_ECOLI	0.3371	0.2500	0.0000	0.0000	0.4886	0.1546	0.3030	0.0513
#21	ARAC_ECOLI	0.4719	0.2857	0.0000	0.0000	0.6364	0.1546	0.3030	0.0514
#22	CYPB_ECOLI	0.2697	0.2262	0.0000	0.0000	0.3523	0.1649	0.3131	0.0513
#23	SYE_ECOLI	0.4270	0.3333	0.0000	0.0000	0.4886	0.1753	0.3131	0.0535
#24	SYFA_ECOLI	0.4607	0.4167	0.0000	0.0000	0.6591	0.1753	0.3131	0.0000
#25	PTHP_ECOLI	0.3371	0.3452	0.0000	0.0000	0.4091	0.1856	0.3232	0.0518
#26	CHEZ_ECOLI	0.4607	0.4881	0.0000	0.0000	0.4432	0.1856	0.3232	0.0515
#27	SYQ_ECOLI	0.3146	0.2024	0.0000	0.0000	0.5114	0.1959	0.3333	0.0514
#28	SYY_ECOLI	0.3371	0.3333	0.0000	0.0000	0.5568	0.1959	0.3333	0.0514
#29	PTHP_ECOLI	0.4494	0.2976	0.0000	0.0000	0.6250	0.1959	0.3333	0.0000
#30	PTWB_ECOLI	0.6404	0.4524	0.0000	0.0000	0.4205	0.2577	0.3333	0.0000
#31	PTCA_ECOLI	0.5618	0.4167	0.0000	0.0000	0.3068	0.2062	0.3434	0.0000
#32	PNOM_ECOLI	0.3258	0.3690	0.0000	0.0000	0.4659	0.2062	0.3434	0.0000
#33	AMPD_ECOLI	0.3258	0.1429	0.0000	0.0000	0.5000	0.2062	0.3434	0.0000
#34	SYK1_ECOLI	0.2697	0.2976	0.0000	0.0000	0.5568	0.2062	0.3434	0.0000
#35	CHEB_ECOLI	0.3820	0.3452	0.0000	0.0000	0.4318	0.2168	0.3535	0.0000

Figure 5: Imported Localization Sites Dataset in JNN Environment

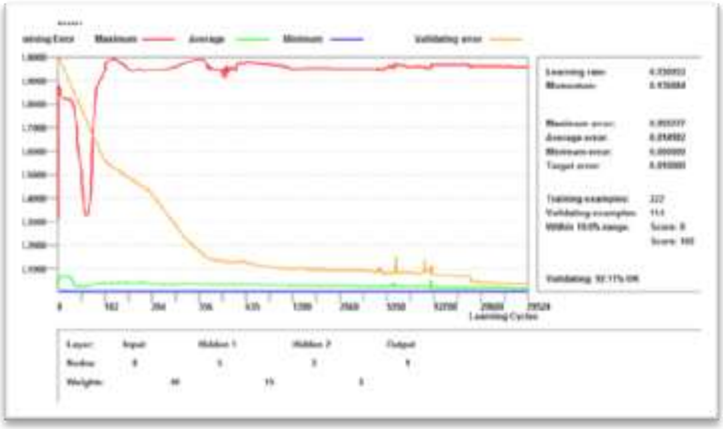
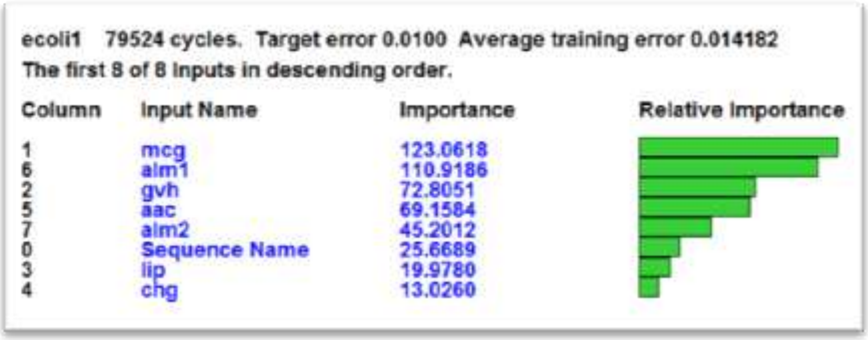


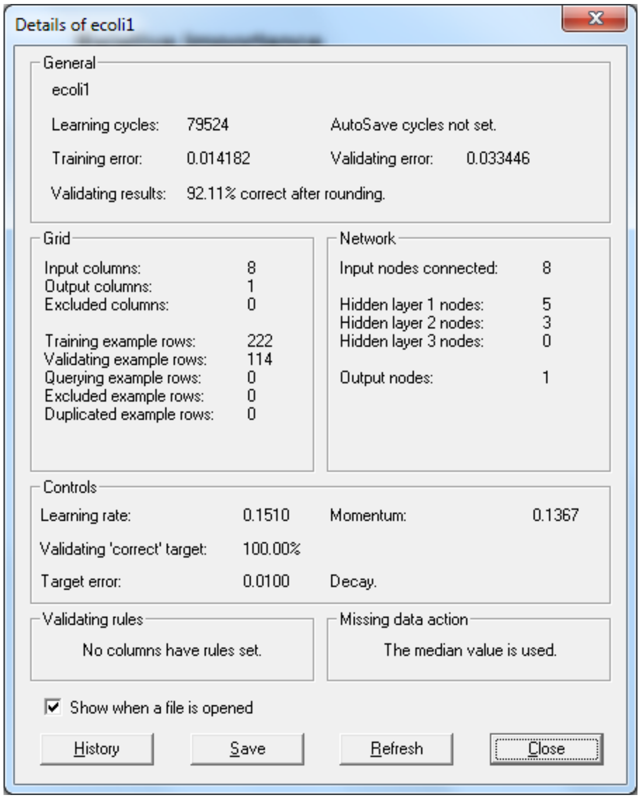
Figure 6: Training and validating the proposed ANN model in JNN environment



ecoli1 79524 cycles. Target error 0.0100 Average training error 0.014182
The first 8 of 8 inputs in descending order.

Column	Input Name	Importance	Relative Importance
1	mcg	123.0618	
6	alm1	110.9186	
2	gvh	72.8051	
5	aac	69.1584	
7	alm2	45.2012	
0	Sequence Name	25.6689	
3	lip	19.9780	
4	chg	13.0260	

Figure 7: Most influential factors in the Localization Sites Dataset



Details of ecoli1

General
ecoli1
Learning cycles: 79524 AutoSave cycles not set.
Training error: 0.014182 Validating error: 0.033446
Validating results: 92.11% correct after rounding.

Grid
Input columns: 8
Output columns: 1
Excluded columns: 0
Training example rows: 222
Validating example rows: 114
Querying example rows: 0
Excluded example rows: 0
Duplicated example rows: 0

Network
Input nodes connected: 8
Hidden layer 1 nodes: 5
Hidden layer 2 nodes: 3
Hidden layer 3 nodes: 0
Output nodes: 1

Controls
Learning rate: 0.1510 Momentum: 0.1367
Validating 'correct' target: 100.00%
Target error: 0.0100 Decay.

Validating rules
No columns have rules set.

Missing data action
The median value is used.

☑ Show when a file is opened

Figure 8: The summary of the proposed model

CONCLUSION

Proteins perform many important tasks in living organisms, such as catalysis of biochemical reactions, transport of nutrients, and recognition and transmission of signals. The plethora of aspects of the role of any particular protein is referred to as its “function.” One aspect of protein function that has been the target of intensive research by computational biologists is its subcellular localization. Proteins must be localized in the same subcellular compartment to cooperate toward a common physiological function. We proposed and ANN model for predicting Protein Localization Sites in Cells. We have collected the dataset Protein Localization Sites in Cells from UCI Machine Learning Repository. The outcome of training and validating the model is an accuracy rate of 92.11%.

References

1. Adams, et al. (1992). Sequence identification of 2,375 human brain genes. Nature 355: 632-634.

2. Baker, K. P., and Schatz, G. (1991). Mitochondrial proteins essentiall for viability mediate protein import into yeast mitochondria. Nature 349: 205-208.
3. Baranski, et. al. (1990). Generation of a lysosomal enzyme targeting signal in the secretory protein pepsinogen. Cell 63: 281-291.
4. Barker, et al. (1990). Protein sequence database. Methods Enzymol. 183: 31-49.
5. Klionsky, D., and Emr, S. D. (1990). A new class of lysosomal/vacuolar protein sorting signals. J. Biol. Chem. 265: 5349-5352.
6. Machamer, C. E., and Rose, J. K. (1987). A specific transmembrane domain of a coronavirus E1 glycoprotein is required for its retention in the Golgi region. J. Cell Biol. 105: 1205-1214.
7. McGeoch, D. J. (1985). On the predictive recognition of signal peptide sequences. Virus Res. 3: 271-286.
8. Nakai, K., and Kanehisa, M. (1988). Prediction of in-vivo modification sites of proteins from their primary structures. J. Biochem. (Tokyo) 104: 693-699.
9. Nakai, K., and Kanehisa, M. (1991). Expert system for predicting protein localization sites in Gram-negative bacteria. Proteins 11: 95-110.
10. Osumi, T., and Fujiki, Y. (1990). Topogenesis of peroxisomal proteins. BioEssays 12: 217-222.
11. UCI Machine Learning repository (https://archive.ics.uci.edu/ml/datasets.html)
12. EasyNN Tool.