Using classifier fusion techniques for protein secondary structure prediction

Majid Kazemian and Behzad Moshiri

Control and Intelligent Processing Center,
School of Electrical and Computer Engineering,
University of Tehran, Tehran, Iran
E-mail: akazemi2@uiuc.edu E-mail: moshiri@ut.ac.ir

Vasile Palade*

Computing Laboratory, Oxford University,
Parks Road, Oxford, OX1 3QD, UK
E-mail:vasile.palade@comlab.ox.ac.uk
*Corresponding author

Hamid Nikbakht

Laboratory of Biophysics and Molecular Biology,
Institute of Biochemistry and Biophysics,
University of Tehran, Tehran, Iran
E-mail: h.nikbakht@gmail.com

Caro Lucas

Control and Intelligent Processing Center,
School of Electrical and Computer Engineering,
University of Tehran, Tehran, Iran
E-mail: lucas@ipm.ir

Abstract: Classifier fusion techniques are gaining more popularity for their capability of improving the accuracy achieved by individual classifiers. A common approach is to combine the classifiers' outcome using simple methods, such as majority voting. In this paper, we build a meta-classifier by fusing some already well-known classifiers for protein structure prediction. Each individual classifier outputs a unique structure for every input residue. We have used the confusion matrix of each protein secondary structure classifier, which is representative of classifiers' expertness, as a general reusable pattern for converting its simple class-label assignment to class-preference score. The results obtained using several classifier fusion operators have been compared, on some standard datasets from the EVA server, with simple majority voting and with the results provided by the individual classifiers. The comparative analysis showed that the Choquet fuzzy integral operator had the highest improvement with respect to accuracy, multi-class sensitivity and specificity criteria over both the best performing individual classifier and the other fusion operators, while all of the classifier fusion techniques yielded some improvements too.

Keywords: protein secondary structure prediction; classifier fusion; Choquet fuzzy integral operator; meta-classifier; confusion matrix; multi-class sensitivity.

Reference to this paper should be made as follows: Kazemian, M., Moshiri B., Palade, V., Nikbakht H., Lucas, C. (2010) 'Using classifier fusion techniques for protein secondary structure prediction', *Int. J. Computational Intelligence in Bioinformatics and Systems Biology*, Vol. 1, No. 4, pp.418–434.

Biographical notes: Majid Kazemian is currently a PhD candidate in Computer Science at the University of Illinois at Urbana-Champaign. He received his MSc in Robotics and Machine Intelligence from the University of Tehran, Iran, in 2007. His research interests include bioinformatics, DNA sequence analysis and data mining.

Behzad Moshiri is currently a Full Professor at the School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran. He has been the head of Machine Intelligence & Robotics division of this school. He received his MSc and PhD from UMIST, UK, in 1987 and 1991, respectively. He is a Senior Member of IEEE since 2006. He is the author and co-author of more than 270 articles including 65 journal papers and several book chapters. His research interests include advanced industrial control design, advanced instrumentation design, sensor data fusion, intelligent transportation systems, mechatronics and bioinformatics.

Vasile Palade is currently working with the Computing Laboratory, Oxford University, UK. He obtained his PhD in Intelligent Systems from the University of Galati, Romania, in 1999. His research interests are in the area of computational intelligence with application to bioinformatics, fault diagnosis, web usage mining, among others. He published more than 70 papers in journals and conference proceedings as well as several books. He is an IEEE Senior Member.

Hamid Nikbakht is currently doing his PhD in Biology (Bioinformatics) at Concordia University, Montreal, Canada. He is studying the evolution of genomic nucleotide content. He received his Masters in Cell and Molecular Biology from University of Tehran in 2005. His research interests include biological sequence analysis, protein structure (prediction/design), simulating evolutionary forces and studying their effects on genomes.

Caro Lucas is currently a Full Professor at the School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran. He received his PhD in Control Systems Engineering from the University of California, Berkeley, in 1976 and his MS in Electrical Engineering from the School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran in 1973. He has served as an Editorial Board Member for several engineering journals. His research interests cover a broad range in machine learning and intelligent control system and design.

1 Introduction

Prediction of protein structures is one of the main challenges in theoretical chemistry and bioinformatics today. It refers to finding the protein's three-dimensional (tertiary) structure from its amino acid sequence (primary structure). Accurate predictions are essential in medicine and biotechnology to help design novel drugs and discover new enzymes. Because of the dramatic advances in high throughput sequencing technologies, the gap between the number of known protein sequences [~5 million (UniProt, 2008)] and their resolved structures [~50 k (Berman et al., 2000)] is rapidly increasing, despite significant improvements in structure resolving methods (Bairoch and Apweiler, 1999). As an initial step in predicting the tertiary structure, researchers usually examine the secondary structural composition of the protein.

Protein secondary structures are recurring patterns formed by interaction between (neighboring) residues. They primarily consist of three patterns: alpha helices (H), beta strands (E) and coils (C). Since the patterns are determined by the properties of the contextual residues, we can think of the properties as features and of the secondary structures as classes, and frame the secondary structure prediction as a classification problem.

Although there are many protein secondary structure prediction software publicly available today (Rost et al., 2004; Karplus et al., 2003; Pollastri and McLysaght, 2005; Argos et al., 1978; Cai et al., 2003; Kim, 2004), it might be difficult and completely impractical to try to improve the performance of a single classifier over a certain limit in solving a complex problem. Rather, the solution may be found in the combination of existing reasonably performing classifiers, with the aim of improving the overall classification result. Different classifiers performing on different parts of the input space may implement different aspects of the problem (Kazemian et al., 2007), and assuming enough coverage of the input space by individual classifiers, their combination should reduce the overall classification error (Ho, 1994). Information fusion techniques have been intensively investigated in recent years and their applicability to classification problems has been widely examined as a natural need for better classification accuracy (Ruta and Gabrys, 2000; Xu et al., 1992; Ranawana and Palade, 2005).

Based on the output of the underlying classifiers used in the combination process, classifier fusion methods can be divided into three main categories (Xu et al., 1992): methods that are dealing with the class labels (Type 1), methods that are using the rank list of the preferred class-labels (Type 2) and methods working with the preference score ('probability') of the class-labels (Type 3). One common approach for combining Type 1 classifiers is majority voting, which has its own limitations (e.g., in the absence of a clear majority winner, it is not obvious how a majority voter should decide) (Kuncheva et al., 2003). Several better techniques have been developed for fusing classifiers' decisions, but most of them require the classifiers' output to be of the Type 2 or Type 3 form (Robles et al., 2004). It is notable that most protein secondary structure classifiers belong to the Type 1 classification category. In this paper we combine several well-known existing protein secondary classifiers, in a multi-classifier system, by first converting the individual classifier outputs from Type 1 to Type 3 and, then, use several classifier fusion techniques to combine the outputs of individual classifiers.

The paper is organized as follows. The general architecture of the proposed meta-classifier have been explained in Section 2. Section 3 describes three different classifier fusion techniques including ordered weighted averaging, Dempster's combination rule and Choquet fuzzy integral operator, and also demonstrates the application of these methods in the protein secondary structure classifier fusion context. The individual classifiers and the datasets are briefly described in Section 4. Section 5 presents some criteria for measuring the accuracy of the classifiers. Section 6 reveals the results of the fusion methods used and, finally, some conclusions are drawn in Section 7.

2 The meta-classifier system

2.1 Converting and organizing classifiers' outputs

Confusion matrices are the most common visualization tool for characterizing the behavior of a classifier. They represent the predicted classes (rows of the matrix) versus the actual classes (columns of the matrix). The (i,j)th element in the matrix shows the number of times the label i is assigned to the actual label j. Therefore, higher diagonal values correspond to better classification.

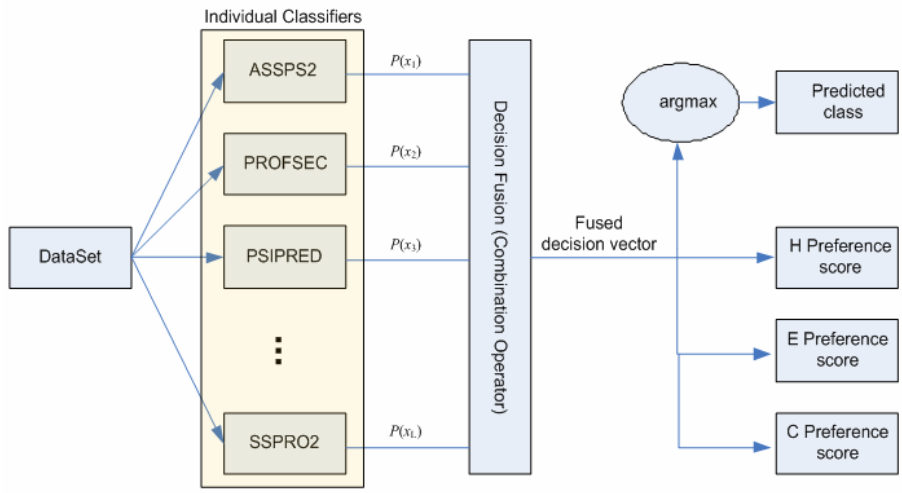
As it was mentioned before, the outputs of most protein secondary structure classifiers are of the Type 1 form. (Xu et al., 1992) suggested that, under certain assumptions, the confusion matrix of a Type 1 classifier can be used to convert its class-label output to class-preference output, as explained below. The assumptions are that, firstly, the confusion matrix can capture the classifier's behavior, and, secondly, the classifier's behavior does not change over time. Here, we used the same approach to alter the output of our classifiers.

Given the output of a Type 1 classifier (let say l) and assuming that the output is correct, we normalized the column l of the confusion matrix to get the corresponding class-preference scores. Subsequently, we organized the outputs of all classifiers in a decision profile which is a compact representation of multiple classifiers' outputs in a matrix format (Kuncheva et al., 2001). In decision profile (DP) matrices, each row represents an individual classifier's output and each column represents the amount of 'confidence' from all classifiers to a certain class. For example, in Type 1 classification, all elements in a row would be zero except the element that corresponds to the correct class-label.

2.2 General architecture of the protein secondary structure meta-classifier

Let $X = \{x_1, x_2, \ldots, x_L\}$ be the set of L classifiers, $\mathbf{C} = \{H, E, C\}$ be the set of class-labels corresponding to secondary structural elements, and $P(x_i) = \{p(x_i^H), p(x_i^E), p(x_i^C)\}$ be the output of classifier i, where $p(x_i^c)$ indicates the preference score given by classifier x_i to the class-label $c \in \mathbf{C}$. We represent the output of multiple classifiers in a decision profile as $\mathrm{DP}(X) = [P(x_1)\ P(x_2)\ \ldots\ P(x_i)\ \ldots\ P(x_L)]^{\mathrm{T}}$. The columns of $\mathrm{DP}(X)$ have been independently fused using some fusion operator (see Section 3). After the fusion process, the secondary structure of each certain amino acid is extracted from the maximum membership value of the fused result. The general architecture of the proposed multi-classifier approach is shown in Figure 1.

Figure 1 The meta-classifier system and the fusion scheme



Note: For information about individual classifiers see Section 4.2.

3 Fusion techniques

Information fusion deals with synergistic combination of different data sources, such as databases, classifiers, etc., to provide a better understanding of the problem (Mongi and Rafael, 1992). Classifier fusion techniques can be categorized as conventional or intelligent approaches. Here, we demonstrate the ordered weighted averaging operator and the Dempster's combination rule for the conventional approaches and the Choquet integral operator for the intelligent ones.

3.1 Ordered weighted averaging

The Ordered Weighted Averaging (OWA) operator was first introduced by Yager to solve the multi-criteria aggregation problem (Ronald, 1988). OWA's versatility allows it to cover the range between satisfying all criteria ('and') or satisfying at least one criterion in a parameterized manner. OWA maps L-dimensional real-value inputs to a real-value output, which makes it an appropriate tool for combining Type 3 classifiers (Kazemian et al., 2005). OWA is defined in a general form as follows:

$$OWA(p_1, p_2, ..., p_L) = \sum_{j=1}^{L} w_j p_{\sigma(j)} \tag{1}$$

where σ is a permutation that orders the input values in ascending order: $p_{\sigma(1)} \leq p_{\sigma(2)} \leq \cdots \leq p_{\sigma(L)}$. All weights are non-negative ($w_i \geq 0$) and sum to one. By adjusting the weights, w_j, in OWA, we can achieve a wide range of well-known operators such as max, min, median (Detyniecki, 2001). It is worthwhile to mention that OWA is always between the minimum and the maximum of the input values. The general algorithm for classifier fusion using OWA operator is shown in Figure 2.

Figure 2 The general algorithm of OWA based classifier fusion

1 *Choose L weights such that:*

$$\forall i \in \{1,2,...,L\}; w_i \geq 0, \text{ and } \sum_{i=1}^{L} w_i = 1$$

[we used exponential weights explained in (Filev and Yager, 1998)]

2 *For* each class-label c ∈ {H,E,C}

a. *Sort* $p(x_i^c)$, *i=1,2,...,L in descending order, such that*

$$a_1 = \max_i p\left(x_i^c\right), \quad \text{and} \quad a_L = \min_i \, p\left(x_i^c\right)$$

b. *Calculate the total preference score* ($\mu(X^c)$) *for class c as:*

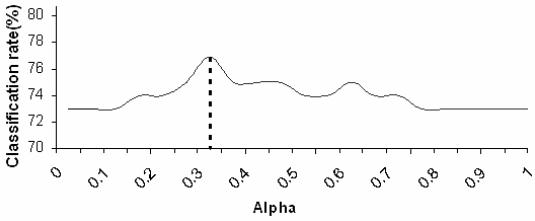
$$\mu\left(X^c\right) = \sum_{i=1}^{L} w_i a_i$$

3 *Obtain the final class-label by:*

$$\arg\max_k \mu\left(X^k\right)$$

To obtain the OWA weights, we used a certain family of OWA operators called optimistic exponential OWA (Filev and Yager, 1998). The weights were defined in a similar way to those used in the method of exponential smoothing, and parameterized by one variable, $\alpha \in [0, 1]$ (Filev and Yager, 1998). In order to find the parameter α, we varied it from zero to one with step size of 0.01, calculated the corresponding correct classification rate, and then found the value of α for which the maximum correct classification rate has occurred. Figure 3 shows the correct classification rate as a function of α.

Figure 3 Correct classification rate for varying α, in optimistic OWA



3.2 Dempster's rule of combination

Dempster's combination rule (DCR) is a generalization of Bayes' rule (Dempster, 2008), which combines the evidences from multiple independent sources. It has been successfully applied to the multiple classifiers combination problem (Xu et al., 1992). In our classifier fusion benchmark, the evidences are the preference scores in each column of the decision profile matrix.

Let $p(x_1^c)$ and $p(x_2^c)$ denote the preference scores of class c given by two classifiers. The adapted version of Dempster's combination rule can be written as follows:

$$p\left(x_{12}^c\right) = p\left(x_1^c\right) \oplus p\left(x_2^c\right) = \frac{p\left(x_1^c\right) \times p\left(x_2^c\right)}{\sum_{k \in \{H,E,C\}} p\left(x_1^k\right) \times p\left(x_2^k\right)} \tag{2}$$

The associative and commutative properties of the DCR rule make it possible to sequentially combine multiple evidences from different classifiers. The general algorithm for classifier fusion using DCR operator is shown in Figure 4.

Figure 4 Dempster's combination rule for classifier fusion

1 *For* each class-label c ∈ {H,E,C}

   a. *Set* $\mu(X^c)=p(x_1^c)$

   b. *For classifiers l = 2 to L*

$$\mu\left(X^c\right) = \mu\left(X^c\right) \oplus p\left(x_l^c\right)$$

2 *Obtain the final class-label by:*

$$\arg\max_k \mu\left(X^k\right)$$

3.3 Choquet fuzzy integral operator

Fuzzy Integrals are defined as the integration of a function with respect to a fuzzy measure (or a λ-fuzzy measure). In this sense, they are analogous to Lebesgue integrals with respect to an ordinary measure. Choquet integral (Choquet, 1954) is one of the most common fuzzy integral operators that has been successfully applied to the classifier combination problem (Kuncheva, 2001, Wang et al., 1998).

In the classifier fusion benchmark, the function that we are integrating over is the preference scores of the independent classifiers to a certain class-label, with respect to a λ-fuzzy measure defined over the space of the classifiers.

$$\mu\left(X^c\right) = \sum_{j=1}^{L} \left(p\left(x_{\sigma(j)}^c\right) - p\left(x_{\sigma(j-1)}^c\right) \right) g\left(A_{\sigma(j)}^c\right) \tag{3}$$

where $p\left(x_i^c\right)$ is the preference score of classifier x_i to class c; σ is a permutation that orders the preference scores in an ascending order $\left(p\left(x_{\sigma(1)}^c\right) \le p\left(x_{\sigma(2)}^c\right) \le \ldots \le p\left(x_{\sigma(L)}^c\right)\right)$, and $p\left(x_{\sigma(0)}^c\right) = 0$. Also $A_i^c = \{x_i^c, x_{i+1}^c, \ldots, x_L^c\}$, and g is a λ–fuzzy measure introduced by (Sugeno, 1977) and calculated recursively as follows:

$$g(A_i^c) = p^i + g\left(A_{i-1}^c\right) + \lambda\, p^i\, g\left(A_{i-1}^c\right) \tag{4}$$

where p^i is the importance of classifier x_i, and λ is the unique root, greater than –1, of the following equation (Tahani and Keller, 1990);

$$\lambda + 1 = \prod_{j=1}^{L} \left(1 + \lambda p^j\right) \qquad \lambda \neq 0 \tag{5}$$

We have calculated the importance of classifier x_i as the correct classification rate of x_i on the training data. On the other hand, the normalized summation of the diagonal elements of each classifier's confusion matrix (on training data) represents the total importance of the classifier. The general algorithm for classifier fusion using Choquet fuzzy integral is shown in Figure 5 (Kuncheva, 2001).

Figure 5 Choquet fuzzy integral operator in classifier fusion

1 *Estimate the importance of classifiers*(p^1, p^2, ...,p^L) *from training data as explained above*
2 *Calculate* λ *using Eq.5*
3 *Calculate fuzzy densities using Eq.4*
4 *Calculate the total preference score for class c using Eq.3*
5 *Obtain the final class-label by:*

$$\arg\max_k \mu\left(X^k\right)$$

4 Datasets and classifiers used

4.1 Datasets

Our evaluation was carried out on few datasets from EVA[1] server. EVA is a web service which continuously and automatically pulls newly resolved protein structures back from PDB[2], sends the corresponding required information to the prediction servers, gets their results back, evaluates them extensively and displays the final results on a web interface (Rost and Eyrich, 2001; Koh et al., 2003).

Table 1 The selected datasets from EVA server

EVA data sets			
Name	*No of proteins*	*No of residues*	*Description*
Set 1	30	More than 4,000	Cumulative results from 1999 to October 2002
Set 2	134	More than 16,000	Cumulative results from 1999 to October 2002
Set 3	80	More than 8,000	Cumulative results from October 2002
Set 4	175	More than 17,000	Cumulative results from October 2002

Notes: The datasets are located at:
http://cubic.bioc.columbia.edu/eva/sec_2002_10/common.html. The overlaps of Set 3 and Set 4 with Set 1 and Set 2 have been removed.

We have randomly selected one third of the proteins in each dataset as the training data, for constructing the confusion matrix, finding the parameter α and calculating the importance of each individual secondary structure classifier. The remaining two thirds

have been used for evaluating the proposed approach and reporting the results (results on the training data are not shown).

4.2 Classifiers used

Table 2 shows some information about the selected protein secondary structure prediction (classification) servers used in our classifier fusion system.

Table 2 Protein secondary structure classifiers used as individual classifiers

Secondary structure prediction servers		
Name	*Location*	*Prediction method*
APSSP2	Institute of Microbial Technology, India	EBL^3 + neural network
PROFSEC	Columbia University, USA	Profile-based neural network
PSIPRED	University College London, UK	Neural network
SAM-T99	University of California, Santa Cruz, USA	Hidden Markov model
SSPRO2	University of California, Irvine, USA	Recurrent neural network
PHDPSI	University of Columbia, USA	Profile-based neural network
PROF_KING	University of Wales, UK	Cascading different classifiers

- *APSSP2* combines the results of standard neural network (NN) and modified version of example based learning (EBL) trained on proteins with high resolution in PDB. It leverages the value of homology-based methods (e.g., EBL), considering the increasing number of known protein structures (Raghava, 2000).
- *PROFSEC* employs a cascade of three NNs: the first one is a simple feed-forward NN, which maps a set of local and global protein characteristics to its corresponding secondary structure. The second NN improves the output of the first NN by applying the natural constraints between adjacent predictions, and the third NN corrects some obvious prediction errors (Rost, 2001; Rost, 1996; Rost, 2005).
- *PSIPRED* utilizes two sequential NNs: The first NN receives PSI-BLAST's profile as an input and provides an initial prediction. The second NN improves the results of the first NN by checking for invalid structures (Jones, 1999).
- *SAM-T99* uses multiple sequence alignment generated by profile hidden Markov models (HMM) to predict secondary structures (Karplus et al., 1998).
- *SSPRO2* combines bidirectional recurrent NNs and PSI-BLAST profiles for predicting the secondary structure (Pollastri et al., 2002).
- *PHDPSI* improves the PHD method (Rost, 1996) by applying the information from multiple sequence alignment obtained by PSI-BLAST's profiles (Przybylski and Rost, 2002).
- *PROF_KING* cascades different types of classifiers together and combines their results using a NN and linear discrimination method. It improves the prediction results by "*exploiting the production of uncorrelated errors*" from different kinds of classifiers (Ouali and King, 2000).

5 Evaluation metrics

- *Three-class classification accuracy* (Q_3) criterion is the most common measure used for secondary structure prediction (Rost and Sander, 1993), and defined as follows:

$$Q_3 = 100 \times \frac{1}{N_{res}} \times \sum_{i=1}^{3} M_{ii} \tag{6}$$

where M_{ii} is the number of residues observed in class i and classified as i, and N_{res} is the total number of residues.

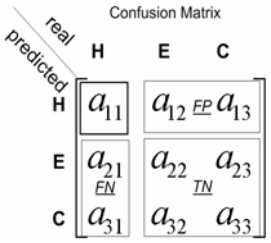
- *Per-class accuracy* criterion for class i is defined as the percentage of correctly classified residues in the class i, to all residues observed in class i (Rost and Sander, 1993).

$$Q_i^{\%obs} = 100 \times \frac{M_{ii}}{obs_i} \tag{7}$$

where M_{ii} is the number of residues observed in class i and classified as i, and obs_i is the total number of residues observed in class i.

- *Multi-class specificity (MC-sp) and Multi-class sensitivity (MC-sen): specificity and sensitivity* are common metrics for assessing the performance of binary classifiers (Altman and Bland, 1994). Here, we have adapted the specificity (sensitivity) measure of a binary classifier to our multi-class classifier problem. We first break the multi-class classifier to several binary classifiers, we then calculate the specificity (sensitivity) of each binary classifier and, finally, we obtain the multi-class specificity (sensitivity) by averaging the specificity of the binary classifiers. Figure 6 shows how we can break the confusion matrix of a three-class classifier when the actual class label is H.

Figure 6 The confusion matrix of a three-class classifier is shown. The matrix has been pictured in a form of binary classifier, where H is the 'positive' class, and E and C belong to the 'negative' class. TP(a_{11}), FP($a_{21} + a_{31}$), FN($a_{12} + a_{13}$) and TN($a_{22} + a_{23} + a_{32} + a_{33}$) are true positives, false positives, false negatives, and true negatives regions, respectively.



6 Results

A summary of the classification results obtained using the selected classifiers, which were described in Section 4, is presented in Table 3. These results demonstrate that the

best classifier for all of the four datasets is PSIPRED. Although its classification accuracy is the highest among other secondary structure classifiers, our multi-classifier approach has further improved this. The results of ordered weighted averaging, Dempster's combination rule, Choquet fuzzy integral and majority voting based classifier fusion methods are presented in Table 4. The multi-class specificity, multi-class sensitivity, and accuracy are also shown in Figures 7, 9, 11 and 13 for EVA set 1 to 4, respectively. These results show that the Choquet based classifier fusion provided the best results with respect to accuracy, multi-class specificity and sensitivity. While other fusion methods have provided some improvements too, the Choquet based classifier fusion method showed the best improvements: 2.24%, 2.57%, 2.14%, and 2.06% with respect to accuracy; 1.17%, 1.79%, 1.32% and 1.2% with respect to multi-class specificity; 2.35%, 2.57%, 2.35%, and 2.06% with respect to multi-class sensitivity, compared to PSIPRED on the four chosen datasets.

Good results have been achieved for α-helix and β-strand structure classification by ordered weighted averaging and Choquet fuzzy integral classifier fusion systems. The results of Choquet fusion method have been improved by 5.9% for alpha helix and 4.33% for β-strand compared to PSIPRED in the EVA set 1 (Figure 8). Similar results have been obtained for the other datasets (Figures 10, 12 and 14).

Table 3 Prediction results of the individual classifiers on four EVA datasets

	Q_3	$Q_h^{\%obs}$	$Q_e^{\%obs}$	$Q_c^{\%obs}$	*MC-sp*	*MC-sen*
			EVA set 1			
APSSP2	74.49	78.00	65.65	77.01	87.04	74.08
PROFSEC	74.71	75.38	74.48	74.05	87.12	74.24
PSIPRED	74.78	78.53	68.25	75.67	87.18	74.36
SAM-T99	74.63	82.60	63.12	75.06	87.13	74.27
SSPRO2	73.58	78.14	62.79	76.45	86.60	73.21
			EVA set 2			
PROFSEC	74.43	77.32	70.17	74.18	87.21	74.43
PSIPRED	74.56	78.29	66.54	75.49	87.28	74.56
SAM-T99	73.97	81.25	63.42	73.24	86.99	73.97
SSPRO2	74.00	79.04	63.73	74.96	87.00	74.00
			EVA set 3			
PSIPRED	77.62	83.96	68.66	75.02	88.81	77.62
PHDPSI	73.29	75.35	68.73	73.15	86.65	73.29
PROFSEC	75.43	76.16	72.82	75.81	87.71	75.49
SAMT99	77.48	85.79	61.90	75.69	88.74	77.48
PROF_KING	73.54	73.25	69.08	75.75	86.77	73.54
			EVA set 4			
PSIPRED	78.08	84.85	69.06	75.37	89.04	78.08
PHDPSI	74.49	79.91	67.16	72.38	87.25	74.49
PROFSEC	76.54	78.19	72.23	76.90	88.27	76.54
SAMT99	77.55	87.06	64.23	74.09	88.78	77.55
PROF_KING	72.80	71.42	68.70	76.22	86.40	72.80

The comparison between majority voting and Choquet operator shows that the Choquet based fusion has an improvement of 1.34% compared to majority voting for the EVA set 1, which is not very impressive at a first look, but, by better analysing the results, we found out that the Choquet based fusion provided an improvement of 4.26% and 6.21% in α-helix and β-strand structures, respectively, compared to majority voting. For the other datasets, similar results are obtained and presented in Table 4.

Table 4 Results for different classifier combination methods

	Q_3	$Q_h^{\%obs}$	$Q_e^{\%obs}$	$Q_c^{\%obs}$	*MC-sp*	*MC-sen*
			EVA set 1			
Majority voting	75.68	80.17	66.37	77.68	87.88	75.77
OWA	76.47	84.31	73.08	71.92	87.98	75.96
DCR	76.42	81.12	71.52	75.81	87.87	75.75
Choquet	77.02	84.43	72.58	73.43	88.35	76.71
			EVA set 2			
Majority voting	76.12	84.33	69.61	72.62	88.06	76.12
OWA	76.72	84.47	77.99	69.08	88.36	76.72
DCR	76.59	85.32	66.14	75.68	88.24	76.59
Choquet	77.13	85.45	70.61	73.56	89.07	77.13
			EVA set 3			
Majority voting	78.77	83.42	70.91	77.40	89.38	78.77
OWA	79.08	83.62	76.76	74.20	89.45	78.84
DCR	79.58	84.84	61.34	82.05	89.79	79.58
Choquet	79.76	87.99	68.28	76.96	90.13	79.97
			EVA set 4			
Majority voting	78.95	84.35	70.87	77.21	89.47	78.95
OWA	79.75	86.83	71.95	77.56	89.97	79.78
DCR	79.58	85.12	64.47	81.09	89.79	79.58
Choquet	80.14	87.97	70.83	80.92	90.24	80.14

Figure 7 The MC-sp, MC-sen and Q_3 of the five individual classifiers and of the classifier fusion systems on EVA set 1

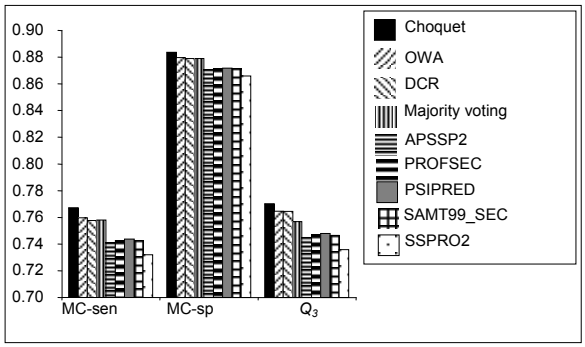


Figure 8 Comparison of $Q_h^{\%obs}$, $Q_e^{\%obs}$, $Q_c^{\%obs}$ and Q_3 between the best performing individual classifier and the classifier fusion systems on EVA set 1

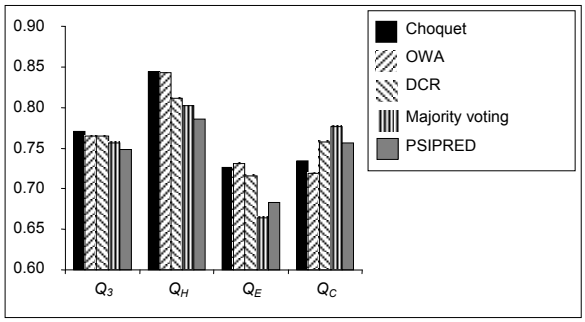


Figure 9 The MC-sp, MC-sen and Q3 of the five individual classifiers and of the classifier fusion systems on EVA set 2

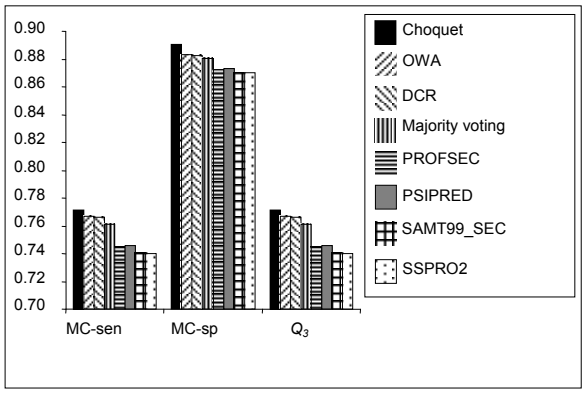


Figure 10 Comparison of $Q_h^{\%obs}$, $Q_e^{\%obs}$, $Q_c^{\%obs}$ and Q_3 between the best performing individual classifier and the classifier fusion systems on EVA set 2

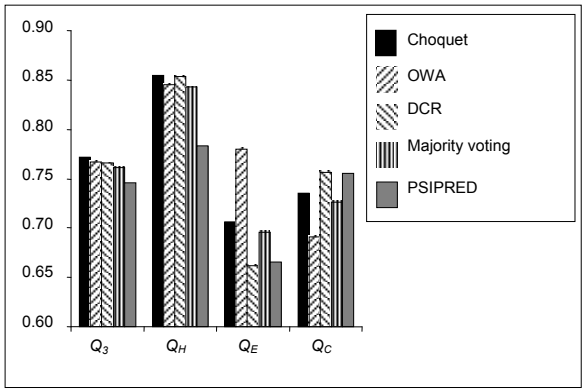


Figure 11 The MC-sp, MC-sen and Q_3 of the five individual classifiers and of the classifier fusion systems on EVA set 3

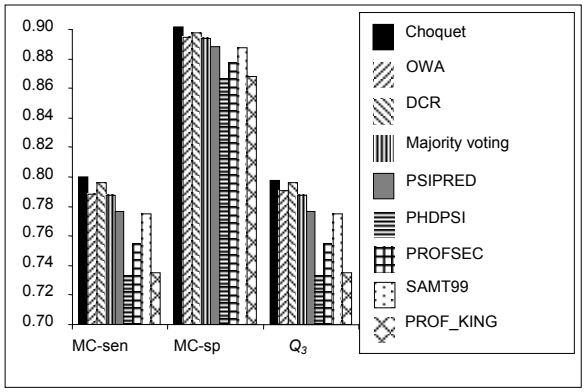


Figure 12 Comparison of $Q_h^{\%obs}$, $Q_e^{\%obs}$, $Q_c^{\%obs}$ and Q_3 between the best performing individual classifier and the classifier fusion systems on EVA set 3

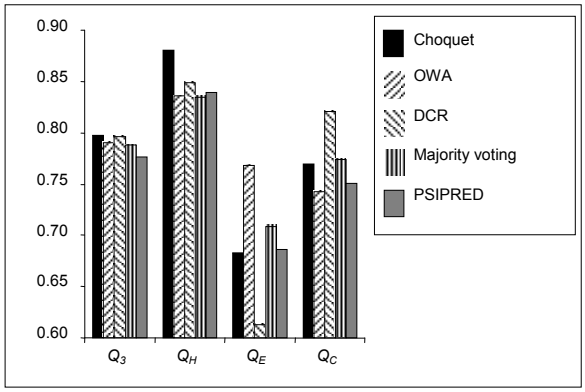


Figure 13 The MC-sp, MC-sen and Q_3 of the five individual classifiers and of the classifier fusion systems on EVA set 4

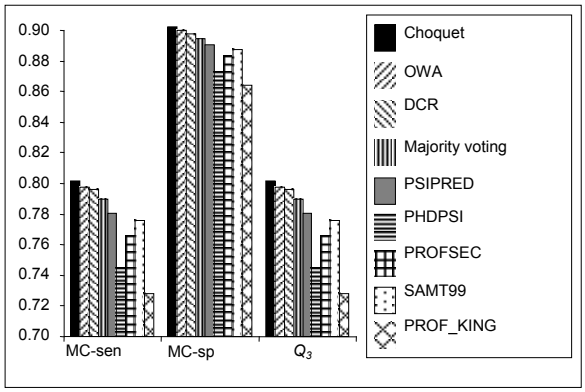
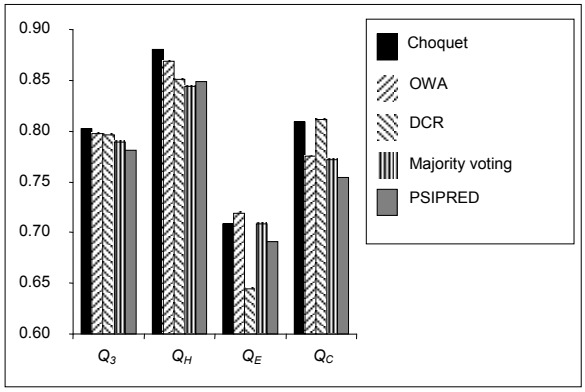


Figure 14 Comparison of $Q_h^{\%obs}$, $Q_e^{\%obs}$, $Q_c^{\%obs}$ and Q_3 between the best performing individual classifier and the classifier fusion systems on EVA set 4



7 Conclusions and further research

Classifier fusion is an important step towards building a meta-classifier. A meta-classifier combines the results of several existing classifiers and returns results that should be more accurate. In this paper, the decision profile of several well-known protein secondary structure classifiers have been extracted from their confusion matrices, and several classifier fusion techniques, including ordered weighted averaging, Dempster's combination rule and Choquet fuzzy integral operator, have been used to combine these decisions. The results showed that the Choquet fuzzy integral operator provided the best accuracy overall.

The combination of more existing individual classifiers led us to better prediction results for protein secondary structure. Improving existing individual models or developing new better models is more expensive than simply combining existing individual models. In addition, developing more models for the same problem might be more costly than using a single classifier, but if the individual classifiers are properly combined and diverse enough from one another, we can achieve better overall prediction results using less trained classifiers. Therefore, training individual classifiers to be integrated in a multi-classifier system can be less time consuming than training one single very well performing model.

There are two key challenges in the classifier combination problem. First, classifier fusion, in which the results of individual classifiers are combined to achieve the final decision. In this regard, developing various intelligent classifier fusion methods is a good avenue for future research. Second, classifier selection, in which the best individual classifiers are selected to contribute to the final decision. In this paper, we focused on the first aspect of classifier combination only. The confusion matrix of each classifier was used as being representative of its expertness, and did not contain the confidence of each decision, separately. If the classifier confidences or more details of regional expertness are available, the fusion results can be expected to be improved even more than what has been reported in this paper.

References

Altman, D.G. and Bland, J.M. (1994) 'Diagnostic tests 1: sensitivity and specificity', *BMJ*, Vol. 308, pp.1552.

Argos, P., Hanei, M. and Garavito, R.M. (1978) 'The Chou-Fasman secondary structure prediction method with an extended data base', *FEBS Letters*, Vol. 93, pp.19–24.

Bairoch, A. and Apweiler, R. (1999) 'The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999', *Nucleic Acids Res*., Vol. 27, pp.49–54.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) 'The protein data bank', *Nucleic Acids Res*., Vol. 28, pp.235–242.

Cai, Y.D., Liu, X.J. and Chou, K.C. (2003) 'Prediction of protein secondary structure content by artificial neural network', *J Comput Chem*, Vol. 24, pp.727–731.

Choquet, G. (1954) 'Theory of capacities', *Annales de l'institut Fourier*, Vol. 5, pp.131–259.

Dempster, A. (2008) 'Upper and lower probabilities induced by a multivalued mapping', Classic Works of the Dempster-Shafer Theory of Belief Functions.

Detyniecki, M. (2001) *Fundamentals on Aggregation Operators*, AGOP, Berkeley.

Filev, D. and Yager, R.R. (1998) 'On the issue of obtaining OWA operator weights', *Fuzzy Sets and Systems*, Vol. 94, pp.157–169.

Ho, T.K. (1994) 'Decision combination in multiple classifier systems', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, pp.66–75.

Jones, D.T. (1999) 'Protein secondary structure prediction based on position-specific scoring matrices', *J Mol Biol*., Vol. 292, pp.195–202.

Karplus, K., Barrett, C. and Hughey, R. (1998) 'Hidden Markov models for detecting remote protein homologies', *Bioinformatics*, Vol. 14, pp.846–856.

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M. and Hughey, R. (2003) 'Combining local-structure, fold-recognition, and new fold methods for protein structure prediction', *Proteins*, Vol. 53, No. 6, pp.491–496.

Kazemian, M., Moshiri, B., Nikbakht, H. and Lucas, C. (2005) 'Protein secondary structure classifiers fusion using OWA', *ISBMDA*, pp.338–345.

Kazemian, M., Moshiri, B., Nikbakht, H. and Lucas, C. (2007) 'A new expertness index for assessment of secondary structure prediction engines', *Comput. Biol. Chem*., Vol. 31, pp.44–47.

Kim, S. (2004) 'Protein beta-turn prediction using nearest-neighbor method', *Bioinformatics*, Vol. 20, pp.40–44.

Koh, I.Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A. and Rost, B. (2003) 'EVA: evaluation of protein structure prediction servers', *Nucleic Acids Res*., Vol. 31, pp.3311–3315.

Kuncheva, L.I. (2001) 'Combining classifiers: soft computing solutions', in Pal, S.K. (Ed.): *Pattern Recognition: From Classical to Modern Approaches*, World Scientific.

Kuncheva, L.I., Bezdek, J.C. and Duin, R.P.W. (2001) 'Decision templates for multiple classifier fusion: an experimental comparison', *Pattern Recognition*, Vol. 34, pp.299–314.

Kuncheva, L.I., Whitaker, C.J., Shipp, C.A. and Duin, R.P.W. (2003) 'Limits on the majority vote accuracy in classifier fusion', *Pattern Analysis and Applications*, Vol. 6, pp.22–31.

Mongi, A.A. and Rafael, C.G. (1992) 'Data fusion in robotics and machine intelligence', Academic Press Professional, Inc.

Ouali, M. and King, R.D. (2000) 'Cascaded multiple classifiers for secondary structure prediction', *Protein Sci*., Vol. 9, pp.1162–1176.

Pollastri, G. and Mclysaght, A. (2005) 'Porter: a new, accurate server for protein secondary structure prediction', *Bioinformatics*, Vol. 21, pp.1719–1720.

Pollastri, G., Przybylski, D., Rost, B. and Baldi, P. (2002) 'Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles', *Proteins*, Vol. 47, pp.228–235.

Przybylski, D. and Rost, B. (2002) 'Alignments grow, secondary structure prediction improves', *Proteins*, Vol. 46, pp.197–205.

Raghava, G.P.S. (2000) 'Protein secondary structure prediction using nearest neighbor and neural network approach', CASP4.

Ranawana, R. and Palade, V. (2005) 'A neural network based multi-classifier system for gene identification in DNA sequences', *Neural Comput. Appl.*, Vol. 14, pp.122–131.

Robles, V., Larranaga, P., Pena, J.M., Menasalvas, E., Perez, M.S., Herves, V. and Wasilewska, A. (2004) 'Bayesian network multi-classifiers for protein secondary structure prediction', *Artif. Intell. Med.*, Vol. 31, pp.117–136.

Ronald, R.Y. (1988) 'On ordered weighted averaging aggregation operators in multicriteria decisionmaking', *IEEE Trans. Syst. Man Cybern.*, Vol. 18, pp.183–190.

Rost, B. (1996) 'PHD: predicting one-dimensional protein structure by profile-based neural networks', *Methods Enzymol*, Vol. 266, pp.525–539.

Rost, B. (2001) 'Review: protein secondary structure prediction continues to rise', *J Struct. Biol.*, Vol. 134, pp.204–218.

Rost, B. (2005) 'How to use protein 1-D structure predicted by PROFphd', *The Proteomics Protocols Handbook*.

Rost, B. and Eyrich, V.A. (2001) 'EVA: large-scale analysis of secondary structure prediction', *Proteins*, Suppl. 5, pp.192–199.

Rost, B. and Sander, C. (1993) 'Prediction of protein secondary structure at better than 70% accuracy', *J Mol Biol.*, Vol. 232, pp.584–599.

Rost, B., Yachdav, G. and Liu, J. (2004) 'The PredictProtein server', *Nucleic Acids Res.*, Vol. 32, pp.W321–W326.

Ruta, D. and Gabrys, B. (2000) 'An overview of classifier fusion methods', *Computing and Information Systems*, Vol. 7, pp.1–10.

Sugeno, M. (1977) 'Fuzzy measures and fuzzy integrals: a survey', in Gupta, M.M., Saridis, G.N. and Gaines, B.R. (Eds.): *Fuzzy Automata and Decision Processes*.

Tahani, H. and Keller, J.M. (1990) 'Information fusion in computer vision using the fuzzy integral', *Systems, Man and Cybernetics, IEEE Transactions*, Vol. 20, pp.733–741.

Uniprot (2008) 'The universal protein resource (UniProt)', *Nucleic Acids Res*., Vol. 36, pp.D190–D195.

Wang, D., Keller, J.M., Carson, C.A., Mcado-Edwards, K.K. and Bailey, C.W. (1998) 'Use of fuzzy-logic-inspired features to improve bacterial recognition through classifier fusion', *IEEE Trans Syst. Man Cybern. B Cybern.*, Vol. 28, pp.583–591.

Xu, L., Krzyzak, A. and Suen, C.Y. (1992) 'Methods of combining multiple classifiers and their applications to handwriting recognition', *Systems, Man and Cybernetics, IEEE Transactions*, Vol. 22, pp.418–435.

Notes

1 Evaluation of automatic protein structure prediction.

2 Protein data bank.

3 Example based learning.