PROTEINS: Structure, Function, and Bioinformatics 65:453–462 (2006)

Measures for the Assessment of Fuzzy Predictions of Protein Secondary Structure

Julian Lee*

Department of Bioinformatics and Life Science, Bioinformatics and Molecular Design Technology Innovation Center and Computer Aided Molecular Design Research Center, Soongsil University, Seoul 156-743, Korea

ABSTRACT **Many of the recent secondary structure prediction methods incorporate the idea of fuzzy set theory, where instead of assigning a definite secondary structure to a query residue, probability for the residue being in each of the conformational states is estimated. Moreover, continuous assignment of conformational states to the experimentally observed protein structures can be performed in order to reflect inherent flexibility. Although various measures have been developed for evaluating performances of secondary structure prediction methods, they depend only on the most probable secondary structures. They do not assess the accuracy of the probabilities produced by fuzzy prediction methods, and they cannot incorporate information contained in continuous assignments of conformational states to observed structures. Three important measures for evaluating performance of a secondary structure prediction algorithm, *Q* score, Segment OVerlap (SOV) measure, and the *k*-state correlation coefficient (Corr), are deformed into fuzzy measures *F* score, Fuzzy OVerlap (FOV) measure, and the fuzzy correlation coefficient (Forr), so that the new measures not only assess probabilistic outputs of fuzzy prediction methods, but also incorporate information from continuous assignments of secondary structure. As an example of application, prediction results of four fuzzy secondary structure prediction methods, PSIPRED, PROFking, SABLE, and PREDICT, are assessed using the new fuzzy measures. Proteins 2006;65:453–462.** © 2006 Wiley-Liss, Inc.

Key words: secondary structure prediction; assessment; evaluation

INTRODUCTION

The prediction of the three-dimensional structure of a protein from its amino acid sequence is one of the most important problems in bioinformatics. As a first step toward solving this problem, many algorithms for statistically predicting the local secondary structure, instead of the full global tertiary structure, have been developed.[1–18] The most common definition of the secondary structure is based on Dictionary of Secondary Structure of Proteins (DSSP),[19] where the secondary structure is classified as eight states. By grouping these eight states into three classes, Coil (C), Helix (H), and Extended (E), one obtains three state classification, which is more widely used. Therefore, the goal of the secondary structure prediction is to assign one of the three possible states to each residue of the query protein. Typical secondary structure prediction algorithms apply pattern recognition algorithms such as artificial neural network, *k*-nearest neighbor method, or support vector machine to the query protein sequence or sequence profile obtained by multiple sequence alignment with related sequences.

Many secondary structure prediction algorithms not only can assign a definite secondary structure class to a query residue, but also estimate probability for the residue being in each of the secondary structural classes. These methods can be considered as incorporating the fuzzy set theory[20,21] where the predicted secondary structure of a residue does not belong to a definite conformational state, but has fuzzy membership to all three conformational states. Therefore, this class of methods is called the fuzzy prediction methods in this work. The *k*-nearest neighbor method incorporating the concept of fuzzy set is explicitly called the fuzzy *k*-nearest neighbor method.[22] The probabilities produced by a fuzzy prediction method contain detailed information on possible secondary structure of a query residue, which is not contained in the secondary structure finally produced by the prediction algorithm.

Moreover, since the experimentally observed protein structure itself is not rigid, one can use continuous DSSP (DSSPcont)[23] instead of standard DSSP, so that more information on experimental structure can be maintained. Just as in the case of fuzzy prediction, a continuous assignment of the secondary structure produces the probability for a residue being in each of the conformational states. Therefore, continuous assignment of secondary structure such as DSSPcont can be considered as fuzzy observation, in contrast to crisp observation described by standard discrete DSSP assignments.

There are various applications of secondary structure prediction, such as homology modeling, threading, tertiary

Grant sponsor: Soongsil University Research Fund.

*Correspondence to: Julian Lee, Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Korea. E-mail: jul@ssu.ac.kr

Received 21 January 2006; Revised 29 May 2006; Accepted 5 July 2006

Published online 31 August 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21164

© 2006 WILEY-LISS, INC.

structure prediction of a new fold, and remote homology detection for the transfer of functional annotation between sequences, and some of these applications require not only the final discrete secondary structural classes produced by the prediction algorithm and discrete assignments, but also the full information contained in the fuzzy prediction and observation. One such application is the tertiary structure prediction based on fragment assembly,[24–30] where the local structures are generated according to the probabilities estimated by the secondary structure prediction methods. Therefore, for this class of methods, the correlation between fuzzy prediction and fuzzy observation is more important than that between their crisp counterparts. However, measures for evaluating secondary structure prediction algorithms that have been used so far,[31–36] called crisp measures in this work, compare only the final output of the most probable secondary structure predicted, with the discrete assignment of the experimental secondary structure. Clearly, we need measures to evaluate the accuracy of the probabilities of the secondary structural classes estimated by an algorithm, which also incorporate the continuous assignment of secondary structure.

In this work, I consider three important crisp measures for evaluating secondary structure prediction methods, Q score, Segment OVerlap (SOV) measure, and the k-state correlation coefficient (Corr), and deform them to obtain new measures that compare fuzzy prediction results with fuzzy observation. I will call the resulting modified measures as fuzzy measures, named F score, Fuzzy OVerlap (FOV) measure, and the fuzzy correlation coefficient (Forr). They are elaborated in the next section along with their crisp counterparts.

METHODS

Q and F Scores

The Q score is the simplest measure for evaluating secondary structure prediction performance. It is given by the percentage of residues predicted correctly. For a given secondary structure class S, one defines

$$Q_S \equiv 100\% \times \frac{N_{\mathrm{T}+}(S)}{O_+(S)} \tag{1}$$

where $O_+(S)$ is the observed number of residues in the class S, and $N_{\mathrm{T}+}(S)$ is the number of residues correctly predicted to be in the class S. The Q_3 score, the measure of overall performance without reference to a specific secondary structure class, is defined as

$$Q_3 \equiv \sum_{S\in\{\mathrm{C,H,E}\}} \frac{O_+(S)}{N} \quad Q_S = 100\% \times \frac{N_{\mathrm{corr}}(3)}{N} \tag{2}$$

where N is the total number of residues of the query protein, and $N_{\mathrm{corr}}(3) = \Sigma_{S\epsilon\{\mathrm{C,H,E}\}} N_{\mathrm{T}+}(S)$ is the total number of correctly predicted residues regardless of their secondary structures.

It is rather straightforward to deform Q_S and Q_3 to obtain fuzzy measures. We define

$$F_S \equiv 100\% \times \frac{\tilde{N}_{\mathrm{T}+}(S)}{\tilde{O}_+(S)},$$
$$F_3 \equiv 100\% \times \frac{\tilde{N}_{\mathrm{corr}}(3)}{N}, \tag{3}$$

where

$$\tilde{N}_{\mathrm{T}+}(S) \equiv \sum_j \mathrm{Pr_P}(j;S)\,\mathrm{Pr_O}(j;S),$$
$$\tilde{O}_+(S) \equiv \sum_j \mathrm{Pr_O}(j;S), \tag{4}$$

with $\mathrm{Pr_P}(j;S)$ and $\mathrm{Pr_O}(j;S)$ denoting the probability that the j-th residue belongs to the secondary structure class S as estimated by the fuzzy prediction and observation, respectively. Similarly,

$$\tilde{N}_{\mathrm{corr}}(3) \equiv \sum_{S\in\{\mathrm{C,H,E}\}} \tilde{N}_{\mathrm{T}+}(S)$$
$$= \sum_{S\in\{\mathrm{C,H,E}\}} \sum_{j=1}^{N} \mathrm{Pr_P}(j;S)\,\mathrm{Pr_O}(j;S). \tag{5}$$

By construction, the lower and upper bound of F scores are 0 and 100%. It is evident that when the probabilities consist of 0s and 1s only, implying prediction and observation with 100% confidence, the F scores reduce to the Q scores.

As an example, let us consider the prediction result given in Figure 1. The Q scores for this prediction are

$$Q_{\mathrm{C}} \equiv 100\% \times \frac{2}{4} = 50\%,$$
$$Q_{\mathrm{H}} \equiv 100\% \times \frac{5}{8} = 62.5\%,$$
$$Q_3 \equiv 100\% \times \frac{7}{12} = 58.3\%. \tag{6}$$

The Q_{E} score is undefined since both the numerator and the denominator vanish, due to the fact that there

$\mathrm{Pr_O}$(C)	0.8 0.2 0.2 0.2 0.0 0.0 0.0 0.0 0.0 0.8 1.0 1.0
$\mathrm{Pr_O}$(H)	0.2 0.8 0.8 0.8 1.0 1.0 1.0 1.0 1.0 0.2 0.0 0.0
$\mathrm{Pr_O}$(E)	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
Observed	C H H H H H H H H C C C
Predicted	C C C C H H H H H H H C
$\mathrm{Pr_P}$(C)	0.8 0.5 0.5 0.5 0.1 0.1 0.1 0.1 0.1 0.2 0.2 0.8
$\mathrm{Pr_P}$(H)	0.1 0.4 0.4 0.4 0.8 0.8 0.8 0.8 0.8 0.5 0.5 0.1
$\mathrm{Pr_P}$(E)	0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.3 0.3 0.1

Fig. 1. An example of the fuzzy prediction. The values of the crisp and fuzzy measures for this result are shown in Table I.

is no residue observed to be in the class E. On the other hand, the F scores for this example are

$$\begin{aligned} F_{\mathrm{C}} &\equiv 100\% \times \frac{0.8^2+0.2\times0.5\times3+0.8\times0.2+0.2+0.8}{0.8+0.2\times3+0.8+2.0} \\ &= 50.0\%, \\ F_{\mathrm{H}} &\equiv 100\% \times \frac{0.2\times0.1+0.8\times0.4\times3+1.0\times0.8\times5+0.2\times0.5}{0.2+0.8\times3+1.0\times5+0.2} \\ &= 65.1\%, \\ F_3 &\equiv 100\% \times [0.8^2+0.2\times0.1+(0.2\times0.5+0.8\times0.4)\times3 \\ &\quad +0.8\times5+(0.8\times0.2+0.2\times0.5)+0.2+0.8]/12 \\ &= 59.8\%, \end{aligned} \tag{7}$$

where again F_{E} is undefined for this example.

SOV and FOV Measures

In contrast to the Q score, the SOV measure[31,34] is not based simply on the number of correct residues, but also puts emphasis on the continuity of a secondary structure element. For example, SOV(3) gives a big penalty for a prediction where a long helix is falsely predicted as two short helices because of one misclassified residue in the middle, whereas Q_3 score assesses this result as nearly perfect prediction since only the number of correctly predicted residues is taken into account.

As in the case of Q scores, one defines SOV measure for a particular secondary structure class S as well as that for the overall prediction. First, for a given secondary structural class S (=C,H,E), we define the set of overlapping segments:

$$\mathrm{Seg}(S) = \{(s_1(S), s_2(S)) | s_1(S) \cap s_2(S) \neq \emptyset\}. \tag{8}$$

where ($s_1(S)$, $s_2(S)$) is a pair of observed and predicted secondary structure segments in the class S, which has at least one residue in common. Seg(S) is the set of all such pairs. Then, the SOV measures are defined as

$$\begin{aligned} \mathrm{SOV}(S) &= 100\% \times \frac{1}{N(S)} \\ &\quad \times \sum_{\mathrm{Seg}(S)} \left[\frac{\mathrm{minov}(s_1(S),s_2(S)) + \delta(s_1(S),s_2(S))}{\mathrm{maxov}(s_1(S),s_2(S))} \mathrm{len}(s_1(S)) \right], \\ \mathrm{SOV}(3) &= 100\% \times \frac{1}{N_{\mathrm{all}}} \\ &\quad \times \sum_{S\in\{\mathrm{C,H,E}\}} \sum_{\mathrm{Seg}(S)} \left[\frac{\mathrm{minov}(s_1(S),s_2(S)) + \delta(s_1(S),s_2(S))}{\mathrm{maxov}(s_1(S),s_2(S))} \right. \\ &\qquad \left. \times \mathrm{len}(s_1(S)) \right], \end{aligned} \tag{9}$$

where len($s_1(S)$) and len($s_2(S)$) are the number of residues in the segments $s_1(S)$ and $s_2(S)$ respectively; minov($s_1(S)$, $s_2(S)$) is the length of actual overlap of a given pair $s_1(S)$ and $s_2(S)$; maxov($s_1(S)$, $s_2(S)$) is the length of the total extent of residues, which belong to either $s_1(S)$ or $s_2(S)$, and

$$\begin{aligned} \delta(s_1,s_2) = \min[&(\mathrm{maxov}(s_1,s_2) - \mathrm{minov}(s_1,s_2)); \mathrm{minov}(s_1,s_2); \\ &\mathrm{int}(\mathrm{len}(s_1)/2); \mathrm{int}(\mathrm{len}(s_2)/2)]. \end{aligned} \tag{10}$$

Also, the normalization factors $N(S)$ and N_{all} are defined as

$$\begin{aligned} N(S) &= \left[\sum_{\mathrm{Seg}(S)} \mathrm{len}(s_1(S)) + \sum_{\mathrm{Seg}'(S)} \mathrm{len}(s'_1(S)) \right], \\ N_{\mathrm{all}} &= \sum_{S\in\mathrm{C,H,E}} N(S), \end{aligned} \tag{11}$$

where Seg′(S) is the set of observed segments $s'_1(S)$ that have no overlap with predicted segments of secondary structure S. To preserve main features of the SOV measure, we keep the original definition of the segments and the sets of segments, $s_1(S)$, $s_2(S)$, Seg(S), Seg′(S), when we deform SOV to the fuzzy measure FOV, but redefine maxov and minov. As in the case of Q score, this is done by summing probabilities. That is, we define

$$\begin{aligned} \mathrm{minov}'(s_1(S),s_2(S)) &\equiv \sum_{j\in s_1(S)\cap s_2(S)} \min(\mathrm{Pr_O}(j;S), \mathrm{Pr_P}(j;S)) \\ \mathrm{maxov}'(s_1(S),s_2(S)) &\equiv \sum_{j\in s_1(S)\cup s_2(S)} \max(\mathrm{Pr_O}(j;S), \mathrm{Pr_P}(j;S)). \end{aligned}$$

Then the FOV measure is defined as

$$\begin{aligned} \mathrm{FOV}(S) &= 100\% \times \frac{1}{N(S)} \\ &\quad \times \sum_{\mathrm{Seg}(S)} \left[\frac{\mathrm{minov}'(s_1(S),s_2(S)) + \delta'(s_1(S),s_2(S))}{\mathrm{maxov}'(s_1(S),s_2(S))} \mathrm{len}(s_1(S)) \right], \\ \mathrm{FOV}(3) &= 100\% \times \frac{1}{N_{\mathrm{all}}} \\ &\quad \times \sum_{S\in\{\mathrm{C,H,E}\}} \sum_{\mathrm{Seg}(S)} \left[\frac{\mathrm{minov}'(s_1(S),s_2(S)) + \delta'(s_1(S),s_2(S))}{\mathrm{maxov}'(s_1(S),s_2(S))} \right. \\ &\qquad \left. \times \mathrm{len}(s_1(S)) \right], \end{aligned} \tag{12}$$

where δ' is defined in the manner similar to the case of SOV measure, with maxov′ and minov′ used in places of maxov and minov. As in the case of F scores, FOV measures reduce to SOV measures when the estimated probabilities consist of 0s and 1s only.

Let us elaborate by considering the example in Figure 1. There is one pair of overlapping segments of H in the middle, and for this pair we have

$$\begin{aligned} &\mathrm{minov}(s_1(H),s_2(H)) = 5, \quad \mathrm{maxov}(s_1(H),s_2(H)) = 10, \\ &\mathrm{minov}'(s_1(H),s_2(H)) = 0.8\times5 = 4, \\ &\mathrm{maxov}'(s_1(H),s_2(H)) = 0.8\times3 + 1.0\times5 + 0.5\times2 = 8.4, \end{aligned}$$

TABLE I. The Comparison of Performance Measures for the Example in Figure 1

	Q	SOV	Corr	F	FOV	Forr
All	58.3	63.8	0.120	59.8	65.6	0.636
Coil	50.0	31.3	0.120	50.0	30.1	0.554
Helix	62.5	80.0	0.120	65.1	83.3	0.747

The measures are undefined for E, since the probabilities of their appearance vanish in the observed secondary structure.

with $\delta(s_1, s_2) = \delta'(s_1, s_2) = \text{int}[\text{len}(s_2)/2] = 3$. Therefore,

$$\text{SOV}(H) = 100\% \times \frac{1}{8}\left(\frac{5+3}{10} \times 8\right) = 80.0\%,$$
$$\text{FOV}(H) = 100\% \times \frac{1}{8}\left(\frac{4+3}{8.4} \times 8\right) = 83.3\%.$$

Similar calculation for C yields

$$\text{SOV}(C) = 100\% \times \frac{1}{1+3}\left(\frac{1}{4} \times 1 + \frac{1}{3} \times 3\right) = 31.3\%,$$
$$\text{FOV}(C) = 100\% \times \frac{1}{1+3}\left(\frac{0.8}{2.3} \times 1 + \frac{0.8}{2.8} \times 3\right) = 30.1\%.$$

Again, SOV(E) and FOV(E) are undefined for this example since both the numerators and the denominators vanish. The measures for the overall performance are

$$\text{SOV}(3) = 100\% \times \frac{1}{12}\left(\frac{1}{4} \times 1 + \frac{5+3}{10} \times 8 + \frac{1}{3} \times 3\right) = 63.8\%,$$
$$\text{FOV}(3) = 100\% \times \frac{1}{12}\left(\frac{0.8}{2.3} \times 1 + \frac{4+3}{8.4} \times 8 + \frac{0.8}{2.8} \times 3\right) = 65.6\%.$$

k-State Correlation Coefficients

Among the measures we discuss, the k-state correlation coefficients, abbreviated as Corr scores in this work, have the most rigorous foundation in statistical theory. The assignment of a secondary structure to a residue can be considered as a categorical variable with three categories. When we restrict our attention to a specific secondary structural class S and classify a residue according to whether it belongs to the class S or not, the resulting categorical variable has only two categories. We then calculate the correlation coefficient between the observed and predicted variables. In contrast to Q scores and SOV measures, which range from 0 to 100%, the numerical value of a correlation coefficient is between -1 and 1, where 1 is a perfect linear correlation, 0 means no linear correlation, and -1 is a perfect linear anti-correlation. A random prediction results in a correlation coefficient close to 0.

Let us first consider the case of the two category assignment. We consider variables $\mathbf{X}_S(j)$ and $\mathbf{Y}_S(j)$ ($j = 1,2,\ldots,N$), which are two-dimensional vectors. For a given secondary structure S, we assign $\mathbf{X}_S(j) = (1,0)$ if the j-th residue is observed to be in the conformational state S, and $\mathbf{X}_S(j) = (0,1)$ otherwise. Similarly, $\mathbf{Y}_S(j) = (1,0)$ or $(0,1)$ depending on whether or not the j-th residue is predicted to be in the conformational state S. The two-state correlation coefficient between $\mathbf{X}_S$ and $\mathbf{Y}_S$ is then given as[35]

$$\begin{aligned}\text{Corr}(S) &\equiv \frac{\sum_{j=1}^{N}(\mathbf{X}_S(j)-\bar{\mathbf{X}}_S)\cdot(\mathbf{Y}_S(j)-\bar{\mathbf{Y}}_S)}{\sqrt{\sum_{k=1}^{N}(\mathbf{X}_S(k)-\bar{\mathbf{X}}_S)^2}\sqrt{\sum_{m=1}^{N}(\mathbf{Y}_S(m)-\bar{\mathbf{Y}}_S)^2}} \\ &= \frac{\sum_{j=1}^{N}(\mathbf{X}_S(j)\cdot\mathbf{Y}_S(j)-\bar{\mathbf{X}}_S\cdot\bar{\mathbf{Y}}_S)}{\sqrt{\sum_{k=1}^{N}(\mathbf{X}_S(k)^2-\bar{\mathbf{X}}_S^2)}\sqrt{\sum_{m=1}^{N}(\mathbf{Y}_S(m)^2-\bar{\mathbf{Y}}_S^2)}} \\ &= \frac{NN_{T+}(S)-O_+(S)P_+(S)}{\sqrt{O_+(S)O_-(S)P_+(S)P_-(S)}} \\ &= \frac{NN_{\text{corr}}(S)-O_+(S)P_+(S)-O_-(S)P_-(S)}{\sqrt{(N^2-O_+(S)^2-O_-(S)^2)(N^2-P_+(S)^2-P_-(S)^2)}}\end{aligned} \tag{13}$$

where $\bar{\mathbf{X}}_S(\bar{\mathbf{Y}}_S)$ is the average value of $\mathbf{X}_S(k)(\mathbf{Y}_S(k))$, $O_+(S)(P_+(S))$ is the numbers of residues observed (predicted) to be in the class S, and $O_-(S)$, $(P_-(S)))$ is the numbers of residues observed (predicted) not to be in the class S. Also, $N_{\text{corr}}(S) \equiv N_{\text{T}+}(S) + N_{\text{T}-}(S)$ with $N_{\text{T}+}(S)$ being the number of correctly predicted residues in the class S (true positives), and $N_{\text{T}-}(S)$ the number of residues correctly identified as something other than the class S (true negatives). It should be noted that since the secondary structure prediction is viewed as a two-class problem in calculating the correlation coefficient in Eq. (13), usually called the Matthews correlation coefficient, no distinction is made between states other than S.

To obtain the three-state correlation coefficient for all the secondary structural classes, we construct variables $\mathbf{X}$ and $\mathbf{Y}$ whose values are three-dimensional vectors. We assign $\mathbf{X}(j)(\mathbf{Y}(j)) = (1,0,0)$, $(0,1,0)$ or $(0,0,1)$ depending on whether the j-th residue is observed (predicted) to be in the class C, H, or E, to get [36]

$$\begin{aligned}\text{Corr}(3) &\equiv \frac{\sum_{j=1}^{N}(\mathbf{X}(j)\cdot\mathbf{Y}(j)-\bar{\mathbf{X}}\cdot\bar{\mathbf{Y}})}{\sqrt{\sum_{k=1}^{N}(\mathbf{X}(k)^2-\bar{\mathbf{X}}^2)}\sqrt{\sum_{m=1}^{N}(\mathbf{Y}(m)^2-\bar{\mathbf{Y}}^2)}} \\ &= \frac{NN_{\text{corr}}(3)-\sum_{S\in\{\text{C,H,E}\}}O_+(S)P_+(S)}{\sqrt{(N^2-\sum_{S'\in\{\text{C,H,E}\}}O+(S')^2)(N^2-\sum_{S''\in\{\text{C,H,E}\}}P_+(S'')^2)}}\end{aligned} \tag{14}$$

where as before, N denotes the number of all residues, and $N_{\text{corr}}(3) \equiv \Sigma_{S\in\{\text{C,H,E}\}}N_{\text{T}+}(S)$ the number of those predicted correctly. We see that Eq. (14) is the generalization of the last expression in Eq. (13) to the case of three-class assignment.

Both in Eq. (13) and Eq. (14) each component of the variables $\mathbf{X}_{(S)}(j)$ and $\mathbf{Y}_{(S)}(j)$ can be interpreted as representing the probability that the j-th residue belongs to a specific class, obtained from the observation and prediction respectively, when the confidence is 100%. Therefore, for fuzzy prediction, where the prediction is made with nonzero probabilities for all of the secondary structural classes, it is clear that the components of $\mathbf{Y}_{(S)}(j)$ should be replaced by the probabilities estimated by the fuzzy prediction algorithm. Similarly, for fuzzy observation, $\mathbf{X}_{(S)}(j)$ should be replaced by the probabilities produced by the DSSPcont assignment.

TABLE II. Average Scores of Prediction on EVA Common Set 1 for the Four Prediction Methods

	PSIPRED	PROFKING	SABLE	PREDICT
Q_3	76.6 (1.2)	71.0 (1.5)	76.8 (1.2)	72.7 (1.3)
Q_C	73.9 (1.6)	76.9 (1.5)	75.2 (2.0)	72.5 (1.9)
Q_H	83.8 (1.9)	69.5 (2.7)	81.6 (1.9)	83.1 (1.6)
Q_E	63.1 (3.8)	61.9 (4.1)	62.6 (3.9)	46.4 (3.6)
SOV(3)	75.9 (1.7)	69.2 (2.0)	74.8 (1.9)	66.4 (1.8)
SOV(C)	71.5 (1.9)	70.2 (2.0)	70.2 (2.3)	66.2 (2.1)
SOV(H)	83.7 (2.2)	73.4 (3.0)	81.9 (2.2)	79.0 (2.4)
SOV(E)	65.4 (4.1)	61.6 (4.6)	65.2 (4.4)	49.1 (3.5)
Corr(3)	0.571 (0.021)	0.498 (0.024)	0.562 (0.024)	0.509 (0.022)
Corr(C)	0.568 (0.019)	0.493 (0.023)	0.555 (0.023)	0.508 (0.021)
Corr(H)	0.624 (0.023)	0.575 (0.029)	0.630 (0.024)	0.558 (0.024)
Corr(E)	0.550 (0.038)	0.471 (0.042)	0.559 (0.041)	0.494 (0.031)
F_3	69.1 (1.1)	62.5 (1.1)	65.3 (1.1)	57.3 (0.9)
F_C	65.6 (1.1)	65.7 (0.9)	61.9 (1.2)	57.6 (0.9)
F_H	77.3 (1.6)	62.3 (2.0)	72.7 (1.6)	65.2 (1.3)
F_E	56.9 (2.8)	56.0 (2.8)	56.3 (2.4)	39.7 (1.5)
FOV(3)	69.0 (1.7)	62.6 (2.0)	67.8 (1.9)	56.5 (2.0)
FOV(C)	63.9 (1.8)	63.1 (1.9)	61.4 (2.3)	56.4 (2.1)
FOV(H)	79.2 (2.2)	67.8 (3.0)	77.4 (2.3)	70.8 (2.6)
FOV(E)	56.4 (3.8)	56.0 (4.3)	57.4 (4.1)	35.7 (3.2)
Forr(3)	0.662 (0.034)	0.590 (0.034)	0.637 (0.035)	0.610 (0.033)
Forr(C)	0.665 (0.015)	0.600 (0.017)	0.637 (0.021)	0.610 (0.014)
Forr(H)	0.703 (0.023)	0.651 (0.024)	0.705 (0.023)	0.649 (0.021)
Forr(E)	0.645 (0.031)	0.563 (0.038)	0.640 (0.036)	0.632 (0.023)

The values in the parentheses are the standard errors.

Thus, for the case of Matthews correlation coefficient, we replace $\mathbf{X}_S(j)$ and $\mathbf{Y}_S(j)$ by $\tilde{\mathbf{X}}_S(j) \equiv (\mathrm{Pr_O}(j;S),\ 1 - \mathrm{Pr_O}(j;S))$ and $\tilde{\mathbf{Y}}_S(j) \equiv (\mathrm{Pr_P}(j;S),\ 1 - \mathrm{Pr_P}(j;S))$ to obtain the two-state fuzzy correlation coefficient, abbreviated as Forr score in this work:

$$\mathrm{Forr}(S) = \frac{\sum_{j=1}^{N}(\tilde{\mathbf{X}}_S(j)\cdot\tilde{\mathbf{Y}}_S(j) - \bar{\tilde{\mathbf{X}}}_S\cdot\bar{\tilde{\mathbf{Y}}}_S)}{\sqrt{\sum_{k=1}^{N}(\tilde{\mathbf{X}}_S(k)^2 - \tilde{\mathbf{X}}_S^2)}\sqrt{\sum_{m=1}^{N}(\tilde{\mathbf{Y}}_S(m)^2 - \tilde{\mathbf{Y}}_S^2)}} = \frac{N\tilde{N}_{T+}(S) - \tilde{O}_+(S)\tilde{P}_+(S)}{\sqrt{(NO'_+(S) - \tilde{O}_+^2(S))(NP'_+(S) - \tilde{P}_+^2(S))}} \quad (15)$$

where

$$\tilde{O}_+(S) \equiv \sum_{j=1}^{N}\mathrm{Pr_O(j;S)},$$
$$O'_+(S) \equiv \sum_{j=1}^{N}\mathrm{Pr_O(j;S)}^2,$$
$$\tilde{P}_+(S) \equiv \sum_{j=1}^{N}\mathrm{Pr_P(j;S)},$$
$$P'_+(S) \equiv \sum_{j=1}^{N}\mathrm{Pr_P}(j;S)^2,$$
$$\tilde{N}_{T+}(S) \equiv \mathrm{Pr_P}(j;S)\,\mathrm{Pr_O}(j;S). \quad (16)$$

Similarly, for the three-state fuzzy correlation coefficient, we replace $\mathbf{X}(j)$ and $\mathbf{Y}(j)$ by $\tilde{\mathbf{X}}(j) \equiv (\mathrm{Pr_O}(j;\mathrm{C}),\ \mathrm{Pr_O}(j;\ \mathrm{H}),\ \mathrm{Pr_O}(j;\ \mathrm{E}))$ and $\tilde{\mathbf{Y}}(j) \equiv (\mathrm{Pr_P}(j;\mathrm{C}),\ \mathrm{Pr_P}(j;\mathrm{H}),\ \mathrm{Pr_P}(j;\ \mathrm{E}))$ to get

$$\mathrm{Forr}(3) = \frac{\sum_{j=1}^{N}(\tilde{\mathbf{X}}(j)\cdot\tilde{\mathbf{Y}}(j) - \tilde{\mathbf{X}}\cdot\tilde{\mathbf{Y}})}{\sqrt{\sum_{k=1}^{N}(\tilde{\mathbf{X}}(k)^2 - \tilde{\mathbf{X}}^2)}\sqrt{\sum_{m=1}^{N}(\tilde{\mathbf{Y}}(m)^2 - \tilde{\mathbf{Y}}^2)}} = \frac{N\tilde{N}_{\mathrm{corr}}(3) - \sum_S \tilde{O}_+(S)\tilde{P}_+(S)}{\sqrt{(N\tilde{N}_O - \sum_{S'\in\{\mathrm{C,H,E}\}}\tilde{O}_+(S')^2)(N\tilde{N}_P - \sum_{S''\in\{\mathrm{C,H,E}\}}\tilde{P}_+(S'')^2)}} \quad (17)$$

where

$$\tilde{N}_O \equiv \sum_{S\in\{\mathrm{C,H,E}\}} O'_+(S) = \sum_{j=1}^{N}\sum_{S\in\{\mathrm{C,H,E}\}}\mathrm{Pr_O}(j;S)^2$$
$$\tilde{N}_P \equiv \sum_{S\in\{\mathrm{C,H,E}\}} P'_+(S) = \sum_{j=1}^{N}\sum_{S\in\{\mathrm{C,H,E}\}}\mathrm{Pr_P}(j;S)^2 \quad (18)$$

By construction, the fuzzy correlation coefficients (15,17) reduce to the crisp correlation coefficients (13,14) when the estimated probabilities consist of 0s and 1s only.

For the example in Figure 1, we have

$$N_{\mathrm{corr}}(3) = N_{\mathrm{corr}}(C) = N_{\mathrm{corr}}(H) = 7,$$
$$N_{\mathrm{corr}}(E) = 12.$$

Also,

$$O_+(C) = 4, \quad O_+(H) = 8, \quad O_+(E) = 0,$$
$$P_+(C) = 5, \quad P_+(H) = 7, \quad P_+(E) = 0.$$

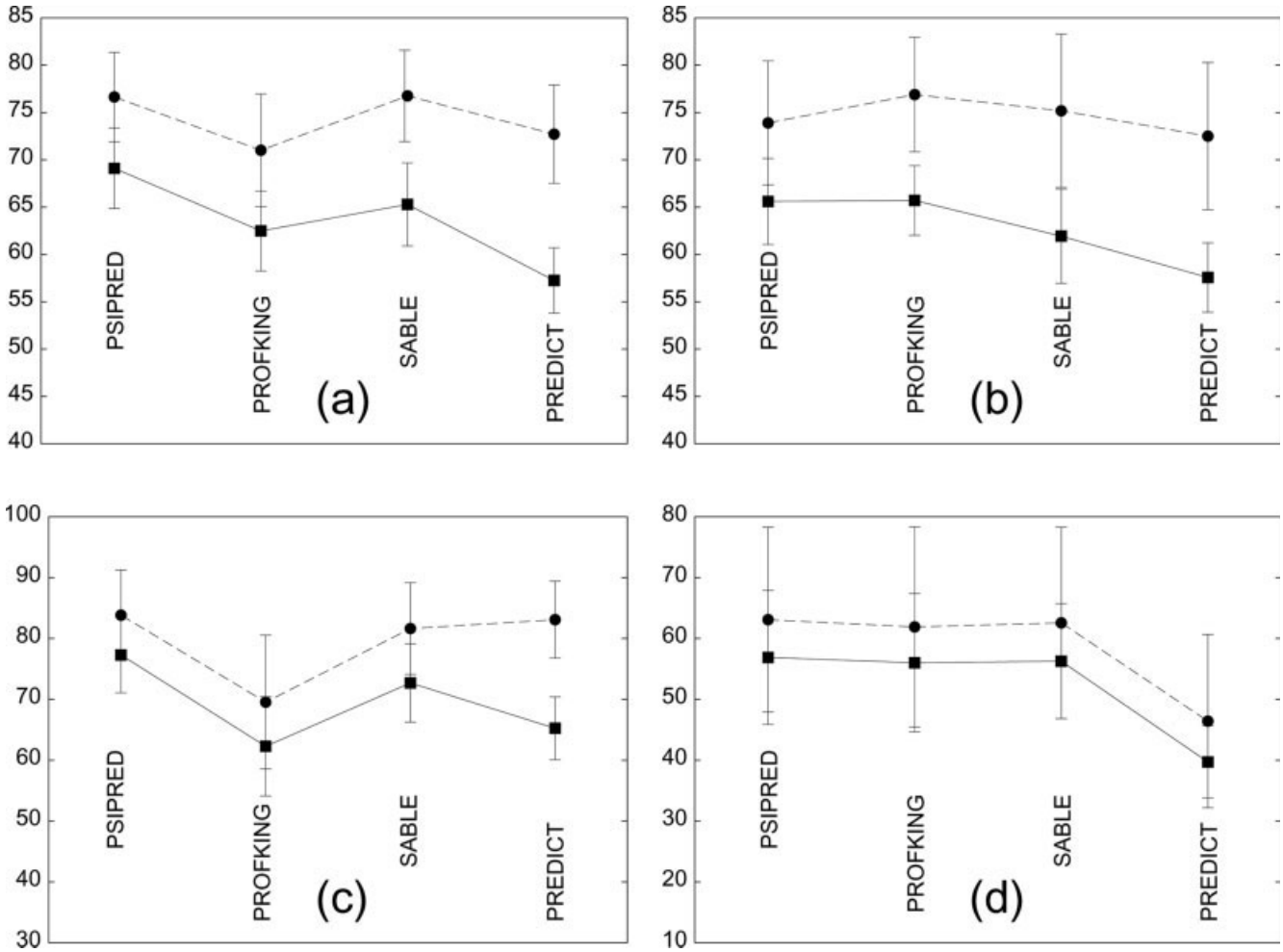

Fig. 2. Plots of average Q (dashed lines with filled circles) and F (solid lines with filled boxes) scores for the four prediction methods, for (a) three states, (b) coil, (c) helix, and (d) extended β-sheet. In these and the following figures, error bars are 4.002 × stderr (See text).

Therefore,

$$\mathrm{Corr}(C) = \mathrm{Corr}(H) = \mathrm{Corr}(3) = \frac{12 \times 7 - 4 \times 5 - 8 \times 7}{\sqrt{(12^2 - 4^2 - 8^2)(12^2 - 5^2 - 7^2)}} = 0.120$$

and Corr(E) is undefined since both the numerator $12 \times 12 - 0 \times 0 - 12 \times 12 = 0$ and the first factor of the denominator $\sqrt{12^2 - 0^2 - 12^2} = 0$ vanish. Introducing the fuzziness, we have

$$\tilde{O}_+(C) = 0.8 + 0.2{\times}3 + 0.8 + 1.0{\times}2 = 4.2,$$
$$\tilde{O}_+(H) = 0.2 + 0.8{\times}3 + 1.0{\times}5 + 0.2 = 7.8,$$
$$\tilde{O}_+(E) = 0.0,$$
$$\tilde{P}_+(C) = 0.8 + 0.5{\times}3 + 0.1{\times}5 + 0.2{\times}2 + 0.8 = 4.0,$$
$$\tilde{P}_+(H) = 0.1 + 0.4{\times}3 + 0.8{\times}5 + 0.5{\times}2 + 0.1 = 6.4,$$
$$\tilde{P}_+(E) = 0.1{\times}9 + 0.3{\times}2 + 0.1 = 1.6,$$

$$O'_+(C) = 0.8^2 + 0.2^2 \times 3 + 0.8^2 + 1.0^2 \times 2 = 3.40,$$
$$O'_+(H) = 0.2^2 + 0.8^2 \times 3 + 1.0^2 \times 5 + 0.2^2 = 7.00,$$
$$O'_+(E) = 0.0,$$
$$P'_+(C) = 0.8^2 + 0.5^2{\times}3 + 0.1^2 \times 5 + 0.2^2 \times 2 + 0.8^2 = 2.16,$$
$$P'_+(H) = 0.1^2 + 0.4^2 \times 3 + 0.8^2{\times}5 + 0.5^2{\times}2 + 0.1^2 = 4.20,$$
$$P'_+(E) = 0.1^2 \times 9 + 0.3^2 \times 2 + 0.1^2 = 0.28,$$
$$\tilde{N}_O = O'_+(C) + O'_+(H) + O'_+(E) = 10.40.$$
$$\tilde{N}_P = P'_+(C) + P'_+(H) + P'_+(E) = 6.64,$$

and

$$\tilde{N}_{\mathrm{T}+}(C) = 0.8^2 + 0.2 \times 0.5 \times 3 + 0.8 \times 0.2 + 0.20 + 0.80 = 2.10,$$
$$\tilde{N}_{\mathrm{T}+}(H) = 0.2 \times 0.1 + 0.8 \cdot 0.4 \times 3 + 1.0 \times 0.8 \times 5 + 0.2 \times 0.5 = 5.08$$
$$\tilde{N}_{\mathrm{T}+}(E) = 0.0,$$
$$\tilde{N}_{\mathrm{corr}}(3) = 2.10 + 5.08 = 7.18.$$

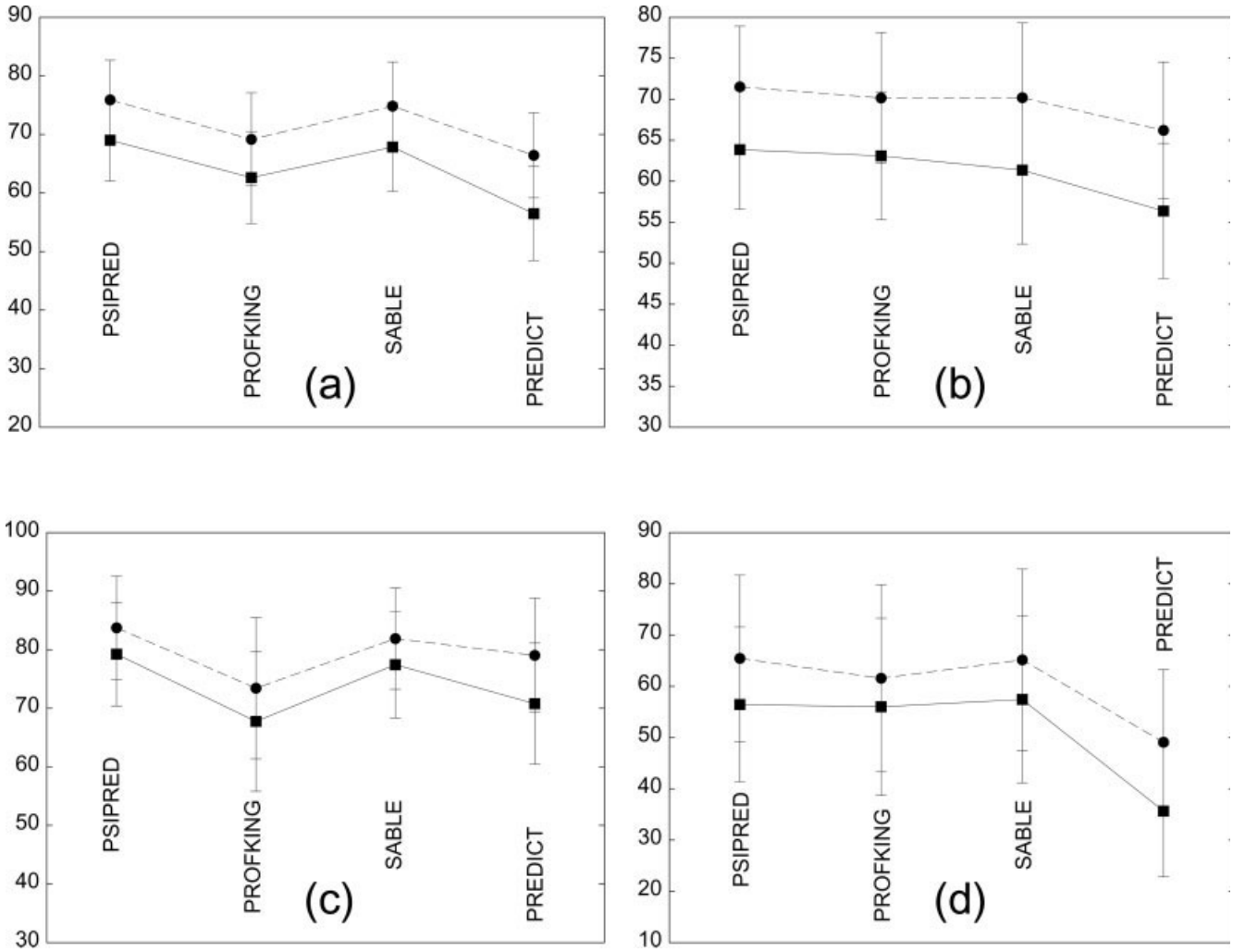

Fig. 3. Plots of average SOV (dashed lines with filled circles) and FOV (solid lines with filled boxes) scores for the four prediction methods, for (a) three states, (b) coil, (c) helix, and (d) extended β-sheet.

Therefore,

$$\mathrm{Forr}(C) = \frac{12 \times 2.10 - 4.2 \times 4.0}{\sqrt{(12 \times 3.40 - 4.2^2)(12 \times 2.16 - 4.0^2)}} = 0.554.$$

$$\mathrm{Forr}(H) = \frac{12 \times 5.08 - 7.8 \times 6.4}{\sqrt{(12 \times 7.00 - 7.8^2)(12 \times 4.20 - 6.4^2)}} = 0.747,$$

$$\mathrm{Forr}(3) = \frac{12 \times 7.18 - 4.2 \times 4.0 - 7.8 \times 6.4}{\sqrt{(12 \times 10.40 - 4.2^2 - 7.8^2)(12 \times 6.64 - 4.0^2 - 6.4^2 - 1.6^2)}} = 0.636. \quad (19)$$

where again, Forr(E) is undefined since both its numerator and denominator vanish. The scores for the example of Figure 1 are summarized in Table I.

RESULTS AND DISCUSSION

As an example of application of the fuzzy measures introduced in this work, I assessed performances of four secondary structure prediction methods that estimate probabilities for all the secondary structure classes, and are relatively easy to install and run on a local computer, PSIPRED (v2.3)[4], PROFking (v1.0)[5], SABLE (v2.0)[6,7], and PREDICT (v1.0)[8]. The test set used was EVA common Set 1[37]. After removing proteins with chain breaks, 76 proteins remain in the set. It should be emphasized that the result is not to be considered as an extensive test of these prediction methods, and the actual performances of the prediction algorithms depend on their versions and the set of proteins used for the test.

The average values of Q, SOV, Corr, F, FOV, and Forr measures for the four methods, for each of the three secondary structural classes and the overall performance, are shown in Table II and Figures 2–4, and the three-state fuzzy scores for individual proteins are also plotted against their crisp counterparts in Figures 5–7. The numbers in the parentheses in Table II are the standard errors, defined by

$$\mathrm{stderr} \equiv \frac{s}{\sqrt{n}}, \quad (20)$$

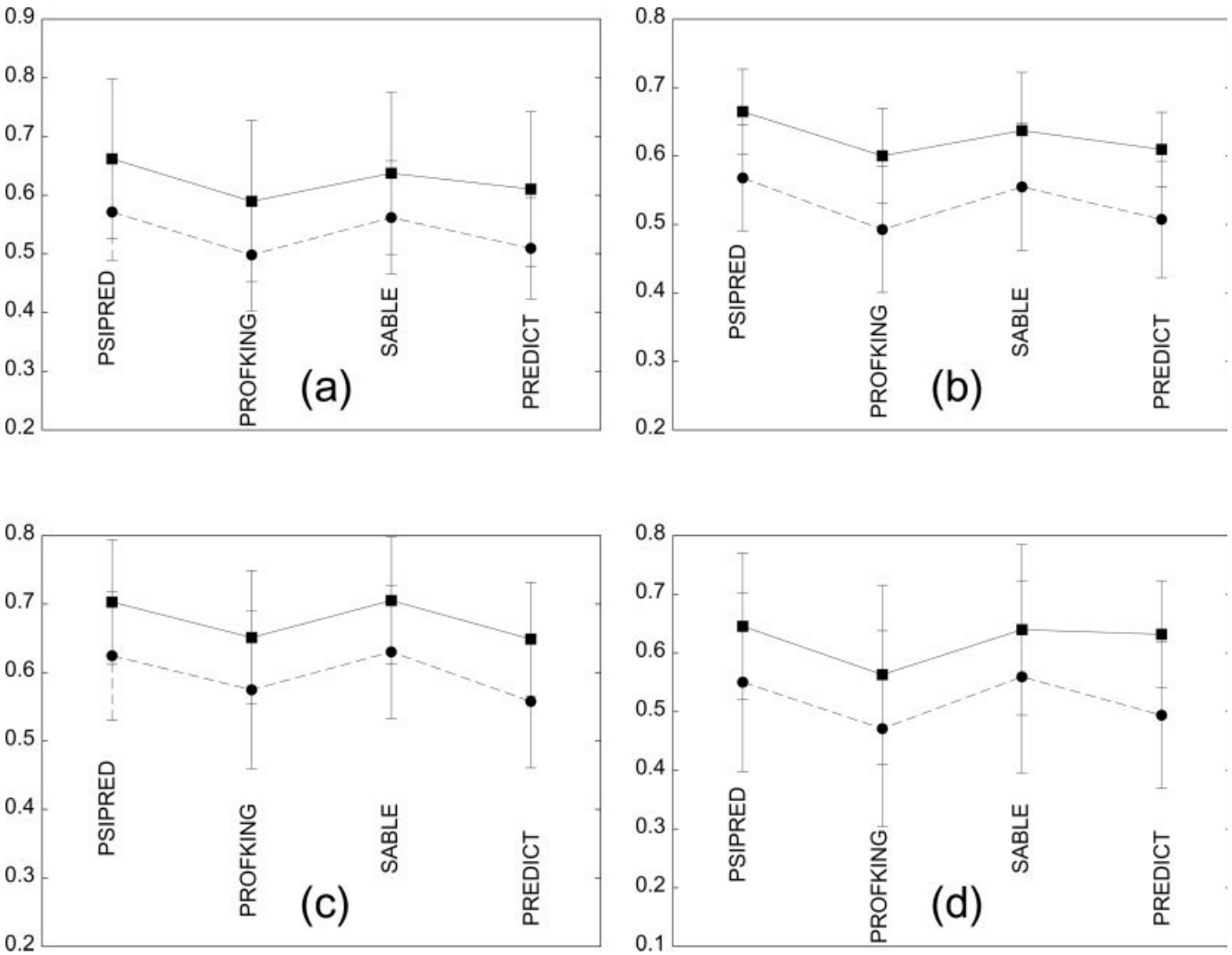

Fig. 4. Plots of average Corr (dashed lines with filled circles) and Forr (solid lines with filled boxes) scores for the four prediction methods, for (a) three states, (b) coil, (c) helix, and (d) extended β-sheet.

where the standard deviation is

$$s \equiv \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}, \tag{21}$$

with X_i denoting a numerical value of a measure, $\bar{X}$ its average over the protein chains, whose number is $n(=60)$. The widths of the error bars in the Figures 2–4 are $2 \times t_{59,0.975} \times \text{stderr} = 4.002 \times \text{stderr}$, which correspond to the 95% confidence intervals for the true mean values of the measures if the sample averages follow the Gaussian distribution. Here, $t_{59,0.975} = 2.001$ is the 97.5 percentile of the student t_{59}-distribution. The fuzzy prediction results were evaluated by comparing with the DSSPcont assignments.

We observe that Q_{E}, F(E), SOV(E), and FOV(E) scores for PREDICT are lower than those for the other methods, implying low sensitivity for the secondary structural class E. However, values of the crisp and fuzzy two-state correlation coefficients, which measure both the sensitivity and specificity at the same time, are comparable to those for the other methods.

We note that although F and FOV scores are lower than their crisp counterparts on average, giving an impression that the performance of the fuzzy prediction is worse than that of the crisp prediction, Forr scores are higher than Corr scores. That is, there is more correlation between the fuzzy prediction and observation, than between their crisp counterparts. In fact, it might not be truly meaningful to directly compare the numerical value of a crisp measure with that of the fuzzy counterpart, except for Corr and Forr scores, both of which derive from a single well-defined quantity in statistical theory, the Pearson correlation coefficient.

Despite relatively high correlations between the crisp measures and their fuzzy counterparts (See Figures 5–7 and their captions. See also Figures 2–4), it should be emphasized that the crisp and fuzzy measures are designed to assess different properties of a secondary

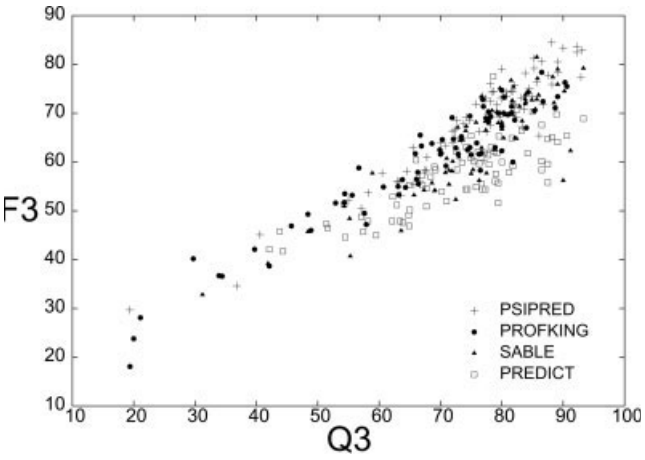

Fig. 5. Plot of Q_3 and F_3 scores for the four prediction methods. The correlations between these measures are 0.94, 0.96, 0.86, and 0.81 for PSIPRED, PROFKING, SABLE, and PREDICT, respectively.

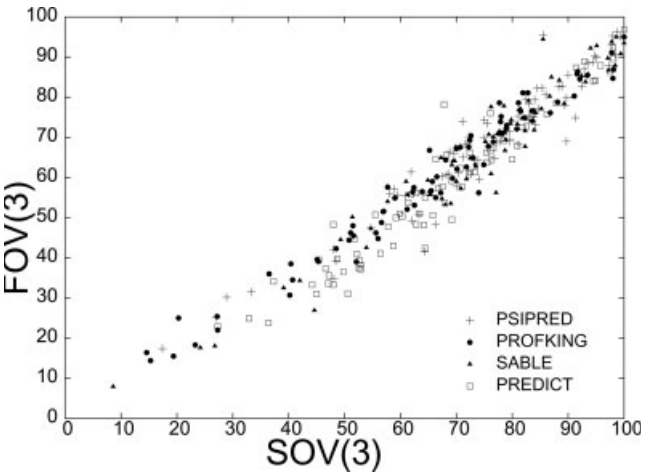

Fig. 6. Plot of SOV (3) and FOV (3) scores for the four prediction methods. The correlations between these measures are 0.96, 0.98, 0.97, and 0.96 for PSIPRED, PROFKING, SABLE, and PREDICT, respectively.

structure prediction algorithm, and fuzzy measures contain information that cannot be obtained from crisp measures. In particular, in the case of three-state correlation, there are protein chains with values of Forr(3) higher than Corr(3), as can be seen from Figure 7, providing examples where assessments using fuzzy measures are different from those using crisp ones. If one intends to use only the final predicted secondary structures, the fuzzy measures are not particularly useful for assessing secondary structure prediction algorithms. However, to fully utilize the estimated probabilities of the secondary structures produced by the fuzzy prediction method, and continuous assignment of observed secondary structure such as DSSPcont, it is more reasonable to use fuzzy prediction methods with high fuzzy scores. Therefore, the fuzzy measures should be considered as being complementary to the crisp measures rather than being replacements.

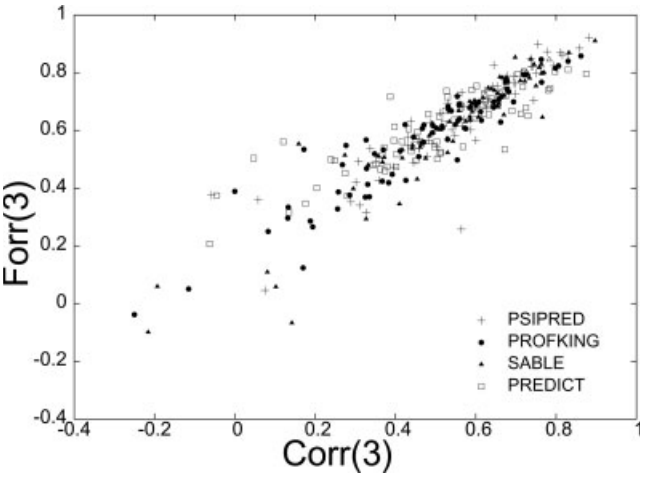

Fig. 7. Plot of Corr (3) and Forr (3) scores for the four prediction methods. The correlations between these measures are 0.89, 0.94, 0.93, and 0.86 for PSIPRED, PROFKING, SABLE, and PREDICT, respectively.

The automatic server and the software for fuzzy prediction assessment using the fuzzy measures are available at http://bioinfo.ssu.ac.kr/~jul/fuzzy/fuzzy.htm. When the estimated probabilities for secondary structural classes are not provided, the server and software produce the crisp measures as outputs. I hope these facilities encourage wide use of the newly invented fuzzy measures for assessments of fuzzy predictions of secondary structure.

ACKNOWLEDGMENTS

The author thanks Jooyoung Lee, Seung-Yeon Kim, and Jaehyun Sim for useful discussions.

REFERENCES

1. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 2005;59:467–475.
2. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;34:508–519.
3. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
4. Jones D. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
5. Ouali M, King R. Cascaded multiple classifiers for secondary structure prediction. Protein Sci 1999;9:1162–1176.
6. Porollo A, Adamczak R, Wagner M, Meller J.Maximum feasibility approach for consensus classifiers: applications to protein structure prediction.CIRAS (conference proceedings) 2003. pp 75–76.
7. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 2004;56:753–767.
8. Joo K, Kim I, Lee J, Kim ,S-Y, Lee SJ, Lee J. Prediction of the secondary structure of proteins using PREDICT, a nearest neighbor method on pattern space. J Korean Phys Soc 2004;45:1441–1449.
9. Kloczkowski A, Ting KL, Jernigan RL, Garnier J. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins 2002;49:154–166.

10. Meiler J, Mueller M, Zeidler A, Schmaeschke F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. J Mol Model 2001;7:360–369.
11. Pollastri G, McLysaght A. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 2004;21:1719,1720.
12. Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. Proteins 2002;46:197–205.
13. Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. Methods Enzymol 1996;266:525–539.
14. Rost B, Liu J. The PredictProtein server. Nucleic Acids Res 2003;31:3300–3304.
15. Sadeghi M, Parto S, Arab S, Ranjbar B. Prediction of protein secondary structure based on residue pair types and conformational states using dynamic programming algorithm. FEBS Lett 2005; 579:3397–3400.
16. Sen TZ, Jernigan RL, Garnier J, Kloczkowski A. GOR V server for protein secondary structure prediction. Bioinformatics 2005; 21:2787,2788.
17. Ward JJ, McGuffin LJ, Buxton BF, Jones DT. Secondary structure prediction with support vector machines. Bioinformatics 2003;19:1650–1655.
18. Wood MJ, Hirst JD. Protein secondary structure prediction with dihedral angles. Proteins 2005;59:476–481.
19. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
20. Zadeh LA. Fuzzy Sets. Inf Control 1965;8:338–353.
21. Bezdek JC. Pattern recognition with fuzzy objective function algorithms. New York: Plenum; 1981.
22. Keller JM, Gray R, Givens JA, Jr. A fuzzy k-nearest neighbor algorithm. IEEE Trans Syst Man Cybernet 1985;15:580–585.
23. Anderson CAF, Palmer AG, Brunak S, Rost B. Continuum secondary structure captures protein flexibility. Structure 2002;10:175–184.
24. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci USA 2003;100:12105–12110.
25. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.
26. Jones DT. Predicting novel protein folds by using FRAGFOLD. Proteins 2001;45(Suppl 5):127–132.
27. Jones DT, McGuffin LJ. Assembling novel protein folds from super-secondary structural fragments. Proteins 2003;53(Suppl 6): 480–485.
28. Lee J, Kim S-Y, Joo K, Kim I, Lee J. Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. Proteins 2004; 56:704–714.
29. Lee J, Kim S-Y, Lee J. Protein structure prediction based on fragment assembly and parameter optimization. Biophys Chem 2005;115:209–214.
30. Lee J, Kim S-Y, Lee J. Protein structure prediction based on fragment assembly and β-strand pairing energy function. J Korean Phys Soc 2005;46:707–712.
31. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol 1994;235:13–26.
32. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16:412–424.
33. Bajic VB. Comparing the success of different prediction software in sequence analysis: a review. Brief Bioinform 2000;1:214–228.
34. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measurement for protein secondary structure prediction assessment. Proteins 1999;34:220–223.
35. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975; 405:442–451.
36. Gorodkin J. Comparing two K-category assignment by a K-category correlation coefficient. Comput Biol Chem 2004;28:367–374.
37. Koh IY, Eyrich V, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Grana O, Pazos F, Valencia A, Sali A, Rost B. EVA: evaluation of protein structure prediction servers. Nucleic Acids Res 2003;31:3311–3315.