Design and Implementation of a Neural Network For the Evaluation of Protein Secondary Structure Prediction Using the DSSP and Chou-Fasman Algorithms

Junger Tang

under the direction of
Dr. R. Mark Adams
Biomolecular Engineering Research Center
Boston University

Research Science Institute
August 2, 1995

Abstract

Neural networks have conventionally been used to predict protein secondary structure. However, they have not been used to improve the predictions of existing methods. This paper presents the design and implementation of a neural network to refine secondary structure prediction using information obtained from the DSSP and Chou-Fasman algorithms. The network was trained with input patterns consisting of an amino acid sequence and a secondary structure prediction and then tested with actual and random input patterns. A statistical analysis of the results suggests that the neural network achieved a greater accuracy using the predictions of the Chou-Fasman algorithm than those of the DSSP algorithm.

1 Introduction

A basic principle of molecular biology is that protein sequence determines protein structure [4]. Knowledge of a protein's amino acid sequence, called the primary structure, makes it possible to predict more complex levels of that protein's structure. Indeed, protein structures are organized in a hierarchy. At the most basic level is the primary structure. The next and most important level, the secondary structure, is mainly composed of alpha helices and beta strands, which are formed from local sequences of amino acids. Even more complex are the tertiary and quartenary structures, each of which is based on the elements of the structure preceding it in the hierarchy[2].

Knowing a protein's secondary structure helps to determine the structural properties of that protein. Several methods have been developed to determine secondary structure, with varying accuracy. One method involves analyzing the X-ray diffraction patterns of crystallized proteins. While X-ray diffraction is rather time-consuming, it is extremely accurate [13]. Another method, structure homology, or threading, utilizes an amino acid sequence with a known secondary structure as a model to predict the secondary structure of another similar sequence [8]. Various theoretical algorithms with high accuracies have also been proposed. Two of the most prominent are the DSSP and Chou-Fasman algorithms. The first algorithm determines secondary structure through knowledge obtained from the three-dimensional protein structure, such as hydrogen bonds and various geometrical features [9]. On the other hand, the Chou-Fasman algorithm predicts secondary structure by using many empirically determined rules in addition to information concerning the primary sequence [3].

A more recent and interesting approach to secondary structure prediction has been the

use of neural networks, which have been found to have respectable accuracy [13]. These information-processing systems consist of a large number of simple interconnected processing units that operate in parallel. All of these units are found in three different types of layers in the network: the input layer, the output layer, and in some cases, the hidden layers in between the input and output layers [6]. Each unit has an internal activation state which fluctuates according to the unit's input: excitatory input increases the activation, while inhibitory input decreases the activation [11]. By changing the activations of the units, neural networks are capable of learning by assimilating past inputs into the activations of each unit [14]. This capability has led to numerous applications in areas such as signal processing and pattern and speech recognition [6].

In this paper, we attempt to improve the predictive capabilities of neural networks for protein secondary structure with the use of predictions made by the DSSP and Chou-Fasman algorithms. With this intent, we evaluate the accuracy that the network achieves while using information obtained from either the DSSP or Chou-Fasman algorithms, in determining whether or not the secondary structure prediction of a given amino acid sequence is valid. The architecture of the network, which was constructed using the C programming language, is described and the results produced by the network are discussed and statistically analyzed.

2 Materials and Methods

2.1 Network Design

The neural network that was used in this investigation consists of a twelve-unit input layer and a one-unit output layer. No hidden layers were incorporated into the network due to the conclusion of Qian and Sejnowski [13] that the peak performance of their network in determining protein secondary structure was nearly independent of the number of hidden units. Furthermore, the network utilizes a feed-forward design, in which signals are transferred forward from the input units to the output unit[12].

Of the twelve units in the input layer, the first eleven units encode a window of eleven residues of an amino acid sequence, composed of five residues on either side of the central residue. The twelfth input represents the Chou-Fasman or DSSP prediction for the secondary structure of the central residue in the amino acid sequence. The output unit represents the prediction made by the neural network as to whether a given amino acid sequence and its secondary structure prediction are valid.

The activation state of each unit, X_i, is a real value between 0 and 1. The strength of the connection, or weight, between a unit j and another unit i is represented by a real number W_{ij}. The activation of a unit can be calculated by summing the products of every unit's output Y_j and weight W_{ij} and then adding a bias term, b_j:

$$X_i = \sum_j W_{ij} Y_j + b_j.$$

Having calculated the activation X_i for a unit, the output of that unit, Y_i can be computed using the logistic sigmoid function

$$Y_i = \frac{1}{1 + e^{-X_i}},$$

and then propagated to the next layer of the neural network.

During each cycle, the inputs are presented to the network. The weights of the units are adjusted at the end of the cycle, and this procedure is repeated. Backpropagation, a type of learning algorithm, is used to optimize the adjustment of the weights. This form of supervised training, in which the desired output is presented to the network along with the inputs [6], was used to train the neural network.

2.2 Network Training and Testing Sets

To train and test the neural network, the amino acid sequences of two globular proteins, flavodoxin and thioredoxin,[1] were obtained from the Protein Data Bank (PDB) at Brookhaven National Laboratory [1]. DSSP and Chou-Fasman computer programs were then used to analyze the amino acid sequences and predict the secondary structure of each residue that comprised the sequence.

The creation of input patterns for propagation through the network was accomplished by writing a computer program in the BASIC programming language that has several objectives. The program first parses the DSSP and Chou-Fasman secondary structure predictions and identifies the residues that have different predictions, of which both are one of the following:

[1] A listing of the amino acid sequences of flavodoxin and thioredoxin is in Appendix A.

α helix, β sheet, loop[2], or turn. Next, it takes each identified residue, the five residues on either side of it, and the prediction, and converts those twelve values using a numerical encoding scheme into a format understandable by the network. Tables 1 and 2 show the schemes that were used to convert the amino acids and secondary structure predictions into numerical formats.

Input	A	C	D	E	...	T	V	W	Y
Output	0.00	0.05	0.10	0.15	...	0.85	0.90	0.95	1.00

Table 1: Amino Acid Encoding Scheme

Input	α Helix	β Sheet	Loop	Turn
Output	0.20	0.40	0.60	0.80

Table 2: Secondary Structure Encoding Scheme

At this point, we have a comprehensive set of input patterns for a particular protein. Finally, the program creates random input patterns by arbitrarily selecting an element for each of the twelve inputs from all of the elements with the same input positions in the input patterns that have just been determined.

The training and testing sets were compiled after running this program on flavodoxin and thioredoxin. Testing both algorithms on two proteins required four different groups of training and testing sets, one for each combination of algorithm and protein. The two training sets from each protein were composed of the protein's first twenty valid input patterns followed by twenty random input patterns derived from the same protein. On the other hand, the two testing sets from each protein consisted of the protein's last ten valid input

[2]While bends are predicted by the DSSP algorithm and loops are predicted by the Chou-Fasman algorithm, the two structures are similar so that the term loop can be understood in this paper to represent either secondary structure.

patterns and ten random input patterns derived from that protein, in addition to the first ten valid input patterns and ten random input patterns derived from the other protein.

3 Results and Data

During each of the four tests, the neural network was trained for one hundred cycles using the appropriate training set before the predictions were made for the testing set. The results of the four tests are shown in Appendix B. To convert the floating-point numerical outputs of the network into definite answers, a threshold value t is calculated by computing the mean of all of the outputs of that particular test. This means that any input pattern with an output greater than or equal to t is considered valid by the network, while any input pattern with an output less than t is deemed invalid.

A summary of the results of Tables 4, 5, 6, and 7 is presented in Table 3. The neural network achieved a higher overall accuracy using the secondary structure predictions of the Chou-Fasman algorithm than those of the DSSP algorithm.

Test(s)	Number Correct	Percentage Correct
DSSP and flavodoxin	18 out of 40	45%
Chou-Fasman and flavodoxin	17 out of 40	42.5%
DSSP and thioredoxin	15 out of 40	37.5%
Chou-Fasman and thioredoxin	19 out of 40	47.5%
Combined results of DSSP and flavodoxin and thioredoxin	33 out of 80	41.25%
Combined results of Chou-Fasman and flavodoxin and thioredoxin	36 out of 80	45%

Table 3: Summary of Results

From these results, we formulated the hypothesis that using the secondary structure predictions of the Chou-Fasman algorithm is more effective than using those of the DSSP

algorithm. To determine if this hypothesis was statistically significant, we analyzed the results using a test of significance involving differences of proportions [15]. This test showed that the level of significance of the hypothesis is 0.3, which provides a 70% confidence level that it is correct.[3]

4 Discussion

The neural network in this study indicates that there is a noticeable difference in the effectiveness of using the secondary structure predictions of the Chou-Fasman algorithm in comparison to that of the predictions made by the DSSP algorithm. However, with a level of significance of 0.3, this finding cannot be considered statistically conclusive.

The source of this inconclusiveness is most likely found in the floating-point encoding schemes used to convert the letters representing the amino acids and secondary structure predictions into a format usable by the network. Sequentially assigning numbers to letters may provide the network with a tendency to consider two elements to be similar if their numerical equivalents are very close, even though the two elements may be dissimilar. For example, an α helix would not be considered more similar to a β sheet than a turn simply because the first two had numerical equivalents of 0.20 and 0.40 and the last had a numerical equivalent of 0.80.

We conducted another experiment to confirm this observation concerning the floating-point encoding scheme. A neural network similar to the one discussed in this paper was designed to recognize actual words given various sequences of six letters. As before, training

[3]For a mathematical derivation of this statistical result, see Appendix C.

and testing sets consisting of forty input patterns each were presented to the network. These sets of input patterns were composed of twenty actual words and twenty random sequences of six letters that had been converted into numerical format using the floating-point encoding scheme. The result was a 55% accuracy rate, which provides further evidence that the encoding scheme of the neural network is likely to be at fault.

Other studies have used binary encoding schemes, in which a group of inputs is assigned to each input unit in the input layer so that every element can be represented by either 0 or 1[8, 12, 13]. Another scheme, Grey coding, involves encoding inputs in a series of binary numbers. Both of these schemes eliminate the problem of interference resulting from elements that are in close proximity to each other numerically.

Several other improvements, besides a modification of the encoding scheme, could be made to improve the accuracy of the neural network. First, the length of the amino acid sequence in the input pattern could be extended, to account for secondary structure influences that exist at a greater distance from the central residue. Second, the number of input patterns in the training set could be increased, allowing the neural network to experience a larger variety of inputs. Finally, the output of the network could be refined by increasing the number of cycles that the network is programmed to undergo.

While neural networks have conventionally been designed to predict secondary structure [8, 12, 13], they have not been used to refine the predictions made by other methods. For that reason, much further study can be made in this area. Methods of protein secondary structure prediction other than the DSSP and Chou-Fasman algorithms could be evaluated, as well as the implementation of different types of neural networks in attaining those evaluations.

5 Conclusion

This study suggests that using the secondary structure predictions of the Chou-Fasman algorithm results in a greater accuracy for the neural network than using the predictions of the DSSP algorithm. This finding is very interesting, because the DSSP algorithm has traditionally been considered to be a standard in predicting secondary structure, even though it is not infallible.

Over the course of this investigation, we conducted four tests using the proteins flavodoxin and thioredoxin. Each test involved training the network with a set of input patterns of amino acid sequences and a secondary structure prediction, and then testing the network with a set of actual and random input patterns. The neural network achieved a 41.25% accuracy using the DSSP algorithm and a 45% accuracy with the Chou-Fasman algorithm. These results were analyzed using a level of significance statistical test, which showed that there is a 70% confidence level in our conclusion. The basis for this inconclusive result seems to be the floating-point numerical encoding scheme, which influenced the outputs produced by the network. This hypothesis was supported by another experiment that we conducted involving six-letter words. Methods of extending and improving this investigation were also discussed.

While not statistically conclusive, the results of this study do possess some utility to scientists and researchers who work with protein sequences and need to determine protein secondary structure. Further research that extends the results of this investigation is definitely needed.

6 Acknowledgments

I would especially like to thank my mentor Dr. R. Mark Adams for his invaluable guidance and support during the course of this research project, and for his kindness of always being willing to answer my questions. Also, I would like to thank Tom Graf, who provided much computer assistance in retrieving protein sequences from the PDB, using the DSSP and Chou-Fasman algorithms, and working in the UNIX environment. In addition, I would like to thank Robert Pacyga for his direction during the many stages of this research project. Finally, I would like to thank the Research Science Institute (RSI), Dr. Mark Saul, the director of RSI, the Center for Excellence in Education (CEE), and Joann P. DiGennaro, the president of CEE, for making it possible for me to conduct this research project at the Biomolecular Engineering Center (BMERC) at Boston University.

References

[1] Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. (1977) The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures, *Journal of Molecular Biology*, Vol. 112, pp. 535-542.

[2] Branden, Carl, and John Tooze. (1991) *Introduction to Protein Structure* (Garland Publishing, New York), pp. 11-31.

[3] Chou, Peter Y., and Gerald D. Fasman. (1974) Prediction of Protein Conformation, *Biochemistry*, pp. 222-245.

[4] Cohen, Bruce I., and Fred E. Cohen. (1994) Predictions of Protein Secondary and Tertiary Structure, *Biocomputing: Informatics and Genome Projects* (Academic Press, Boston), pp. 203-232.

[5] Creighton, Thomas E. (1993) *Proteins: Structures and Molecular Properties* (W. H. Freeman, New York), pp. 217-225.

[6] Fausett, Laurence. (1994) *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications* (Prentice-Hall, Englewood Cliffs, NJ).

[7] Harbison, Samuel P., and Guy L. Steele, Jr. (1987) *C: A Reference Manual, 2nd ed.* (Prentice-Hall, Englewood Cliffs, NJ).

[8] Holley, L. Howard and Martin Karplus. (1989) Protein Secondary Structure Prediction With a Neural Network, *Proceedings of the National Academy of Sciences (USA)*, Vol. 86, pp. 152-156.

[9] Kabsch, Wolfgang and Christian Sander. (1983) Dictionary of Secondary Structure Pattern Recognition of Hydrogen-Bonded and Geometrical Features, *Biopolymers*, Vol. 22, pp. 2577-2637.

[10] Kernighan, Brian W., and Dennis M. Ritchie. (1988) *The C Programming Language, 2nd ed.* (Prentice-Hall, Englewood Cliffs, NJ).

[11] Khanna, Tarun. (1990) *Foundations of Neural Networks*.

[12] Kneller, D. G., F. E. Cohen, and R. Langridge. (1990) Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network, *Journal of Molecular Biology*, Vol. 214, pp. 171-182.

[13] Qian, Ning, and Terrence J. Sejnowski. (1988) Predicting the Secondary Structure of Globular Proteins Using Neural Network Models, *Journal of Molecular Biology*, Vol. 202, pp. 865-884.

[14] Rumelhart, David E., and James L. McClelland. (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (MIT Press, Cambridge, MA), Vol. 1, pp. 45-76.

[15] Spiegel, Murray R. (1975) *Schaum's Outline of Theory and Problems of Probablility and Statistics* (McGraw-Hill, New York), pp. 211-220.

[16] Zar, Jerrold H. (1974) *Biostatistical Analysis* (Prentice-Hall, Englewood Cliffs, NJ).

Appendices

Appendix A

1 MKIVYWSGTG NTEKMAELIA KGIIESGKDV NTINVSDVNI DELLNEDILL

51 IGCSAMGDEV LEESEFEPFI EEISTKISGK KVALFGSYGW GDGKWMRDFE

101 ERMNGYGCVV VETPLIVQNE PDEAEQDCIE FGKKIANI

Amino acid sequence of the protein flavodoxin (PDB Code: 3fxn)

1 SDKIIHLTDD SFDTDVLKAD GAILVDFWAE WCGPCKMIAP ILDEIADEYQ

51 GKLTVAKLNI DQNPGTAPKY GIRGIPTLLL FKNGEVAATK VGALSKGQLK

101 EFLDANLA

Amino acid sequence of the protein thioredoxin (PDB Code: 2trx)

Appendix B

Tables 4, 5, 6, and 7 display the results of the four tests that were conducted using the neural network. For each input pattern[4] from the testing sets, the output from the network, the correct answer as to whether the input pattern is valid or not, the answer predicted by the network, and the matches of the two answers are displayed.

[4]For greater clarity and utility, the input patterns are not listed in numerical format.

Input Pattern	Network Output	Correct Answer	Predicted Answer	Match
ERMNGYGCVVVT	0.60	Valid	Valid	Y
RMNGYGCVVVET	0.40	Valid	Invalid	N
GCVVVETPLIVB	0.28	Valid	Invalid	N
VQNEPDEAEQDH	0.05	Valid	Invalid	N
QNEPDEAEQDCH	0.06	Valid	Invalid	N
NEPDEAEQDCIH	0.88	Valid	Valid	Y
EPDEAEQDCIEH	0.08	Valid	Invalid	N
PDEAEQDCIEFH	0.57	Valid	Valid	Y
DEAEQDCIEFGH	0.36	Valid	Invalid	N
EAEQDCIEFGKH	0.41	Valid	Invalid	N
KIIHLTDDSFDT	0.11	Valid	Invalid	N
IIHLTDDSFDTT	0.51	Valid	Valid	Y
IHLTDDSFDTDT	0.24	Valid	Invalid	N
HLTDDSFDTDVH	0.73	Valid	Valid	Y
LTDDSFDTDVLH	0.05	Valid	Invalid	N
TDDSFDTDVLKH	0.30	Valid	Invalid	N
DDSFDTDVLKAH	0.89	Valid	Valid	Y
DSFDTDVLKADT	0.64	Valid	Valid	Y
TDVLKADGAILB	0.89	Valid	Valid	Y
DVLKADGAILVB	0.40	Valid	Invalid	N
TVTFGPFVTEWT	0.43	Invalid	Valid	N
VLVKEGLGFGFH	0.19	Invalid	Invalid	Y
QGLADVDGQQPS	0.59	Invalid	Valid	N
GDATYAALDDWH	0.55	Invalid	Valid	N
IPQPKAKTKWGS	0.67	Invalid	Valid	N
ALAEGDIGVKLH	0.37	Invalid	Invalid	Y
KQKDGLPPLGYH	0.60	Invalid	Valid	N
QDFLIDTGIDGH	0.58	Invalid	Valid	N
FLPVDPDKDGGH	0.57	Invalid	Valid	N
AFNLLAFLYGWS	0.66	Invalid	Valid	N
GNGGGEEGPILH	0.28	Invalid	Invalid	Y
IDVSIKVLKVNH	0.80	Invalid	Valid	N
EVVCVCEDSQNH	0.23	Invalid	Invalid	Y
FKCEKVNEKLMT	0.06	Invalid	Invalid	Y
VNVIEGSKCSIH	0.70	Invalid	Valid	N
DGISYVGCVSKS	0.04	Invalid	Invalid	Y
GEYEIGINIILH	0.79	Invalid	Valid	N
IVDEEDIMLIWB	0.61	Invalid	Valid	N
GEIGVVAVIGQT	0.72	Invalid	Valid	N
DLEIAAVFYWLH	0.34	Invalid	Invalid	Y

Table 4: Results of using the DSSP predictions for flavodoxin. $t = 0.40$

Input Pattern	Network Output	Correct Answer	Predicted Answer	Match
ERMNGYGCVVVL	0.43	Valid	Valid	Y
RMNGYGCVVVES	0.65	Valid	Valid	Y
GCVVVETPLIVS	0.17	Valid	Invalid	N
VQNEPDEAEQDL	0.03	Valid	Invalid	N
QNEPDEAEQDCL	0.12	Valid	Invalid	N
NEPDEAEQDCIL	0.71	Valid	Valid	Y
EPDEAEQDCIEL	0.20	Valid	Invalid	N
PDEAEQDCIEFT	0.68	Valid	Valid	Y
DEAEQDCIEFGT	0.53	Valid	Valid	Y
EAEQDCIEFGKT	0.44	Valid	Valid	Y
KIIHLTDDSFDL	0.13	Valid	Invalid	N
IIHLTDDSFDTL	0.47	Valid	Valid	Y
IHLTDDSFDTDL	0.25	Valid	Invalid	N
HLTDDSFDTDVL	0.49	Valid	Valid	Y
LTDDSFDTDVLL	0.13	Valid	Invalid	N
TDDSFDTDVLKL	0.60	Valid	Valid	Y
DDSFDTDVLKAL	0.70	Valid	Valid	Y
DSFDTDVLKADL	0.09	Valid	Invalid	N
TDVLKADGAILT	0.59	Valid	Valid	Y
DVLKADGAILVT	0.24	Valid	Invalid	N
TVTFGPFVTEWH	0.40	Invalid	Invalid	Y
VLVKEGLGFGFS	0.24	Invalid	Invalid	Y
QGLADVDGQQPT	0.81	Invalid	Valid	N
GDATYAALDDWT	0.33	Invalid	Invalid	Y
IPQPKAKTKWGH	0.62	Invalid	Valid	N
ALAEGDIGVKLL	0.32	Invalid	Invalid	Y
KQKDGLPPLGYT	0.81	Invalid	Valid	N
QDFLIDTGIDGS	0.51	Invalid	Valid	N
FLPVDPDKDGGL	0.62	Invalid	Valid	N
AFNLLAFLYGWH	0.57	Invalid	Valid	N
GNGGGEEGPILS	0.51	Invalid	Valid	N
IDVSIKVLKVNL	0.57	Invalid	Valid	N
EVVCVCEDSQNH	0.23	Invalid	Invalid	Y
FKCEKVNEKLMH	0.05	Invalid	Invalid	Y
VNVIEGSKCSIL	0.76	Invalid	Valid	N
DGISYVGCVSKH	0.09	Invalid	Invalid	Y
GEYEIGINIILL	0.47	Invalid	Valid	N
IVDEEDIMLIWH	0.76	Invalid	Valid	N
GEIGVVAVIGQH	0.59	Invalid	Valid	N
DLEIAAVFYWLS	0.41	Invalid	Invalid	Y

Table 5: Results of using the Chou-Fasman predictions for flavodoxin. $t = 0.37$

Input Pattern	Network Output	Correct Answer	Predicted Answer	Match
GTAPKYGIRGIT	0.06	Valid	Invalid	N
TLLLFKNGEVAT	0.67	Valid	Valid	Y
LLFKNGEVAATS	0.34	Valid	Invalid	N
LFKNGEVAATKS	0.79	Valid	Valid	Y
FKNGEVAATKVS	0.17	Valid	Invalid	N
KNGEVAATKVGS	0.53	Valid	Valid	Y
NGEVAATKVGAS	0.64	Valid	Valid	Y
KVGALSKGQLKH	0.09	Valid	Invalid	N
VGALSKGQLKEH	0.21	Valid	Invalid	N
GALSKGQLKEFH	0.36	Valid	Invalid	N
WSGTGNTEKMAH	0.14	Valid	Invalid	N
ELIAKGIIESGH	0.43	Valid	Valid	Y
LIAKGIIESGKH	0.22	Valid	Invalid	N
IAKGIIESGKDH	0.40	Valid	Valid	Y
AKGIIESGKDVH	0.39	Valid	Valid	Y
KGIIESGKDVNT	0.52	Valid	Valid	Y
GIIESGKDVNTT	0.55	Valid	Valid	Y
VNTINVSDVNIH	0.12	Valid	Invalid	N
NTINVSDVNIDH	0.03	Valid	Invalid	N
TINVSDVNIDEH	0.35	Valid	Invalid	N
GNGGGEEGPILH	0.27	Invalid	Invalid	N
IDVSIKVLKVNH	0.48	Invalid	Valid	Y
EVVCVCEDSQNH	0.21	Invalid	Invalid	N
FKCEKVNEKLMT	0.32	Invalid	Invalid	N
VNVIEGSKCSIH	0.59	Invalid	Valid	Y
DGISYVGCVSKS	0.07	Invalid	Invalid	N
GEYEIGINIILH	0.50	Invalid	Valid	Y
IVDEEDIMLIWB	0.47	Invalid	Valid	Y
GEIGVVAVIGQT	0.23	Invalid	Invalid	Y
DLEIAAVFYWLH	0.58	Invalid	Valid	N
TVTFGPFVTEWT	0.32	Invalid	Invalid	Y
VLVKEGLGFGFH	0.43	Invalid	Valid	N
QGLADVDGQQPS	0.37	Invalid	Valid	N
GDATYAALDDWH	0.26	Invalid	Invalid	Y
IPQPKAKTKWGS	0.61	Invalid	Valid	N
ALAEGDIGVKLH	0.39	Invalid	Valid	N
KQKDGLPPLGYH	0.32	Invalid	Invalid	Y
QDFLIDTGIDGH	0.60	Invalid	Valid	N
FLPVDPDKDGGH	0.10	Invalid	Invalid	Y
AFNLLAFLYGWS	0.55	Invalid	Valid	N

Table 6: Results of using the DSSP predictions for thioredoxin. $t = 0.42$

Input Pattern	Network Output	Correct Answer	Predicted Answer	Match
GTAPKYGIRGIL	0.03	Valid	Invalid	N
TLLLFKNGEVAS	0.50	Valid	Valid	Y
LLFKNGEVAATH	0.28	Valid	Invalid	N
LFKNGEVAATKH	0.73	Valid	Valid	Y
FKNGEVAATKVH	0.11	Valid	Invalid	N
KNGEVAATKVGH	0.40	Valid	Valid	Y
NGEVAATKVGAH	0.41	Valid	Valid	Y
KVGALSKGQLKT	0.21	Valid	Invalid	N
VGALSKGQLKET	0.34	Valid	Invalid	N
GALSKGQLKEFT	0.57	Valid	Valid	Y
WSGTGNTEKMAL	0.15	Valid	Invalid	N
ELIAKGIIESGS	0.64	Valid	Valid	Y
LIAKGIIESGKS	0.31	Valid	Invalid	N
IAKGIIESGKDS	0.54	Valid	Valid	Y
AKGIIESGKDVL	0.59	Valid	Valid	Y
KGIIESGKDVNL	0.37	Valid	Invalid	N
GIIESGKDVNTL	0.51	Valid	Valid	Y
VNTINVSDVNIS	0.26	Valid	Invalid	N
NTINVSDVNIDS	0.04	Valid	Invalid	N
TINVSDVNIDES	0.48	Valid	Valid	Y
GNGGGEEGPILS	0.37	Invalid	Invalid	Y
IDVSIKVLKVNL	0.69	Invalid	Valid	N
EVVCVCEDSQNH	0.31	Invalid	Invalid	Y
FKCEKVNEKLMH	0.22	Invalid	Invalid	Y
VNVIEGSKCSIL	0.70	Invalid	Valid	N
DGISYVGCVSKH	0.06	Invalid	Invalid	N
GEYEIGINIILL	0.67	Invalid	Valid	N
IVDEEDIMLIWH	0.42	Invalid	Valid	N
GEIGVVAVIGQH	0.14	Invalid	Invalid	Y
DLEIAAVFYWLS	0.78	Invalid	Valid	N
TVTFGPFVTEWH	0.15	Invalid	Invalid	Y
VLVKEGLGFGFS	0.50	Invalid	Valid	N
QGLADVDGQQPT	0.40	Invalid	Valid	N
GDATYAALDDWT	0.41	Invalid	Valid	N
IPQPKAKTKWGH	0.45	Invalid	Valid	N
ALAEGDIGVKLL	0.48	Invalid	Valid	N
KQKDGLPPLGYT	0.61	Invalid	Valid	N
QDFLIDTGIDGS	0.74	Invalid	Valid	N
FLPVDPDKDGGL	0.09	Invalid	Invalid	Y
AFNLLAFLYGWH	0.46	Invalid	Valid	N

Table 7: Results of using the Chou-Fasman predictions for thioredoxin. $t = 0.42$

Appendix C

Using the predictions of the Chou-Fasman algorithm, the neural network was correct 36 out of 80 times. When the network used the predictions made by the DSSP algorithm, it was correct 33 out of 80 times.

Let P_1 and P_2 denote respectively the proportions of correct answers of the Chou-Fasman and DSSP algorithms: $P_1 = \frac{36}{80} = .45$ and $P_2 = \frac{33}{80} = .41$. Furthermore, let n_1 and n_2 denote the sample sizes, which are both 80 in this case. We begin by calculating the average proportion p of correct answers: $p = \frac{36+33}{80} = .43$. Let $q = 1$ - $p = .57$.

Now we determine the standard deviation, σ_s, of the difference of proportions, P_1 - P_2:

$$\sigma_{P_1-P_2} = \sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})} = \sqrt{(.43)(.57)(\frac{1}{80} + \frac{1}{80})} \approx .078.$$

The next step is to calculate the standardized variable Z:

$$Z = \frac{P_1 - P_2}{\sigma_{P_1-P_2}} = \frac{.45 - .41}{.078} \approx .51$$

Finally, we use a table of one-tailed proportions of the normal curve [16], which gives the level of significance given a standardized variable, Z. We find that the level of significance of our hypothesis is approximately .3.