A Modified Definition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment

Adam Zemla,[1] Česlovas Venclovas,[1] Krzysztof Fidelis,[1*] and Burkhard Rost[2]
[1]*Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California*
[2]*European Molecular Biology Laboratory, Heidelberg, Germany*

ABSTRACT **We present a measure for the evaluation of secondary structure prediction methods that is based on secondary structure segments rather than individual residues. The algorithm is an extension of the segment overlap measure Sov, originally defined by Rost et al. (J Mol Biol 1994;235:13–26). The new definition of Sov corrects the normalization procedure and improves Sov's ability to discriminate between similar and dissimilar segment distributions. The method has been comprehensively tested during the second Critical Assessment of Techniques for Protein Structure Prediction (CASP2). Here, we describe the underlying concepts, modifications to the original definition, and their significance. Proteins 1999;34:220–223.**

INTRODUCTION

Although segmented in nature, the secondary structure of proteins has long been predicted and analyzed on a per residue basis. Often, this does not capture the "usefulness" of secondary structure predictions. For example, when predictions of secondary structure are used in a subsequent prediction of protein tertiary (or 3D, i.e., three-dimensional) structure, the correct identification of type and location of secondary structure elements appears often more critical than the assignment of conformational state at the level of individual residues. Indeed, Q_3, the traditional per-residue measure, defined as a fraction of residues predicted correctly in three conformational states (helix, strand, and other, i.e. non-regular structure), many times leads to a distorted picture of how well a prediction corresponds to the real 3D structure (e.g. Refs. 1–3). For example, assigning the entire myoglobin chain as a single helix gives a per-residue score of ca. 80%.[4] Despite a very unrealistic nature of such an assignment, its rating would exceed the overall accuracy of the presently most successful prediction methods (e.g. Ref. 5).

Another issue that affects the assessment of secondary structure prediction is associated with conformational variation observed at secondary structure segment ends. Even for homologous protein pairs with very similar sequences, elements of secondary structure frequently differ in the exact position of their ends (e.g. Refs. 1, 6). Thus, it may not be critical to predict segment ends exactly. Since the overall 3D structure readily accommodates such limited variation, it seems entirely reasonable to make a similar allowance at the level of secondary structure assessment. This line of thought may be extended even further by considering the secondary structure classification itself. In this simplified view of protein structure, it is sometimes difficult to make an unequivocal conformational assignment, especially at segment ends. In effect, algorithms which translate a given 3D structure into a string of secondary structure symbols also tend to differ in the exact placement of segment ends.[7]

The necessity for a structurally more meaningful measure of secondary structure prediction accuracy has been pointed out by a number of authors.[1–4, 8–17] Such a structurally oriented measure should, at least, account for the following:

- Type and position of secondary structure segments rather than a per-residue assignment of conformational state.
- Natural variation of segment boundaries among families of homologous proteins.
- Ambiguity in the position of segment ends due to differences of approach in secondary structure classification.

All these points were addressed by the segment overlap measure Sov.[1] Consequently, it has been able to effectively capture structurally important features (secondary structure segments) and to reduce the significance of those that are structurally less important (small variation of segment length and position). As such, Sov has been selected as one of the prediction evaluation criteria for the second Critical Assessment of Techniques for Protein Structure Prediction (CASP2).[18] While extensive testing during this large scale prediction experiment has fully substantiated the requirements outlined above, it revealed the necessity to amend the original definition. As first proposed, the measure does not have a strictly defined upper limit and thus it is not directly comparable with other measures of prediction accuracy. In this paper we introduce modifications which remedy this deficiency.

Adam Zemla is on leave from Institute of Mathematics, Polish Academy of Sciences, Śniadeckich 8, Warsaw, Poland.
Česlovas Venclovas's permanent address is Institute of Biotechnology, Graičiūno 8, 2028 Vilnius, Lithuania.
*Correspondence to: Krzysztof Fidelis, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551. E-mail: kaf@sb1.llnl.gov
Received 20 April 1998; Accepted 14 September 1998

DEFINITION OF THE NEW MEASURE

We begin with a relatively simple case by first focusing on the single state secondary structure assignments, for example of helices (Eqs. 1–3). We then generalize the definition for multi-state assignments; specifically, in Eqs. 4, 5, and 3 we give the definition for three-state assignments: helix (H), strand (E), and other, i.e. non-regular structure, which for historical reasons is being referred to as coil (C). Our three-state definition replaces the old definition of Sov (subsequently referred to as Sov'94, Rost et al.,[1] Eq. 1, and also A3 in Appendix I). We will use s_1 and s_2 to denote segments of secondary structure in conformational state *i* (i.e. H, E, or C). Segments s_1 and s_2 correspond to the two secondary structure assignments being compared. The first assignment is considered a reference and is typically based on experiment, the second is the one being evaluated. The two assignments are further referred to as "observed" and "predicted," respectively. Let (s_1,s_2) denote a pair of overlapping segments, S(*i*) - the set of all the overlapping pairs of segments (s_1,s_2) in state *i*, i.e.:

$$S(i) = \{(s_1, s_2): s_1 \cap s_2 \neq \varnothing, \quad s_1 \text{ and } s_2 \text{ are both in conformational state } i\},$$

S'(*i*) - the set of all segments s_1 for which there is no overlapping segment s_2 in state *i*, i.e.:

$$S'(i) = \{s_1: \forall\, s_2,\ s_1 \cap s_2 = \varnothing, \quad s_1 \text{ and } s_2 \text{ are both in conformational state } i\}.$$

For state *i* the segment overlap measure is then defined as:

$$Sov(i) = 100 \times \frac{1}{N(i)} \sum_{S(i)} \left[\frac{\text{minov}(s_1, s_2) + \delta(s_1, s_2)}{\text{maxov}(s_1, s_2)} \times len(s_1) \right] \quad (1)$$

with the normalization value N(*i*) defined as:

$$N(i) = \sum_{S(i)} len(s_1) + \sum_{S'(i)} len(s_1). \quad (2)$$

The sum in Eq. 1 and the first sum in Eq. 2 are taken over all the segment pairs in state *i* which overlap by at least one residue, the second sum in Eq. 2 is taken over the remaining segments in state *i* found in the reference assignment, len(s_1) is the number of residues in segment s_1, minov(s_1,s_2) is the length of the actual overlap of s_1 and s_2, i.e. for which both segments have residues in state *i*, maxov(s_1,s_2) is the total extent for which either of the segments s_1 and s_2 has a residue in state *i*, and δ (s_1,s_2) is defined as:

$$\delta(s_1, s_2) = \min\{(\text{maxov}(s_1, s_2) - \text{minov}(s_1, s_2)); \text{minov}(s_1, s_2); \text{int}(\text{len}(s_1)/2); \text{int}(\text{len}(s_2)/2)\}, \quad (3)$$

where min{*x1; x2; x3;...; xn*} is the minimum of *n* integers.

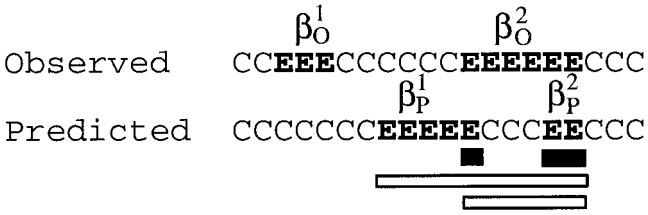

Fig. 1. Illustration of a Sov(E) calculation. Black and white bars correspond respectively to minov and maxov in the overlapping segment pairs from observed and predicted structures.

The measure defined in Eqs. (1–3) is easily extended to evaluate multi-state secondary structure assignments. In particular, for the three state case of helix (H), strand (E), and coil (C) we define:

$$Sov = 100 \times \left[\frac{1}{N} \sum_{i \in \{H,E,C\}} \sum_{S(i)} \frac{\text{minov}(s_1, s_2) + \delta(s_1, s_2)}{\text{maxov}(s_1, s_2)} \times len(s_1) \right] \quad (4)$$

where the normalization value *N* is a sum of *N(i)* over all three conformational states:

$$N = \sum_{i \in \{H,E,C\}} N(i) \quad (5)$$

The quality of match of each segment pair is taken as a ratio of the overlap of the two segments (minov(s_1,s_2)), and the total extent of that pair (maxov(s_1,s_2)). The definition allows to improve this ratio by extending the overlap by the value of δ(s_1,s_2). The normalization procedure assures that Sov values always are within range 0–100 and thus can be used in percentage scale to allow direct comparison with other prediction evaluation measures, for example Q_3. Compared to Sov'94 there are two specific changes. First the definition of δ (Eq. 3) is made symmetric with respect to observed and predicted segments (c.f. Rost et al.,[1] p. 22); second, the normalization factor N equal to the total number of residues is replaced by Eqs. 2 and 5. Properties of the new measure are further discussed in a separate section.

To illustrate the calculation of Sov let us consider a prediction given in Figure 1 and evaluate the strand assignment, i.e. calculate Sov(E). In the observed structure the first strand $\beta_O{}^1$ belongs to the set S'(E) because it does not produce any overlapping pair, the second strand produces two of them: ($\beta_O{}^2$, $\beta_P{}^1$) and ($\beta_O{}^2$, $\beta_P{}^2$). The value of Sov(E) is calculated as follows:

$$Sov(E) = 100 \times \frac{1}{6 + 6 + 3} \times \left(\frac{1 + 1}{10} + \frac{2 + 1}{6} \right) \times 6 = 28.0$$

As in Sov (Eq. 4) all three conformational states are assigned equal weight, coil regions are treated in the same way as strands or helices. Thus the value of Sov calculated

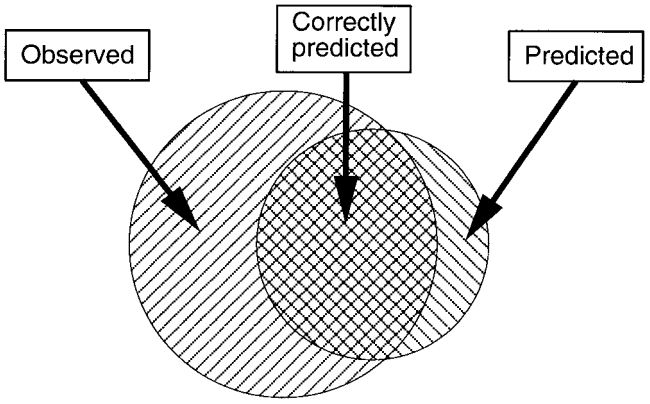

Fig. 2. Graphical representation of the difference between $Sov^{observed}$ and $Sov^{predicted}$. The overlap between two circles corresponds to the overlap between observed and predicted helices. Area of the circle overlap, however, occupies a different fraction of observed and predicted assignments.

for all three conformational states for this prediction is equal to 39.4.

$Sov^{observed}$ vs. $Sov^{predicted}$

So far, Sov has been defined to evaluate the correctness of segment prediction with respect to a reference assignment ($Sov^{observed}$). In addition to this basic measure it is possible to calculate an alternative version ($Sov^{predicted}$) which provides a value indicating what fraction of predicted segments is correct. The latter measure is calculated with s_1 standing for predicted segments and s_2 for observed, and corresponds to $Sov^{predicted}$ defined in Rost et al.,[1] as well as e.g. "probability of correct prediction" for a single conformational state in Kabsch and Sander.[19] $Sov^{predicted}$ is especially useful in methods development, for example in instances where an over- or underprediction of a particular state is suspected. Indeed, perhaps the easiest way to understand the relationship between these two measures is to analyze the prediction success for one particular conformational state, let's say helices (Fig. 2). In $Sov^{observed}$ the overlaps resulting from prediction of helical segments are evaluated against helices in the observed structure. In $Sov^{predicted}$, on the other hand, for the same overlaps the frame of reference is the predicted structure. With the exception of the case where predicted structure is the same as observed, the two versions do not necessarily produce the same results. It should be noted that for segments as well as for individual residues "fraction correct of predicted"[19] can be indicative of the method's predictive power only if it is used together with "fraction correct of observed." For example, if a protein is predicted to have only one helix, which happens to coincide with one of five helices in the real structure, then $Sov^{predicted}$ for helices is calculated as 100% or very close. This remarkable result, however, does not seem impressive at all when contrasted with the value of $Sov^{observed}$ for this prediction, which is very poor.

PROPERTIES OF THE NEW MEASURE

The new measure preserves the basic concepts of Sov'94, however it differs in two important respects: the normalization procedure and the definition of δ, a degree of variation allowed at the segment edges.

The new normalization value is calculated with respect to the segments of the reference assignment (observed), with every segment taken into account at least once. If for a given reference segment, more than one prediction segment has to be considered, the sum is extended accordingly, i.e. the reference segment is summed for each overlapping pair. Consequently, the normalization procedure lowers the prediction score for both erroneous partitioning and non-prediction of segments. The normalization procedure now reflects the pairwise nature of segment comparison; hence, the new Sov's range is from 0 to 100%, depending on prediction accuracy.

In the original Sov, the δ extension had been designed to allow some restricted variation at the edges of secondary structure elements thus putting emphasis on correct prediction of segments rather than on assignment of conformational state to individual residues. In the new definition the allowed variation is limited even further, such that in a pair of segments of observed and predicted secondary structure δ can never exceed one-half of the shorter segment. In practice, this means that if the segments have an extensive overlap, with only small differences at their edges, they are considered identical, i.e. the segment prediction is perfect. On the other hand, if the overlap of corresponding segments is only minor, it cannot be "extended" by any significant amount and produce an artificially improved score. We believe that the re-definition of δ in Eq. 3 strikes a reasonable compromise between a frozen and an excessively relaxed definition of segment overlap. Although without the δ extension (δ=0) Sov gives adequate results for distinguishing between poor and successful predictions, it completely ignores the natural variation at the edges of secondary structure segments.

Just as Q_3, the segment overlap measure treats helices, strands, and coil on an equal basis (three-state assignment). There are no arbitrary cutoffs on segment length, assuring a continuous, threshold-free assessment of prediction accuracy.

We demonstrate the performance of the new measure with several examples. First, let us examine the case of myoglobin predicted as a single helix. Q_3 for this prediction is 78%, but Sov gives a value of just 16%, reflecting the unrealistic nature of such an assignment. Sov not only penalizes wrong predictions but also performs an effective ranking across a full range of assignment quality. Let us now focus on five prediction examples (Table I) which were selected to demonstrate both the fallacy of Q_3, as well as the necessity to amend the original definition of Sov introduced in the 1994 paper.[1] Predictions 1 and 2 show that the new Sov can effectively discriminate between

TABLE I. Examples of Different Predictions Evaluated Against the Observed Structure Using Sov Defined in This Paper, Sov Defined Earlier,[1] and Q_3

		Sov	Sov ($\delta = 0$)	Sov'94	Sov'94 ($\delta = 0$)	Q_3
Observed	CHHHHHHHHHHC					
Prediction 1	CHCHCHCHCHCC	12.5	12.5	95.8	54.2	58.3
Prediction 2	CCCHHHHHCCCC	63.2	46.5	88.2	46.5	58.3
Prediction 3	CHHHCHHHCHHC	40.6	31.3	150.0	83.3	83.3
Prediction 4	CHHCCHHHHHCC	52.3	38.6	129.2	70.8	75.0
Prediction 5	CCCHHHHHHCCC	80.6	55.6	88.9	55.6	66.7

Segment overlap measures are calculated using both δ and $\delta = 0$.

unrealistic and quite accurate predictions while Q_3 fails. Values calculated with the originally defined Sov ("Sov'94" in Table I) are also not satisfactory, assigning a higher score to a multiply split helix (Prediction 1). Predictions 3–5 illustrate the case where the helix in the observed structure is predicted as either multiple helices (Predictions 3 and 4) or as a single helix. New Sov attributes the highest score to the structurally most consistent assignment, while both Q_3 and Sov'94 identify it as the worst prediction. In addition, this example demonstrates that in some cases the original Sov produces values outside the 0–100% range. Although very short sequences were chosen for the purpose of this illustration, examples presented here do in fact reflect more general characteristics of the measure. Nevertheless, probably the best test of the new measure is the real life prediction experiment. Results for all three algorithms discussed here have been obtained during the second Critical Assessment of Techniques for Protein Structure Prediction (CASP2).[18] Data for over 200 secondary structure predictions, made by a number of research groups on more than 20 different prediction targets is thus available. While specific examples are discussed in a special issue of PROTEINS (e.g. Zemla et al.,[20] in particular c.f. Figure 1 in this paper), all of the submitted predictions, their evaluations, as well as secondary structure assignments obtained from structures determined experimentally, can be examined with a WWW browser (http://PredictionCenter.llnl.gov/casp2/evaluation.html). In addition, the three measures may be compared with the help of a web server that will accept any user supplied prediction (http://PredictionCenter.llnl.gov/local/ss_eval/sspred_evaluation.html). The source code for the Sov program is also available from this site.

Assignment of secondary structure provides a relatively simple structural characterization of a protein. Even so, consistency with the tertiary structure is an important issue, one that should be addressed in prediction assessment. In this respect, the re-defined segment-based Sov provides an assessment of prediction quality that is clearly superior to a residue-based approach, such as Q_3. Although hardly a replacement for an insightful visual inspection, accurate computer algorithms are potentially useful in two important applications. First, they provide evaluation that is free from human bias, second they can be directly used in methods development, as well as other tasks requiring extensive evaluation. For example, in a large-scale prediction experiment, such as CASP, evaluation would be extremely difficult without consistently calculated, automatic measures. We feel that by providing a robust measure of secondary structure prediction quality the new Sov is particularly suited for these applications.

ACKNOWLEDGMENT

This work was performed in part under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

REFERENCES

1. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol 1994;235:13–26.
2. Thornton JM, Flores TP, Jones DT, Swindells MB. Prediction of progress at last. Nature 1991;354:105–106.
3. Defay T, Cohen F. Evaluation of current techniques for ab initio protein structure prediction. Proteins 1995;23:431–445.
4. Barton GJ. Protein secondary structure prediction. Curr Opin Struct Biol 1995;5:372–376.
5. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232: 584–599.
6. Russell RB, Barton GJ. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J Mol Biol 1993;234:951–957.
7. Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon J-P. Comparison of three algorithms for the assignment of secondary structure in proteins: The advantages of a consensus assignment. Protein Eng 1993;6:377–382.
8. Taylor WR, Thornton JM. Recognition of super-secondary structure in proteins. J Mol Biol 1984;173:487–514.
9. Schulz GE. A critical evaluation of methods for prediction of secondary structures. Ann Rev Biophys Chem 1988;17:1–21.
10. Biou V, Gibrat JF, Levin JM, Robson B, Garnier J. Secondary structure prediction: Combination of three different methods. Protein Eng 1988;2:185–191.
11. Cohen FE, Kuntz ID. Tertiary structure prediction. In: Fasman GD, editor. Prediction of protein structure and the principles of protein conformation. New York, London: Plenum; 1989. p. 647–705.
12. Benner SA. Predicting de novo the folded structure of proteins. Curr Opin Struct Biol 1992;2:402–412.
13. Sternberg MJE. Secondary structure prediction. Curr Opin Struct Biol 1992;2:237–241.
14. Zhang X, Mesirov JP, Waltz DL. Hybrid system for protein secondary structure prediction. J Mol Biol 1992;225:1049–1063.
15. Rost B, Sander C, Schneider R Progress in protein structure prediction? TIBS 1993;18:120–123.
16. Wang Z-X. Assessing the accuracy of secondary structure. Nat Struct Biol 1994;3:145–146.
17. Zhu Z-Y. A new approach to the evaluation of protein secondary structure predictions at the level of the elements of secondary structure. Protein Eng 1995;8:103–108.
18. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): Round II. Proteins Suppl 1997;1:2–6.
19. Kabsch W, Sander C. How good are predictions of protein secondary structure? FEBS Letters 1983;155:179–182.
20. Zemla A, Venclovas Č, Reinhardt A, Fidelis K, Hubbard TJ. Numerical criteria for the evaluation of ab initio predictions of protein structure. Proteins Suppl 1997;1:140–150.