# NN-PRED: A Novel Consensus Secondary Structure Prediction Program Using Neural Networks

# NN-PRED: Un Nuevo Programa para la Predicción de la Estructura Secundaria de la Proteína Usando Redes Neuronales

***OSCAR FERNANDO BEDOYA LEIVA***
*Magíster en Ingeniería de Sistemas*
*Docente de la Escuela de Ingeniería de Sistemas y Computación*
*Universidad del Valle*
*oscar.bedoya@correounivalle.edu.co*
*Cali, Colombia*

***EDUARD ALBERTO SATIZÁBAL TASCÓN***
*Ingeniero de Sistemas*
*Universidad del Valle*
*eduard__satizabal@hotmail.com*
*Cali, Colombia*

*Fecha de recibido: 01/03/2012*
*Fecha de aprobación: 09/03/2013*

## ABSTRACT

In this paper, a new program for protein secondary structure prediction is proposed. The program, which is called NN-Pred, allows multiple sequences to be submitted and it returns predictions from five secondary structure prediction algorithms. In addition, NN-Pred calculates a consensus prediction, which is based on a neural network strategy that is used in this paper to improve the prediction accuracy. NN-Pred was obtained by using a methodology called consensus strategy, which tries to make a better prediction by integrating some of the most remarkable existing techniques. The NN-Pred program provides a three-state (alpha-helix, beta-sheet, and other) prediction of secondary structure. According to the test sets, the prediction accuracy of NN-Pred is at least 70%, surpassing most of the existing methods. The experiments showed that neural networks can be used as a consensus strategy to producing accurate models for protein secondary prediction.

**KEYWORDS**: secondary structure prediction, consensus strategy, neural networks.

## RESUMEN

En este artículo se propone un nuevo programa para la predicción de la estructura secundaria de la proteína. El programa, llamado NN-Pred, recibe como entrada múltiples secuencias de ADN y utiliza cinco algoritmos existentes para la predicción de la estructura secundaria de la proteína. Además, NN-Pred calcula una predicción consenso que se basa en una estrategia de redes neuronales y que se plantea en este artículo para mejorar la exactitud en la predicción. NN-Pred se obtuvo usando una metodología conocida como estrategia consenso que intenta obtener un modelo de predicción integrando algunos de los mejores métodos existentes. El programa NN-Pred provee una predicción de tres estados (hélices alfa, hojas beta, y otro) para la estructura secundaria de la proteína. De acuerdo a los resultados de las pruebas realizadas, NN-Pred alcanza una exactitud de predicción de al menos 70.0%, sobrepasando la mayoría de los métodos existentes. Los experimentos realizados mostraron que la técnica de redes neuronales se puede usar como una estrategia consenso para obtener modelos precisos para la predicción de la estructura secundaria.

**PALABRAS CLAVE**: predicción de la estructura secundaria, estrategia consenso, redes neuronales.

# 1. INTRODUCTION

## 1.1 Secondary structure prediction

Secondary structure prediction addresses the problem of assigning a class label (i.e., alpha-helix, beta-sheet, turn or coil) to each residue in a given amino acid sequence. Although several strategies have been proposed to secondary structure prediction, most of the methods take the primary structure as input and try to predict the secondary structure based on the amino acid sequence. The secondary structure prediction problem has been studied for almost forty years [1]-[5] and even recent methods [6]-[10] are not able to reach the prediction accuracy demanded by biologists.

Several solutions for the secondary structure prediction problem have been proposed, and they use different approaches such as statistical, neural networks, and nearest-neighbour methods. The statistical approach is based on trying to calculate the probability of an amino acid to be a specific secondary structure by using sequences whose secondary structure is known. Chou-Fasman [1] is one the most representative programs that uses the statistical approach. Some other prediction programs are based on neural networks such as PHD [4], NNPredict [11], and Psipred [12]. These three programs are inspired by artificial neural networks theory, which establishes that a set of computational neurons is able to learn nonlinear deterministic relationships into the training dataset. Finally, one of the most successful approaches is the nearest-neighbour method. It takes the information from the k nearest amino acids whose secondary structure is trying to be predicted. Some of the most remarkable programs following this approach are SOPM [13], SOPMA [14], and PREDATOR [15].

The main efforts in the secondary structure prediction problem are related to raising the prediction accuracy of the existing methods. There is an import issue in secondary structure prediction; the prediction of the existing methods do not coincide for some residues of many amino acid sequences (i.e., there is no consensus in the existing methods). Having a different prediction even for the same sequence can confuse biologists in real-life situations. For instance, the secondary structure prediction for the sequence TMIPAV obtained by the Chou-Fasman method is HHHEEN, which means residues T, M, I are classified as helix (H); P, A as sheets (E), and V as neither a helix nor a sheet (N). The same amino acid sequence was given to the DSC [16] program and the resulting secondary structures were HHENEE. As can be seen, different classifications were obtained for the same residues I, P, and V.

In this paper, a novel secondary structure prediction program is proposed. The program uses a consensus strategy achieved by using a neural network that is trained to predict the secondary structure of amino acids. There is a major difference between this work and the existing methods; the way of integrating different predictor programs proposed in this paper had never been used. In addition, there is an improvement in accuracy by using the new strategy of combining predictors.

## 1.2 Consensus strategy to classify biological sequences

The consensus strategy has been applied in several prediction problems in bioinformatics [17]-[20]. The hypothesis behind a consensus strategy establishes that a better prediction model can be built by combining several experts rather than using a single one. Each predictor program is considered an expert. The consensus strategy follows the idea of running several prediction programs with the same input and comparing or integrating the outputs to make a better decision than by using individual methods. The consensus decision can be as easy as the majority wins criteria in which the decision taken by most of the n given experts is the consensus decision. However, a consensus decision can be taken based on a more complex model (i.e., a decision tree or a neural network). One of the consensus strategies is called ensemble methods [21]-[23]. Ensemble methods try to integrate the result of different individual experts by using a combination of simple learners. Another consensus strategy is called mixture of experts. Unlike ensemble methods, a mixture of experts model try to cover different input regions with different learners to obtain a better overall accuracy.

Neural networks have been used as a mixture of expert strategy in bioinformatics. In [24], a neural network is proposed as a mixture-of-expert model to integrate ten secondary structure prediction programs. The model obtained has a higher accuracy in comparison with each of the ten individual experts. In [25], a decision tree is proposed to predict genes by combining the outputs of three gene predictors: GeneMark [26], GlimmerM [27], and GenScan [28]. The consensus model outperforms even the best individual method.

In this paper, a new program for secondary structure prediction, called NN-Pred, is proposed. NN-Pred makes decisions based on mixing five experts. Each expert is a well-known secondary structure prediction program. The five experts used in this paper were selected by considering the availability of the source code. The experts selected were Chou-Fasman [1], GOR [29], DSC [16], SIMPA96 [30], and PREDATOR [15], which exhibit an accuracy of 60%, 65%, 70.1%, 67.7%, and 68%, respectively.

There is a major difference between the work presented in this paper and some methods based on mixture of experts that use neural networks such as [24] and [31]. The information used to feed the neural network has a different approach. The current mixture-of-expert methods integrate the output of n given individual predictors. However, the output values integrated are the secondary structure labels (i.e., alpha-helix, beta-sheet, turn or coil). A different approach is used in this paper and it is related to taking the scores calculated by each individual method as the values to be mixed. For instance, when Chou-Fasman predicts a single amino acid as a-helix, instead of taking the predicted label a-helix as the value to integrate, we propose to use the score calculated by the method. In this paper, the integration of experts is done by using the numerical values of the prediction scores instead of the discrete values of the prediction labels. This approach is not as rigid as the current strategies used for mixture of experts in secondary structure prediction.

### 1.3 Evaluating predictions

There are several measures to evaluate the accuracy in the secondary structure prediction problem. The three-state per-residue accuracy ($Q_3$) is frequently used to evaluate and compare secondary structure prediction methods. $Q_3$ (1) gives the percentage of residues predicted correctly for alpha-helix ($q_a$), beta-strand ($q_b$), and other ($q_c$) from the total number of residues (N) in a given amino acid sequence. The $Q_3$ measure indicates the percentage of correctly predicted residues and is defined as follows:

$$Q_3 = \left[\frac{(q_a + q_b + q_c)}{N}\right] \cdot 100 \qquad (1)$$

Another measure that is commonly used is the Matthew Correlation Coefficient (2) [32]. It is a real number between 0 and 1 and depends on the number of true positive ($T_p$), true negative ($T_n$), false positive ($F_p$), and false negative ($F_n$) predicted residues. The Matthew Correlation Coefficient is defined as follows:

$$C = \frac{T_p T_n - F_p F_n}{\sqrt{(T_p + F_p)(T_p + F_n)(T_n + F_p)(T_n + F_n)}} \qquad (2)$$

## 2. METHODOLOGY

### 2.1 Selecting and preparing the datasets

A neural network needs to be trained to be able to learn the information included in patterns. Patterns are the training data, which include the input/output values. During the training process, each pattern is presented to the neural network, which tries to adjust its weights to get the corresponding output in the pattern. After the training process, a set of weights are obtained in such a way that feeding the inputs of each pattern produces the expected output.

The selection of patterns in the training set is a major decision because the generalization capability of the model can be affected. A test set is also needed, which is a dataset with data not included in the training set that is used to measure the accuracy during prediction. The datasets used in this work were composed of 300 proteins. Proteins were extracted from the PDB (Protein Data Bank, http://www.pdb.org/pdb/home/home.do) and include domains from 20 different families. Table 1 shows the datasets used in this work.

**Table 1**. *Datasets*

| Dataset reference | Amount of helixes | Amount of sheets | Amount of neither helix nor sheet residues |
|---|---|---|---|
| dataset 1 | 150 | 150 | 100 |
| dataset 2 | 120 | 120 | 120 |
| dataset 3 | 270 | 270 | 270 |
| dataset 4 | 149 | 108 | 121 |

A total of 80% of each dataset was used as the training set and 20% as the test set. Both training and test datasets were submitted to the DSSP program [33] to obtain the actual secondary structure. All the amino acid sequences were also submitted to five structure prediction programs, those to be mixed by the neural network. This paper only considered the a-helix, b-sheet and "other" as the secondary structure labels. The DSSP structures 3-10 helix (G), pi helix (I), short beta bridge (B), bend (S), Turn (T), and Coil (C) were not included in this work.

### 2.2 Secondary structure prediction programs

The secondary structure prediction programs used in this paper were five well-known methods, Chou-Fasman, GOR, DSC, SIMPA96, and PREDATOR. An explanation of each program is presented as follows.

- Chou-Fasman [1]. It is based on a table whose values indicate the probability that a specific amino acid is an alpha-helix or a beta-sheet. The values are used to make predictions and are associated with an algorithm that indicates the way those values should be used to make the prediction. The prediction accuracy of the Chou-Fasman method ranges from 50% to 60%.

- GOR [29]. It uses a window of 17 amino acids to predict the structure of the residue at position 9 of the window (i.e., the residue in the center of the window). GOR uses two scoring tables for each residue in the window of 17 amino acids, one for alpha-helix and another for beta-sheet. The prediction accuracy of the GOR method is 65%.

- DSC [16]. DSC (Discrimination of Secondary structure Class) is a secondary structure prediction method that uses a window size of 17 amino acids. DSC integrates statistical criteria such as the conformation of each residue and some other criteria based on concepts related to folds. The concepts used by DSC are length of the amino acid at the end of the chain, hydrophobic moments for alpha-helix and beta-sheet, and conservation moments for alpha-helix and beta-sheet. The prediction accuracy of DSC is 70.1%.

- PREDATOR [15]. It is a method for secondary structure prediction based on the strategy of immediate neighbours. PREDATOR predicts the secondary structure of an amino acid analysing the structure of its neighbours. It uses a database of sequences with known secondary structure for short segments of amino acids. In addition, PREDATOR uses statistics to predict pairs of amino acid hydrogen bonds between neighbours. The prediction accuracy of PREDATOR is 68% for individual sequences and 75% for sets of related sequences.

- SIMPA96 [30]. It is a method based on three elements: a database of proteins with known secondary structure, a scoring matrix, and a prediction algorithm. The prediction accuracy of the SIMPA method is 67.7%.

Some of these methods such as DSC, PREDATOR, and SIMPA96 are available via FTP public servers. Some other algorithms such as Chou-Fasman and GOR were developed because there was no source code available.

### 2.3 Preparing the neural network input

Some specific tasks were performed to prepare the data to feed the neural network:

- The source code of the five secondary structure prediction programs were modified to obtain the score of each individual method. The source code was studied and the specific lines in which the score is calculated were identified. It allowed having a numeric value for the secondary structure prediction of each residue.

- Each secondary structure prediction program calculates different values to make a final decision. For instance, Chou-Fasman calculates two values, the sheet score and the helix score. These two scores are calculated using tables proposed by Chou and Fasman and correspond to the probabilities of each amino acid to be a particular secondary structure. Each predictor calculates different values; we decided to take them all as inputs for a neural network. The values obtained from the predictors are presented in Table 2.

**Table 2**. *Values obtained from the secondary structure predictors*

| Secondary structure prediction program | Value | Description |
|---|---|---|
| GOR | GOR_E | Sheet propensity calculated by GOR |
| | GOR_H | Helix propensity calculated by GOR |
| CHOU-FASMAN | CHOU_E | Sheet propensity calculated by CHOU-FASMAN |
| | CHOU_H | Helix propensity calculated by CHOU-FASMAN |
| DSC | DSC_E | Sheet propensity calculated by DSC |
| | DSC_H | Helix propensity calculated by DSC |
| | DSC_N | Propensity to neither sheet nor helix structure calculated by DSC |
| PREDATOR | PREDATOR_E | Sheet propensity calculated by PREDATOR |
| | PREDATOR_H | Helix propensity calculated by PREDATOR |
| | PREDATOR_N | Propensity to neither sheet nor helix structure calculated by PREDATOR |
| SIMPA | SIMPA_E | Sheet propensity calculated by SIMPA |
| | SIMPA_H | Helix propensity calculated by SIMPA |
| | SIMPA_N | Propensity to neither sheet nor helix structure calculated by SIMPA |

- According to the amount of secondary structure prediction programs selected and the values calculated by each of them, a total of 13 scores can be obtained for the same amino acid.

- The training set was submitted to the new version of the prediction algorithms to obtain the scores for each amino acid. Besides, the actual secondary structure was assigned using the DSSP program.

The set composed of the 13 scores corresponds to the input that feeds the neural network. The secondary structure assigned by DSSP is taken as the expected output of the neural network. According to the artificial intelligence theory, a neural network is able to learn from these input/output values to predict the secondary structure of future unseen amino acid sequences.

## 2.4 Obtaining the neural network

A neural network was designed to learn the training dataset and eventually predict secondary structure for unseen residues. The network was composed of three layers (input, hidden, and output). The input layer has 13 units; each unit corresponding to a score calculated by the prediction programs. The hidden layer is composed of five units. The number of neurons in the hidden layer was obtained by using the BIC criterion (Bayesian Information Criterion). The output layer has three neurons (a-helix, b-sheet, and "other"). The output of the network is 1 0 0 when the input corresponds to a helix, 0 1 0 for a sheet, and 0 0 1 when the input is neither a helix nor a sheet. Figure 1 shows the topology of the neural network.

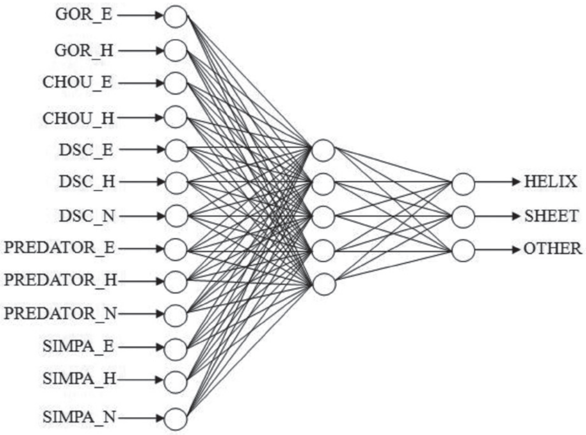


**Figure 1**. *Consensus neural network*

The backpropagation learning algorithm was used. The number of iterations in the learning process was optimized to 200 to save time without losing the learning performance of the network. A logistic function was used as the activation function. In the training process, the whole proteins in the training set were presented to the neural network.

The neural network can be used as a secondary structure predictor by feeding the inputs with the 13 values of a given uncharacterised amino acid sequence. Although there are many secondary structure predictors following the consensus strategy, none of them uses numerical values from the experts as inputs to the neural network.

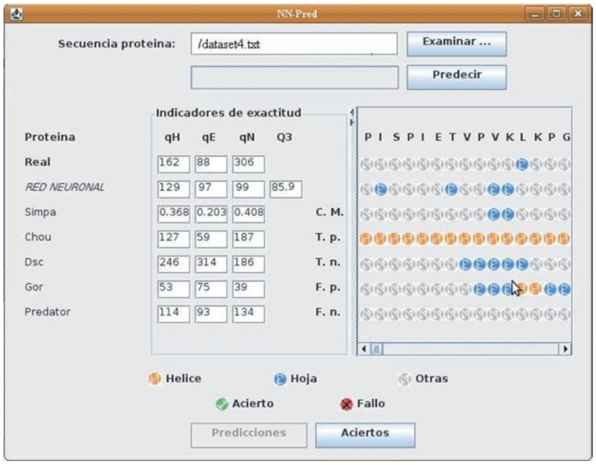
## 2.5 The NN-Pred program

Once the neural network is obtained, a new strategy to predict protein secondary structure is achieved. The neural network was implemented as a program that allows users to submit an amino acid sequence and make the prediction based on the consensus strategy proposed in this paper. The program is called NN-Pred and it is available for academic purposes from the authors upon request.

### *2.5.1 Case study*

A user can input a single sequence or multiple sequences. The NN-Pred program allows running five secondary structure prediction programs: Chou-Fasman, GOR, SIMPA96, PREDATOR, and DSC. In addition, there is an option that includes the consensus decision. In the case study, dataset 1 was used to train and validate the neural network. As can be seen in Figure 2, the user can easily compare the results from all of the predictors. When the consensus prediction is selected, the neural network model is used to calculate the output. As an advantage, the program allows users to compare results with the DSSP output values. The program also presents the DSSP assignment for each residue. Finally, the $Q_3$, $C_H$, $C_E$, and $C_N$ values for each individual prediction program and the consensus decision are also presented. NN-Pred has an easy-to-use interface, which is a helpful tool for comparing the secondary structure prediction programs included in this version of the software.

In this case study, the $Q_3$ measure was calculated to compare the individual methods and the consensus strategy. The structure assigned by DSSP program is taken as the actual secondary structure. Table 3 shows the $Q_3$ values. The values presented for the existing methods were calculated for the specific case study and thus, they may differ from the theoretical accuracy.



NN-Pred

Secuencia proteina: /dataset4.txt

Indicadores de exactitud

| Proteina | qH | qE | qN | Q3 |
|---|---|---|---|---|
| Real | 162 | 88 | 306 | |
| *RED NEURONAL* | 129 | 97 | 99 | 85.9 |
| Simpa | 0.368 | 0.203 | 0.408 | |
| Chou | 127 | 59 | 187 | |
| Dsc | 246 | 314 | 186 | |
| Gor | 53 | 75 | 39 | |
| Predator | 114 | 93 | 134 | |

**Figure 2**. *NN-Pred in a case study*

The highest accuracy is reached by the NN-Pred consensus strategy. As can be seen in Table 3, the individual methods are not as accurate as the neural network proposed in this paper.

**Table 3**. *Accuracy values in the case study*

| Method | $Q_3$ |
|---|---|
| GOR | 42.3% |
| CHOU-FASMAN | 47.3% |
| DSC | 52.1% |
| PREDATOR | 63.5% |
| SIMPA | 66.1% |
| NN-Pred | 82.1% |

## 3. RESULTS

In a first experiment, the accuracy of the consensus neural network included in the NN-Pred program was obtained using the dataset 2. Dataset 2 is composed of 360 residues distributed in 120 helixes, 120 sheets, and 120 amino acids whose secondary structures were neither helix nor sheet. The experiment can be carried out following the methodology explained in Section 2. The source codes of the five structure prediction programs (i.e., GOR, Chou-Fasman, DSC, PREDATOR, and SIMPA) have to be modified. Each predictor calculates a set of internal values before a final prediction is made. Table 2 can be used to define the specific values to be obtained from each predictor. Then, the topology shown in Figure 1 has to be achieved. A neural network simulator can be used to obtain a network with thirteen (13) neurons in the input layer, five (5) neurons for the hidden layer, and three (3) neurons for the output layer. The logistic activation function and the backpropagation learning algorithm should be used. Finally, every residue in the dataset 2 has to be input for the individual predictors in order to obtain the 13 values. These values are used to feed the neural network. When the learning process is completed, a set of weights are obtained, which are used to predict the secondary structure of unseen residues.

The accuracy values of the consensus strategy using the dataset 2 are shown in Table 4. The $Q_3$ and C values reach 100% in training. The previous result shows that the topology used is appropriate for secondary structure prediction. Besides, a test set was presented to the network by using the weights calculated in training. The $Q_3$ and C values for the test set are also shown in Table 4. According to the results, unseen sequences reduce the accuracy of the model. However, the $Q_3$ and C values are still acceptable compared with some of the existing methods.

**Table 4.** *Accuracy values using dataset 2*

| Dataset 2 | $Q_3$ | C | | |
|---|---|---|---|---|
| | | $C_H$ | $C_E$ | $C_N$ |
| Training set | 100% | 100% | 100% | 100% |
| Test set | 71.8% | 50.7% | 47.0% | 55.3% |

A second experiment was related to calculating the capability of the consensus neural network to learn the dataset 3. In this case, a set of 810 amino acids was used. The $Q_3$ and C values are shown in Table 5. $Q_3$ of 100% and C value of 100% were obtained in training. A test set was presented to the network by using the weights calculated in training. The $Q_3$ and C values for the test set are also shown in Table 5.

**Table 5.** *Accuracy values using dataset 3*

| Dataset 3 | $Q_3$ | C | | |
|---|---|---|---|---|
| | | $C_H$ | $C_E$ | $C_N$ |
| Training set | 100% | 100% | 100% | 100% |
| Test set | 80.2% | 75.7% | 70.5% | 72.5% |

Another experiment was related to comparing the individual methods with the consensus neural network. Dataset 4 was used in this experiment. Dataset 4 is composed of 378 residues distributed in 149 helixes, 108 sheets, and 121 amino acids whose secondary structures were neither helix nor sheet. $Q_3$ and C values are shown in Table 6. These values were also calculated for the individual methods (GOR, Chou-Fasman, DSC, PREDATOR, and SIMPA).

**Table 6**. *Comparison of $Q_3$ values*

| Method | $q_H$ | $q_E$ | $q_N$ | $Q_3$ |
|---|---|---|---|---|
| GOR | 95 | 60 | 22 | 31.8% |
| CHOU-FASMAN | 152 | 18 | 10 | 32.3% |
| DSC | 93 | 36 | 189 | 57.1% |
| PREDATOR | 129 | 46 | 170 | 62.0% |
| SIMPA | 113 | 39 | 193 | 62.0% |
| NN-Pred | 129 | 97 | 99 | 85.9% |

NN-Pred achieves a $Q_3$ of 85.9%, which is clearly higher than the best individual expert (i.e. PREDATOR or SIMPA). The overall performance of NN-Pred was more accurate than the individual experts. In general, accuracy of secondary structure prediction is improved drastically when numeric scores, instead of the discrete values, are used. Neural networks showed a high generalization capability, which is usually better than some other classifying techniques (i.e., decision tree or Bayesian networks). Tests allowed finding out that neural networks are suitable as a consensus strategy for secondary structure prediction.

The prediction model proposed in this paper is a consensus neural network. The strategy is based on training a neural network as a model that integrates the output of five secondary structure predictors. The prediction accuracy of the model was tested using four datasets. It can be seen that the accuracy of the model during the prediction phase achieves a $Q_3$ value that was at least 71.8%, which is a high threshold for the secondary structure prediction problem. The prediction model is able to classify unseen amino acids in three classes: helixes, sheets, and secondary structures that are neither helix nor sheet. Finally, Table 6 shows that the prediction model based on a consensus neural network surpasses the accuracy of the individual experts, a result that might expected when a consensus strategy is used.

## 4. CONCLUSIONS

The objective of this study was to obtain a neural network capable of classifying the secondary structure of the protein. The NN-Pred program includes a consensus neural network formed by three layers with 13 nodes in the input layer, five neurons in the hidden layer, and three units in the output. The neural network showed perfect accuracy values during the training process, which means that the topology used was appropriate for the problem of classifying secondary structures of a protein. In the test sets, the $Q_3$ value was at least 71.8%, which is a high threshold for the secondary structure prediction problem. The $Q_3$ values obtained in the experiments are shown in Figure 3. According to the $Q_3$ values, it can be seen that the neural network is capable of identifying a high portion of the helixes and sheets out of the total number of amino acids in the test set.

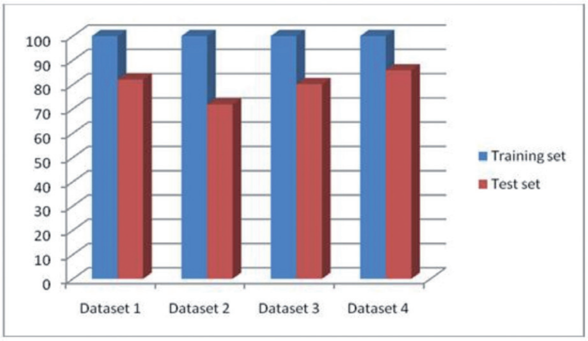


**Figure 3**. *Summary of $Q_3$ values*

The main difference between the existing methods for secondary structure prediction and the strategy proposed in this paper is the way of integrating different predictor programs. In this paper, numerical values are used instead of discrete labels. Analysing Table 6, there is a large gain when numeric scores are considered to make a consensus strategy. As a result, a fair recommendation would be to consider the values behind individual predictors when a consensus strategy is going to be used.

There is also a major achievement in this work; the NN-Pred is capable of learning to guarantee 100% accuracy. It means a biologist could train the network with a particular dataset of interest and use it being absolutely sure the program is correct.

The NN-Pred program allows users to interact with an easy-to-use interface, which can graphically compare the results of both individual methods and the consensus. The program also presents accuracy values to carry a fair comparison of the methods. NN-Pred could be improved by adding some other experts or predictors whose source code is available. Besides, numerical values from individual experts could also be integrated by using support vector machines or some other classification techniques.

## 5. REFERENCES

[1] P. Chou and G. Fasman, “Prediction of protein conformation,” *Biochemistry,* vol. 13, no. 2, 1974, pp. 222-245.

[2] V. Lim, “Structural principles of the globular organisation of protein chains. A stereochemical theory of globular protein secondary structure,” *Journal of Molecular Biology,* vol. 88, no. 1, 1974, pp. 857-872.

[3] J. Garnier, D. Osguthorpe and B. Robson, “Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins,” *Journal of Molecular Biology*, vol, 120, no 1, 1978, pp. 97-120.

[4] B. Rost and C. Sander, “Improved prediction of protein secondary structure by use of sequence profiles and neuronal networks,” *Proceedings of the National Academy of Science*, vol. 90, no. 1, 1990, pp. 7558-7562.

[5] B. Rost, “PHD: predicting one-dimensional protein structure by profile based neural networks,” *Methods Enzimol*, vol. 266, no. 1, 1996, pp. 525-539.

[6] C. Cole, J. Barber and G. Barton, “The JPred 3 secondary structure prediction server,” *Nucleic Acids Res*, vol. 36, no. 1, 2008, pp. 197-201.

[7] M. Osman, M. Abdullah and R. AbdulRashid, “RNA secondary structure prediction using dynamic programming algorithm - A review and proposed work,” *Information Technology*, vol. 2, 2010, pp. 551-556.

[8] D. Mojie, Z. Yanhong, H. Huiyan, “A Protein Secondary Structure Prediction Tool Using Two-Level Strategy to Improve the Prediction Accuracy of Secondary Structures and Structure Boundaries,” *Information Engineering and Computer Science*, vol. 1, 2009, pp.1-4.

[9] H. Tsang and K. Wiese, “SARNA-Predict: Accuracy Improvement of RNA Secondary Structure Prediction Using Permutation-Based Simulated Annealing,” *Computational Biology and Bioinformatics*, vol. 7, no. 4, 2010, pp. 727-740.

[10] B. Tang, X. Wang and X. Wang, “Protein Secondary Structure Prediction Using Large Margin Methods,” *Computer and Information Science*, vol. 1, 2009, pp. 142-146.

[11] D. Kneller, F. Cohen and R. Langridge, “Improvements in protein secondary structure prediction by an enhanced neural network,” *Journal of Molecular Biology*, vol. 214, no. 1, 1990, pp. 171-182.

[12] D. Jones, “Protein secondary structure prediction based on position-specific scoring matrices,” *Journal of Molecular Biology, vol.* 292, 1999, pp. 195-202.

[13] C. Geourjon and G. Deleage, “SOPM: a self-optimized method for protein secondary structure prediction,” *Protein Eng*, vol. 7, no. 2, 1994, pp. 157-164.

[14] C. Geourjon and G. Deleage, “SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments,” *Comput. Appl. Biosci*, vol. 1, no. 6, 1995, pp. 681-684.

[15] D. Frishman and P. Argos, “75% accuracy in protein secondary structure prediction,” *Proteins*, vol. 27, 1997, pp. 329-335.

[16] R. King and M. Sternberg, “Identification and application of the concepts important for accurate and reliable protein secondary structure prediction,” *Protein Science*, vol. 5, 1996, pp. 2298-2310.

[17] A. Thomas and Y. Zheng, “Improved Prediction of HIV-1 Protease Genotypic Resistance Testing Assays using a Consensus Technique,” *Neural Networks*, vol. 1, 2006, pp. 2308-2314.

[18] Y. Zhao and W. Zhengzhi, ”Consensus RNA Secondary Structure Prediction Based on SVMs,“ *Bioinformatics and Biomedical Engineering*, vol. 1, 2008, pp. 101-104.

[19] C. Mazo and O. Bedoya, “PESPAD: una nueva herramienta para la predicción de la estructura secundaria de la proteína basada en árboles de decisión,” *Ingeniería y Competitividad*, vol. 12, no. 2, 2010, pp. 9-22.

[20] Q. Wang, Y. Shang and D. Xu, “Protein structure selection based on consensus,” *Evolutionary Computation,* vol. 1, 2010, pp.1-7.

[21] M. Siek, “Nonlinear multi-model ensemble prediction using dynamic Neural Network with incremental learning,” *The 2011 International Joint Conference on Neural Networks* (IJCNN), vol. 1, 2011, pp. 2873-2880.

[22] Liyu Lin, Shuanqiang Yang, and Ruijuan Zuo, “Protein secondary structure prediction based on multi-SVM ensemble,” *International Conference on Intelligent Control and Information Processing* (ICICIP), vol. 1, 2010, pp. 356-358.

[23] Bidargaddi, N., Chetty M., and Kamruzzaman, J, “An Architecture Combining Bayesian segmentation and Neural Network Ensembles for Protein Secondary Structure Prediction,” *Proceedings of the 2005 IEEE Symposium on*

*Computational Intelligence in Bioinformatics and Computational Biology* (CIBCB), vol. 1, 2005, pp. 1-8.

[24] J. De Haan and J. Leunissen, "Protein Secondary Structure Prediction: Comparison of Ten Common Prediction Algorithms Using a Neural Network," *Nato Science Series Sub Series I Life and behavioural sciences,* vol. 368, 2005, pp. 149-161.

[25] J. Allen, M. Pertea and S. Salzberg, "Computational Gene Prediction Using Multiple Sources of Evidence," *Genome Res*, vol. 14, 2004, pp. 142-148.

[26] A. Lukashin and M. Bordovsky, "GeneMark.hmm: New solutions for gene finding," *Nucleic Acids Res*, vol. 26, 1998, pp. 1107-1115.

[27] M. Pertea and S. Salzberg, "Computational gene finding in plants," *Plant Mol. Biol*, vol. 48, 2002, pp. 39-48.

[28] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA," *J. Mol. Biol*, vol. 268, 1997, pp. 78-84.

[29] J. Garnier, J. Gibrat and B. Robson, "GOR method for predicting protein secondary structure from amino acid sequence," *Methods in Enzymology*, vol. 266, 1996, pp. 540-553.

[30] J. Levin, "Exploring the limits of nearest neighbour secondary structure prediction," *Protein Engineering*, vol. 7, 1997, pp. 771-776.

[31] C. Combet, C. Blanchet, C. Geourjon and G. Deléage, "NPS@: Network Protein Sequence Analysis,“ *TIBS, v*ol. 25, no. 3, 2000, pp. 147-150.

[32] B. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica Biophysica*, vol. 405, 1975, pp. 442-451.

[33] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, 1983, pp. 2577-637.