JOURNAL OF CHEMICAL INFORMATION AND MODELING

Application Note

pubs.acs.org/jcim

conSSert: Consensus SVM Model for Accurate Prediction of Ordered Secondary Structure

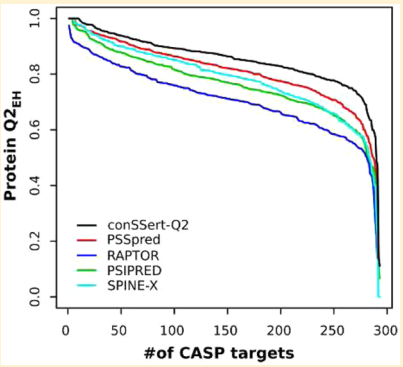
Chris A. Kieslich,[†,‡] James Smadbeck,[§] George A. Khoury,[§] and Christodoulos A. Floudas*[,†,‡]

[†]Artie McFerrin Department of Chemical Engineering and [‡]Texas A&M Energy Institute, Texas A&M University, College Station, Texas 77843, United States

[§]Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08544, United States

Supporting Information

ABSTRACT: Accurate prediction of protein secondary structure remains a crucial step in most approaches to the protein-folding problem, yet the prediction of ordered secondary structure, specifically beta-strands, remains a challenge. We developed a consensus secondary structure prediction method, conSSert, which is based on support vector machines (SVM) and provides exceptional accuracy for the prediction of beta-strands with QE accuracy of over 0.82 and a Q2-EH of 0.86. conSSert uses as input probabilities for the three types of secondary structure (helix, strand, and coil) that are predicted by four top performing methods: PSSpred, PSIPRED, SPINE-X, and RAPTOR. conSSert was trained/tested using 4261 protein chains from PDBSelect25, and 8632 chains from PISCES. Further validation was performed using targets from CASP9, CASP10, and CASP11. Our data suggest that poor performance in strand prediction is likely a result of training bias and not solely due to the nonlocal nature of beta-sheet contacts. conSSert is freely available for noncommercial use as a webservice: http://ares.tamu.edu/conSSert/.



1. INTRODUCTION

Sequence-based secondary structure prediction is an essential initial step in most approaches for prediction of protein tertiary structure (as reviewed in refs 1 and 2). The secondary structure prediction problem is typically simplified to a three class problem, concerned with specifying which of the three conformation types (α-helical, beta-strand, or nonregular) a given amino acid residue will adopt. Early methods focused on propensities for specific amino acids, or segments of amino acids, to form secondary structure, while substantial improvements in prediction accuracy were presented in the 1990s as databases of evolutionary information became more prevalent (as reviewed in refs 3 and 4). Due to the existence of chameleon sequences and inconsistencies in secondary structure assignment of experimental structures, the maximum theoretical accuracy for secondary structure prediction has been proposed to be ~88%.[5,6] However, the accuracies of current state-of-art methods have plateaued at ~80%.[7−13] Of the secondary structure types, beta-strand (strand) has historically been the most difficult to predict, which has typically been explained to be a result of the nonlocal nature of the contacts that form beta-sheets. This is despite the finding that balanced training leads to balanced predictions.[3] Additionally, it has been noted that reported per-residue accuracy, which is the standard measure of performance, tends to favor methods over predicting nonregular structure (coil),[3] due to the fact that coil is the most frequently observed secondary structure. Therefore, there is a clear need for secondary structure prediction methods that are able to accurately predict ordered secondary structure, specifically strands and helices.

Top performing approaches[7−12] utilize PSI-BLAST[14] searches of large databases of protein amino acid sequences to produce profiles, with Jones[15] being the first to introduce this approach. Also, the use of neural networks is fairly ubiquitous within the field of protein secondary structure prediction.[3,7,8,11,15,16] Many of the recent advances have to do with incorporation of additional features/aspects of protein secondary structure. SPINE-X[11] couples the prediction of secondary structure with the prediction of solvent accessibility and backbone torsion angles. SCORPION[7] includes pseudo-potentials based on analysis of high-order inter-residue interactions in addition to PSI-BLAST profiles.

Despite the prevalence of neural networks in the field of protein secondary structure prediction, several previous studies have successfully applied support vector machines[17] (SVMs) to the secondary structure prediction problem.[18−21] Support vector machines were originally designed for binary classification but can be extended to multiclass problems by combining multiple binary classifiers, with one-vs-all and one-vs-one being the most popular approaches. A major advantage of SVMs, and a reason SVM models should be considered for protein secondary structure prediction methods, is a guarantee of global optimality for the trained models.[17]

Received: September 14, 2015

DOI: 10.1021/acs.jcim.5b00566
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

One widely adopted approach for improving secondary structure prediction accuracy is the development of consensus methods by combining the predictions of multiple preexisting methods.[13,20,22−25] Our group has previously developed CONCORD,[13] a consensus secondary structure prediction method based on mixed integer linear optimization. CONCORD uses as input the predictions from seven methods, SSpro,[26] DSC,[27] PROF,[28] PROFphd,[16] PSIPRED,[15] Predator,[29] and GorIV,[30] using both the secondary structure prediction and the confidence score from each. Of the existing consensus methods, one has utilized support vector machines, however, their focus was on the development of multiclass SVM models.[20]

Many previous studies have proposed layered approaches for protein secondary structure prediction in which models designed to identify different characteristics are used as features for the next layer of the prediction scheme.[3,4] While some consensus methods employ a simple voting scheme,[24] in this study, we have developed an additional layer in the prediction scheme in which SVM models are used to combine existing methods. We have selected four methods as input into our consensus model, PSSpred,[9] PSIPRED,[10] RAPTOR,[12] and SPINE-X,[11] which were selected since they represent the latest advances in the field and have varied accuracies in terms of prediction of coil, strand, or helix. The consensus SVM models use as features the predicted probabilities of coil, helix, and strand from each of the four selected methods. We have trained/tested two sets of SVM binary classifiers to emphasize different types of accuracies and have evaluated our models through competing in CASP11.

2. METHODS

2.1. Data Sets. Initial training and benchmarking was performed based on 4261 protein chains from PDBSelect25.[31] The data set was sorted according to CONCORD[13] Q3 accuracy, and separated in to five groups of comparable difficulty and composition containing ∼850 sequences each. The five identified subsets of proteins were used as testing sets for our 5-fold cross-validation (see section 2.3 for more details). Table S1 summarizes the compositions of the PDBSelect25 training/test sets, including the number of residues of each secondary structure type. An additional data set containing 8632 targets from the PISCES[32] protein culling server based on a percent identity cutoff of 25%, resolution cutoff of 3.0 Å, and an *R*-factor cutoff of 1.0 was used for further evaluation of the robustness of developed models. The PISCES data set was sorted by sequence length, and separated into five test sets containing ∼1700 sequences each for an additional run of 5-fold cross validation (Table S6). Finally, an independent test set containing 99 targets from CASP9, 95 targets from CASP10, and 99 targets from CASP11 was also used. Secondary structure assignment was performed using DSSP.[33] On the basis of DSSP, types G, H, and I were classified as helix, types B and E as strands, and all others as coil.

2.2. SVM Models. The developed consensus SVM models used as features the probabilities for strand, helix, and coil from four established methods: PSSpred,[9] PSIPRED,[10] RAPTOR,[12] and SPINE-X.[11] All developed SVM models use the same 12 features, while training regimes were varied. The R[34] package e1071,[35] which is a wrapper for the LIBSVM C library,[36] was used for training the SVM models using the radial basis function kernel. We developed SVM models using the main approaches for addressing multiclass predictions: one-vs-all and one-vs-one. The developed one-vs-all model was designed to emphasize Q2-EH and will be referred as conSSert-Q2. On the other hand, the one-vs-one model was designed to improve Q3 and will be referred as conSSert-Q3.

conSSert-Q2 consists of three binary SVM classifiers of the type coil/not-coil, strand/not-strand, or helix/not-helix. Each classifier was trained on sets of samples composed of 50% of the desired class, and 25% of each of the remaining classes. To ensure that each classifier was trained on the same number of samples, the total number of samples was set to twice the number of strand residues, since strand is the least frequent type of secondary structure. Binary SVM models were trained to predict class probabilities based on the built-in functionality of e1071, which fits the decision values to a logistic distribution using maximum likelihood. To combine the three one-vs-all binary classifiers, the class with the highest probability is assigned. conSSert-Q3, on the other hand, was trained using all residues of each training set, maintaining the natural occurrences of coil, strand, and helix within each set. To train conSSert-Q3, the default multiclass functionality of e1071 and LIBSVM was used. Therefore, conSSert-Q3 is based on three one-vs-one binary classifiers (helix/coil, strand/coil, and helix/strand) that are combined using a voting scheme. Postprocessing, as previously described for CONCORD,[13] was performed for all predictions from conSSert-Q2 and conSSert-Q3. The postprocessing seeks to remove isolated secondary structure elements that arise from the combination of multiple predictions by considering multiple consecutive predictions and amino acid type.

2.3. Evaluation Metrics. The classical accuracy in the field of protein secondary structure prediction is the per-residue measure Q3, which is computed as follows:

$$\mathrm{Q3} = \frac{\sum_{i\in \mathrm{E,H,C}} \sum_{p}^{N_{\mathrm{prot}}} \mathrm{TP}_i(p)}{\sum_{i\in \mathrm{E,H,C}} \sum_{p}^{N_{\mathrm{prot}}} N_i(p)} \quad (1)$$

In this expression, $\mathrm{TP}_i(p)$ is the number of true predictions for secondary structure type i in protein p, while $N_i(p)$ is the number of residues of type i found in protein p. Both the numerator and denominator are summed over the total number of proteins, N_{prot}. Q2-EH is an additional per-residue accuracy measure that places emphasis on prediction of ordered secondary structure, and is calculated similary to eq 1, but for $i \in$ E, H. In addition to Q2-EH and Q3, accuracies were decomposed to analyze the percentage of the residues of observed for each secondary structure type that was predicted to be strand, helix, or coil. This means that we have nine additional metrics of the form {observed secondary structure}-{predicted secondary structure}. If the observed and predicted secondary structures are the same, the metric is the percent of correct predictions, while if the observed and predicted secondary structure differ the metric represents one of the types of misclassifications. For instance, C−C is the percent of residues in the coil conformation that are predicted to be coil, while E−C is the percent of strand residues predicted to be coil. C−C, E−E, and H−H are synonymous with QC, QE, and QH.

The segment overlap score[37] (SOV3) was also computed to quantify the level of accuracy in the prediction of continuous segments of secondary structure. SOV3 is computed as follows,

DOI: 10.1021/acs.jcim.5b00566
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Comparison of per Residue Accuracies of Recent and Historical Methods

method	C−C[c]	C−E[d]	C−H[d]	E−E[c]	E−C[d]	E−H[d]	H−H[c]	H−C[d]	H−E[d]	Q3[c,e]	Q2-EH[c]
conSSert-Q3[a]	0.819	0.086	0.096	0.774	0.205	0.022	0.869	0.117	0.015	**0.826**	0.831
	(0.003)	(0.003)	(0.002)	(0.005)	(0.005)	(0.001)	(0.004)	(0.003)	(0.001)	(0.002)	(0.002)
conSSert-Q2[b]	0.752	0.136	0.112	**0.838**	***0.140***	0.022	**0.883**	***0.095***	0.022	0.821	**0.865**
	(0.005)	(0.005)	(0.003)	(0.005)	(0.004)	(0.001)	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)
SCORPION	0.821	0.096	0.083	0.768	0.208	0.024	0.851	0.130	0.019	0.819	0.818
PSSpred	0.813	0.082	0.105	0.756	0.220	0.024	0.867	0.119	0.014	0.819	0.823
CONCORD	**0.871**	***0.068***	0.061	0.737	0.245	0.018	0.811	0.175	0.014	0.817	0.781
PSIPRED	0.864	0.072	0.063	0.729	0.254	***0.017***	0.803	0.185	0.012	0.809	0.774
SPINE-X	0.799	0.083	0.119	0.686	0.281	0.033	0.865	0.123	***0.011***	0.796	0.794
SSpro	0.805	0.071	0.124	0.646	0.237	0.117	0.820	0.144	0.036	0.772	0.751
RAPTOR	0.854	0.088	***0.057***	0.669	0.304	0.027	0.730	0.234	0.036	0.764	0.706
PROF	0.792	0.137	0.071	0.707	0.274	0.018	0.710	0.250	0.040	0.741	0.709
DSC	0.812	0.073	0.115	0.517	0.393	0.090	0.695	0.265	0.040	0.698	0.624
PROFphd	0.684	0.145	0.170	0.592	0.257	0.151	0.730	0.187	0.083	0.679	0.675
gorIV	0.716	0.130	0.153	0.456	0.344	0.200	0.599	0.281	0.120	0.611	0.542
predator	0.667	0.195	0.138	0.592	0.295	0.113	0.560	0.271	0.169	0.610	0.573

[a]Average and standard deviation (parentheses) of 5-fold cross-validation results from Table S3. [b]Average and standard deviation (parentheses) of 5-fold cross-validation results from Table S2. [c]Bold and underlined values indicate the best accuracy for a given column. [d]Bold and italicized values indicate the lowest error for a given column. [e]Rows are ordered according to Q3 accuracy.

$$\mathrm{SOV3} = \frac{1}{N}\sum_{i\in \mathrm{E,H,C}}\sum_{S(i)}\left[\frac{\min \mathrm{OV}(s_1, s_2) + \delta(s_1, s_2)}{\max \mathrm{OV}(s_1, s_2)}\mathrm{len}(s_1)\right] \quad (2)$$

where $\delta(s_1, s_2)$ is defined as

$$\delta(s_1, s_2) = \min\left\{\begin{array}{c}\max \mathrm{OV}(s_1, s_2) - \min \mathrm{OV}(s_1, s_2)\\ \min \mathrm{OV}(s_1, s_2)\\ \mathrm{int}(0.5 \times \mathrm{len}(s_1))\\ \mathrm{int}(0.5 \times \mathrm{len}(s_2))\end{array}\right\} \quad (3)$$

In eqs 2 and 3, s_1 and s_2 represent the observed and predicted overlapping segments of secondary structure type i, respectively, and $S(i)$ is the set of all pairs of segments (s_1, s_2) of type i with at least one residue in common. len represents the length of the secondary structure segments, more specifically the number of residues. min OV is the actual overlap between s_1 and s_2, and max OV is the maximum of the lengths of s_1 and s_2. SOV3 is normalized by the total number of residues, N. The other SOV measures are computed based on eqs 2 and 3, but the secondary structure types included in set i are varied. Unless specifically stated, all metrics are computed for the entire evaluation set, rather than representing averages of per protein accuracies.

2.4. Feature Selection. To evaluate the importance of each of the input secondary structure prediction methods, more specifically input features, we have applied a nonlinear SVM-based feature selection approach that has been recently developed by our group.[38] The approach is a recursive feature elimination algorithm, where at each iteration, the feature k with the largest magnitude of the criterion, crit_k, is eliminated from the feature basis. The criterion is based on the objective function of the dual SVM formulation with kernel, $K(x_i, x_j)$ and characterizes feature k's importance in a given feature basis as

$$\mathrm{crit}_k = -\frac{1}{2}\sum_i\sum_j \alpha_i^*\alpha_j^* y_i y_j \frac{\partial K(x_i \circ z, x_j \circ z)}{\partial z_k}\Bigg|_{z=1}$$

where α_i^*, α_j^* are the optimal values of the dual variables, y_i, y_j are the class labels (parameters taking on the values of -1, 1), and $z = 1$ indicates $z_k = 1$, $\forall\ k$. Here, the Hadamard product is used to associate each instance vector x_i with a selection vector z:

$$x_i \leftarrow x_i \circ z$$

Feature selection was performed for each of the three binary classifiers of conSSert-Q2 to identify the key features in the prediction of strand, helix, and coil. Data sets consisting of 10 000 samples, 5000 samples of the desired secondary structure type, and 2500 of each of the remaining types, were used in training the SVM models during feature selection. For each of the binary classifiers, feature selection was performed 100 times based on 100 different training sets and the final feature ranking was based on average of the rank for each feature across the 100 training sets.

3. RESULTS

Initially, we evaluated 12 established methods on our PDBselect25 data set of 4261 proteins to identify the top performing methods, and Tables 1 and 2 summarize the results. More specifically we sought to identify the methods that most accurately predict each of the secondary structure types (strand, helix, and coil) in order to select methods to serve as inputs into our consensus SVM models. Only methods that provide individual scores/probabilities for each secondary structure type were considered as potential inputs for our consensus models. Based on these criteria we selected PSSpred, PSIPRED, SPINE-X, and RAPTOR as inputs into our consensus SVM models, as these four methods represent the two most accurate methods for the prediction of each of secondary structure types.

On the basis of the selected input methods, we developed two consensus SVM models tuned to improve either Q2-EH or Q3 per residue accuracies, named conSSert-Q2 and conSSert-Q3 respectively. conSSert-Q2 is based on binary one-vs-all classifiers trained on balanced sets, while conSSert-Q3 is based one-vs-one classifiers that were trained using the natural occurrence of coil, helix, and strand. The methods were evaluated through 5-fold cross validation, where models were

DOI: 10.1021/acs.jcim.5b00566
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Comparison of SOV Accuracies of Recent and Historical Methods

method	SOV-C[c]	SOV-E[c]	SOV-H[c]	SOV3[c]	SOV-EH[c]
conSSert-Q3[b]	**0.740**	0.774	0.851	0.787	0.817
	(0.004)	(0.004)	(0.004)	(0.002)	(0.001)
conSSert-Q2[a]	0.717	**0.811**	**0.855**	**0.789**	**0.835**
	(0.006)	(0.004)	(0.006)	(0.003)	(0.002)
SCORPION	**0.740**	0.760	0.849	0.783	0.810
PSSpred	0.710	0.756	0.845	0.767	0.805
CONCORD	0.718	0.748	0.822	0.761	0.789
PSIPRED	0.707	0.738	0.817	0.752	0.782
SPINE-X	0.683	0.647	0.831	0.721	0.745
RAPTOR	0.692	0.678	0.747	0.707	0.717
PROF	0.665	0.734	0.716	0.701	0.724
SSpro	0.689	0.644	0.754	0.700	0.707
DSC	0.635	0.558	0.694	0.635	0.635
PROFphd	0.602	0.599	0.695	0.633	0.654
predator	0.573	0.599	0.578	0.582	0.587
gorIV	0.570	0.492	0.600	0.560	0.553

[a]Average and standard deviation (parentheses) of 5-fold cross-validation results from Table S4. [b]Average and standard deviation (parentheses) of 5-fold cross-validation results from Table S5. [c]Bold and underlined values indicate best accuracy for a given column.

trained on four-fifths of the data and tested on the remaining one-fifth. The mean cross validation results with respect to per-residue accuracies are summarized in Table 1, and full per-residue cross validation accuracies are provided in Tables S2 and S3.

Comparison of conSSert-Q2 and conSSert-Q3 per-position accuracies shows that conSSert-Q2 provides ∼6% improvement in strand accuracy (Table 1, E−E) and over 3% improvement in Q2-EH. However, this significant improvement in strand accuracy comes with a trade-off, a slight reduction in Q3 and ∼6% reduction in coil accuracy (Table 1, C−C). On the basis of either the conSSert-Q2 or conSSert-Q3 model, strands and helices are rarely misclassified as one another, and the majority of errors involve coil residues being predicted to be helix/strand or helix/strand residues being misclassified as coil. The order of accuracy for conSSert-Q2 is helix is more accurate than strand, and strand is more accurate than coil, while for conSSert-Q3 helix prediction is more accurate than coil, and coil is more accurate than strand. SCORPION and PSSpred, the most recent evaluated methods, were the most accurate of the competing methods with respect to Q3. However, conSSert-Q3 provides ∼0.5% improvement with regard to Q3. PSSpred performed best among the competing methods with regard to Q2-EH at 0.823, but conSSert-Q2 provides an over 4% improvement in Q2-EH. conSSert-Q2's most significant improvement in per-residue accuracy, is an over 7% increase in strand accuracy (Table 1, E−E), while a ∼1.5% increase in helix accuracy still contributes to overall improvements in Q2-EH.

Per-residue accuracies do not always reflect the utility of secondary structure prediction methods, especially in the context of protein threading and protein structure prediction, where accurate identification of continuous segments of secondary structure is highly desired. Therefore, we have also evaluated our conSSert cross validation results with respect to the segment overlap measure (SOV), with mean values across the 5-fold cross validation provided in Table 2, and the full cross validation SOV accuracies provided in Tables S4 and S5. Interestingly, conSSert-Q2 outperforms conSSert-Q3 with regard to all SOV measures except SOV-C. By improving accuracy of helices and strands over coil, conSSert-Q2 actually provides more accurate three-class SOV accuracy (SOV3) than conSSert-Q3, which provides higher three-class per-residue accuracy. Of the competing methods, SCORPION performed best with regard to all SOV measures (Table 2). However, overall, conSSert-Q2 outperforms all of the evaluated methods, providing a 4% improvement in strand overlap score (Table 2, SOV-E) and a 2% improvement in SOV-EH.

To further test the robustness of the conSSert consensus models, we performed an additional round of 5-fold cross validation using a larger data set composed of 8632 proteins from the PISCES protein culling server. In addition to conSSert-Q2, conSSert-Q3, and the four input methods, we also evaluated the accuracy of SCORPION, which was one of the most accurate single methods for our PDBSelect25 data set, as an additional control. The per-residue accuracies for the PISCES data set (Table 3) are very similar to those obtained for the PDBselect25 data set (Table 1). The lack of major change in the accuracies of conSSert when performing cross-validation with the PISCES data set, which is twice as large as the PDBSelect25 data set, confirms the robustness of the conSSert models and suggests that the use of an even larger training set would not improve prediction accuracy.

It is likely that competing methods, including the four methods used as input into conSSert, were trained based on many targets of our data set, implying a potential bias in our comparisons. We have also tested the conSSert-Q2 model, which was trained on the PISCES data set, on targets from

Table 3. Comparison of per Residue Accuracies for the 8632 Proteins of the PISCES Dataset

method	C−C[c]	C−E[d]	C−H[d]	E−E[c]	E−C[d]	E−H[d]	H−H[c]	H−C[d]	H−E[d]	Q3[c]	Q2-EH[c]
conSSert-Q3[a]	0.814	0.081	0.105	0.755	0.218	0.027	0.871	0.116	0.014	**0.823**	0.829
	(0.003)	(0.002)	(0.002)	(0.006)	(0.005)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.003)
conSSert-Q2[b]	0.750	0.135	0.115	**0.828**	***0.148***	0.024	**0.877**	***0.098***	0.024	0.818	**0.860**
	(0.003)	(0.002)	(0.001)	(0.005)	(0.004)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
SCORPION	0.816	0.093	0.091	0.755	0.216	0.029	0.855	0.126	0.019	0.818	0.819
PSSpred	0.806	0.077	0.117	0.739	0.231	0.029	0.869	0.116	0.014	0.816	0.822
PSIPRED	**0.866**	***0.067***	0.067	0.695	0.284	***0.020***	0.799	0.188	***0.012***	0.802	0.762
SPINE-X	0.791	0.078	0.131	0.676	0.286	0.039	0.870	0.118	***0.012***	0.796	0.800
RAPTOR	0.849	0.089	***0.062***	0.662	0.310	0.028	0.727	0.229	0.043	0.759	0.704

[a]Average and standard deviation (parentheses) of 5-fold cross-validation results from Table S8. [b]Average and standard deviation (parentheses) of 5-fold cross-validation results from Table S7. [c]Bold and underlined values indicate best accuracy for a given column. [d]Bold and italicized values indicate lowest error for a given column.

DOI: 10.1021/acs.jcim.5b00566
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

CASP9, CASP10, and CASP11. Table 4 summarizes the overall per-residue accuracies for the three CASP test sets. With regard to Q2-EH, conSSert-Q2 outperforms all input methods on all three CASP data sets, with improvements of ∼4% for every CASP set. However, conSSert-Q2 only provides improvements in Q3 for the CASP9 and CASP11 sets, since the Q3 of conSSert-Q2 on the CASP10 set was slightly worse than PSSpred (Table 4). The improved accuracy for ordered secondary structure of conSSert-Q2 is further illustrated by Figure 1, which shows the distributions of per-target accuracies for Q2-EH, QE, and QH. Figure 1C shows there is clear variability in the helix accuracy of the four input methods, however, all four methods perform comparably in predicting strands. Despite consistently poor prediction of strands by the four input methods, conSSert-Q2 provides substantial improvement in the per target strand accuracies (Figure 1B).

Table 4. Summary of Overall per Residue Accuracies for CASP9, CASP10, and CASP11

	CASP9		CASP10		CASP11	
method	Q3[a]	Q2-EH[a]	Q3[a]	Q2-EH[a]	Q3[a]	Q2-EH[a]
conSSert-Q2	**0.811**	**0.854**	0.818	**0.863**	**0.805**	**0.853**
PSSpred	0.810	0.818	**0.820**	0.823	0.802	0.811
PSIPRED	0.799	0.761	0.809	0.765	0.796	0.766
RAPTOR	0.754	0.699	0.774	0.725	0.761	0.715
SPINEX	0.783	0.786	0.801	0.803	0.787	0.794

[a]Bold and underlined values indicate best accuracy for a given column.

4. DISCUSSION

Based on the accuracies of the recent and historical methods (Tables 1 and 2), it is obvious that significant progress has been made in the area of protein secondary structure prediction. That said, prediction of ordered secondary structure, especially strands, has somewhat lagged behind. Multiple competing methods (e.g., PSIPRED, CONCORD, RAPTOR) have per residue per type accuracies (Table 1) that are ordered by the natural occurrence of each type (i.e., coil is more accurate than helix, helix more accurate than strand). Some recent methods, such as PSSpred and SCORPION, predict helix more accurately than coil, but to our knowledge conSSert-Q2 is the first method to predict helix better than strand, and strand better than coil while providing an accurate prediction. Several of the evaluated methods, including both conSSert models, have SOV per-type accuracies such that helix is predicted best and coil predicted worst. However, some methods, such as SPINE-X and RAPTOR, predict strand least accurately when compared to their accuracies for helix and coil.

By comparing conSSert-Q2 and conSSert-Q3, the trade-offs in per-type accuracies between the two training schemes are apparent. In comparison to PSSpred, the input method with the highest overall Q3 accuracy (Table 1), conSSert-Q3 provides slight improvement in per-protein QC and QE, while performing equivalently with regard to QH. conSSert-Q2, provides improvement in QE and in QH when comparing to PSSpred, the most accurate input method with regard to Q2-EH, while providing the least accurate predictions for coil. However, despite these trade-offs in regard to coil prediction, conSSert-Q2 still out performs all input methods with regard to Q3 (Table 1). Additionally, PSSpred appears to make similar trade-offs in accuracy, since PSSpred is most accurate among the input methods with regard to Q2-EH, and second to last in regard to prediction of coil.

The individual accuracies of the one-vs-all binary classifiers that compose conSSert-Q2 (Figure 2) provide further insight into interplay between per type accuracies. The helix binary classifier outperforms strand and coil for most probability values, suggesting it is easier to distinguish between residues that are helix or not-helix. The strand binary classifier achieves higher precision/recall values at higher probability values, suggesting that the strand classifier is more confident in its predictions than the coil classifier, or in other words, the binary classifiers are able to separate strand/not-strand further than coil/not-coil. However, strand and coil provide similar precision/recall for probability values of ∼0.8 and less, illustrating that both classifiers have similar accuracies when approaching the margin. Since binary classifiers of strand and coil perform similarly, it is clear why the training scheme can be used to emphasize Q3 or Q2-EH or more specifically coil or strand.

To evaluate the contribution of the four input methods utilized by conSSert-Q2 we performed feature selection using a nonlinear SVM-based algorithm developed by our group. Table 5 summarizes the feature ranking for each of the binary classifiers of conSSert-Q2 based on 100 runs of feature selection, where the features are sorted according to the average rank. For all three classifiers the top ranked feature is a predicted probability for the corresponding class, however, predicted probabilities for a secondary structure type other than

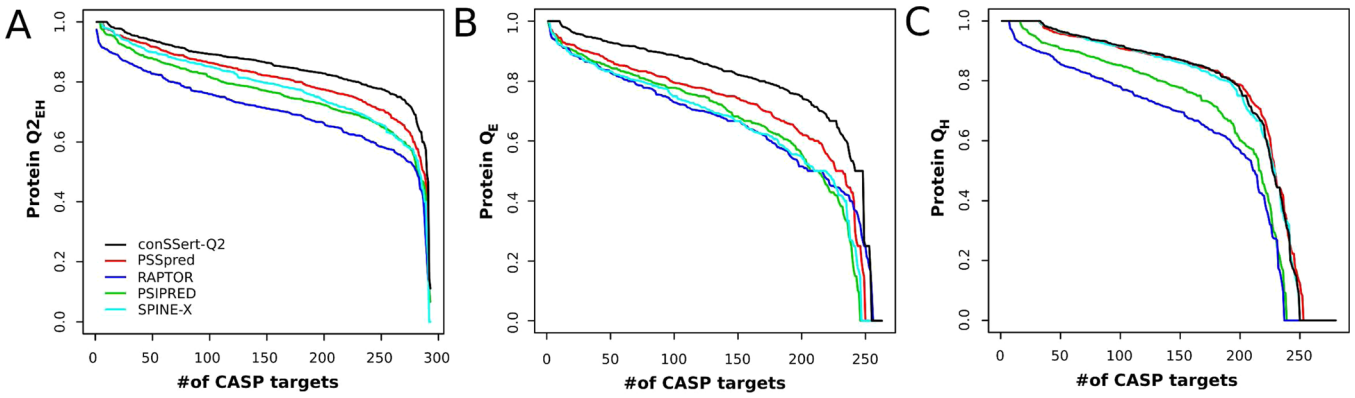


Figure 1. Comparison of per protein accuracy distributions for 293 targets from CASP9, CASP10, and CASP11. Sorted per protein accuracies for prediction of (A) Q2-EH, (B) QE, and (C) QH. The legend in panel A is for all three panels.

DOI: 10.1021/acs.jcim.5b00566
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

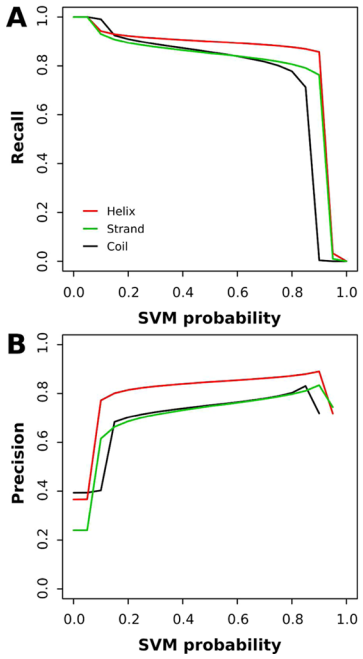


Figure 2. Comparison of recall and precision of conSSert-Q2 binary SVM models. (A) Recall of conSSert-Q2 SVM models as a function of the predicted SVM probabilities: coil in black, strand in green, helix in red. (B) Precision of conSSert-Q2 SVM models as a function of the predicted SVM probabilities: coil in black, strand in green, helix in red. The legend in panel A is for both panels.

the model type are often highly ranked, which is rather unexpected. For example, the third best feature for prediction of coil is the strand probability fom PSSpred and the second best feature for strand prediction is the helix probability from PSSpred. Since the feature selection ranks individual features, rather than methods, a cumulative score was calculated based on the sum of the average ranks of the best feature for each secondary structure type from each method. The resulting scores and ranks are PSSpred 8.4, PSIPRED 8.7, SPINE-X 13.9, and RAPTOR 21.4, where the best possible score is 3 and the worse possible score is 30. This ranking gives some insight into the overall contribution of the methods, but hides some details, such as the fact that the probability for helix from SPINE-X was the second best feature for helix prediction. One thing that is apparent when analyzing the AUC values is that not all features are required to achieve the maximum accuracy, especially for coil.

To further quantify the contributions of the features we performed cross-validation for each set of features, producing models with between 1 and 12 features for each secondary structure type, and calculated area under the receiver operator curve. For example, the strand model based on the top two features, PSI-E and PSS-H, has an AUC of 0.930. Additionally, it is clear that the models for predicting strand and helix are more accurate than the model for predicting coil, regardless of the number of features included. This further suggests that some methods have over emphasized the prediction of coil to improve the Q3 score, rather than methods predicting coil most accurately due to coil being the easiest class to predict. The training scheme of conSSert-Q2 corrects this bias and as a result conSSert-Q2 provides highly accurate prediction of ordered secondary structure.

In this study, we developed two consensus SVM models, conSSert-Q2 and conSSert-Q3, for highly accurate prediction protein secondary structure. We illustrate that simply based on the design of the training scheme it is possible to tune the preference for prediction of helix and strand, which are observed in nature less frequently than coil, but more crucial to ordered protein structures. The significant improvement in strand accuracy is in great need, especially in the area of beta-sheet topology prediction.[2,39−41] conSSert-Q2, which out performs all evaluated methods in regard to Q2-EH and SOV3, is available to the public as a webservice at http://ares.tamu.edu/conSSert/.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00566.

The Supporting Information file includes tables describing the compositions of training and testing sets, as well as the full cross validation results from which means and standard deviations were derived. (PDF)

Table 5. Summary of Feature Ranking for conSSert-Q2 Models

	coil model			strand model			helix model		
no. of features	feature[a]	average rank[b]	AUC[c]	feature[a]	average rank[b]	AUC[c]	feature[a]	average rank[b]	AUC[c]
1	PSI-C	1.2	0.853	PSI-E	3.2	0.917	PSS-H	2.0	0.912
2	PSS-C	1.8	0.866	PSS-H	3.4	0.930	SPX-H	3.7	0.916
3	PSS-E	3.0	0.871	PSS-E	3.5	0.930	PSI-E	4.3	0.926
4	SPX-C	4.0	0.873	PSS-C	5.3	0.933	PSI-H	4.5	0.923
5	RAP-C	5.0	0.874	PSI-C	6.1	0.934	PSS-C	4.5	0.925
6	PSI-E	6.2	0.877	SPX-H	6.2	0.936	PSS-E	5.2	0.926
7	PSI-H	6.9	0.880	SPX-C	6.5	0.940	SPX-C	6.6	0.928
8	RAP-E	7.8	0.882	PSI-H	6.6	0.939	PSI-C	7.7	0.928
9	SPX-E	9.0	0.883	SPX-E	7.0	0.940	RAP-E	8.2	0.932
10	PSS-H	10.0	0.883	RAP-E	8.1	0.940	SPX-E	10.0	0.934
11	SPX-H	11.0	0.883	RAP-H	10.7	0.941	RAP-H	10.4	0.935
12	RAP-H	12.0	0.883	RAP-C	11.4	0.941	RAP-C	11.0	0.934

[a]Features are named as {method}-{structure type}, where PSI = PSIPRED, PSS = PSSpred, SPX = SPINE-X, RAP = RAPTOR, C = coil, E = strand, and H = helix. [b]Average rank is based on 100 runs of feature selection. [c]Area under the ROC curve based on 10 runs of 10-fold cross-validation using 10 000 samples total.

DOI: 10.1021/acs.jcim.5b00566
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Author

*E-mail: floudas@tamu.edu.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Floudas, C. A. Computational Methods in Protein Structure Prediction. *Biotechnol. Bioeng.* **2007**, *97* (2), 207−213.

(2) Khoury, G. A.; Smadbeck, J.; Kieslich, C. A.; Floudas, C. A. Protein Folding and De Novo Protein Design for Biotechnological Applications. *Trends Biotechnol.* **2014**, *32* (2), 99−109.

(3) Rost, B.; Sander, C. Third Generation Prediction of Secondary Structures. *Methods Mol. Biol.* **2000**, *143*, 71−95.

(4) Zhou, Y.; Faraggi, E. *Prediction of One-Dimensional Structural Properties of Proteins by Integrated Neural Networks*; John Wiley & Sons, Inc., 2010; pp 45−74.

(5) Rost, B. Review: Protein Secondary Structure Prediction Continues to Rise. *J. Struct. Biol.* **2001**, *134* (2−3), 204−218.

(6) Zhang, W.; Dunker, A. K.; Zhou, Y. Assessing Secondary Structure Assignment of Protein Structures by Using Pairwise Sequence-Alignment Benchmarks. *Proteins: Struct., Funct., Genet.* **2008**, *71* (1), 61−67.

(7) Yaseen, A.; Li, Y. Context-Based Features Enhance Protein Secondary Structure Prediction Accuracy. *J. Chem. Inf. Model.* **2014**, *54* (3), 992−1002.

(8) Bettella, F.; Rasinski, D.; Knapp, E. W. Protein Secondary Structure Prediction with SPARROW. *J. Chem. Inf. Model.* **2012**, *52* (2), 545−556.

(9) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* **2015**, *12* (1), 7−8.

(10) Buchan, D. W. A.; Minneci, F.; Nugent, T. C. O.; Bryson, K.; Jones, D. T. Scalable Web Services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **2013**, *41* (W1), W349−W357.

(11) Faraggi, E.; Zhang, T.; Yang, Y.; Kurgan, L.; Zhou, Y. SPINE X: Improving Protein Secondary Structure Prediction by Multistep Learning Coupled with Prediction of Solvent Accessible Surface Area and Backbone Torsion Angles. *J. Comput. Chem.* **2012**, *33* (3), 259−267.

(12) Wang, Z.; Zhao, F.; Peng, J.; Xu, J. Protein 8-Class Secondary Structure Prediction Using Conditional Neural Fields. *Proteomics* **2011**, *11* (19), 3786−3792.

(13) Wei, Y.; Thompson, J.; Floudas, C. A. CONCORD: a Consensus Method for Protein Secondary Structure Prediction via Mixed Integer Linear Optimization. *Proc. R. Soc. London, Ser. A* **2012**, *468* (2139), 831−850.

(14) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J. H.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389−3402.

(15) Jones, D. T. Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. *J. Mol. Biol.* **1999**, *292* (2), 195−202.

(16) Rost, B.; Yachdav, G.; Liu, J. F. The PredictProtein Server. *Nucleic Acids Res.* **2004**, *32* (Web Server issue), W321−W326.

(17) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. *A Training Algorithm for Optimal Margin Classifiers*; ACM, 1992.

(18) Ward, J. J.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. Secondary Structure Prediction with Support Vector Machines. *Bioinformatics* **2003**, *19* (13), 1650−1655.

(19) Hua, S. J.; Sun, Z. R. A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach. *J. Mol. Biol.* **2001**, *308* (2), 397−407.

(20) Guermeur, Y.; Pollastri, G.; Elisseeff, A.; Zelus, D.; Paugam-Moisy, H.; Baldi, P. Combining Protein Secondary Structure Prediction Models with Ensemble Methods of Optimal Complexity. *Neurocomputing* **2004**, *56*, 305−327.

(21) Kountouris, P.; Hirst, J. D. Prediction of Backbone Dihedral Angles and Protein Secondary Structure Using Support Vector Machines. *BMC Bioinf.* **2009**, *10*, 437.

(22) Guermeur, Y.; Geourjon, C.; Gallinari, P.; Deleage, G. Improved Performance in Protein Secondary Structure Prediction by Inhomogeneous Score Combination. *Bioinformatics* **1999**, *15* (5), 413−421.

(23) Gupta, A.; Deshpande, A.; Amburi, J. K.; Sabarinathan, R.; Senthilkumar, R.; Sekar, K. CSSP (Consensus Secondary Structure Prediction): a Web-Based Server for Structural Biologists. *J. Appl. Crystallogr.* **2009**, *42*, 336−338.

(24) Cuff, J. A.; Clamp, M. E.; Siddiqui, A. S.; Finlay, M.; Barton, G. J. JPred: a Consensus Secondary Structure Prediction Server. *Bioinformatics* **1998**, *14* (10), 892−893.

(25) Cheng, H.; Sen, T. Z.; Jernigan, R. L.; Kloczkowski, A. Consensus Data Mining (CDM) Protein Secondary Structure Prediction Server: Combining GOR v and Fragment Database Mining (FDM). *Bioinformatics* **2007**, *23* (19), 2628−2630.

(26) Pollastri, G.; Przybylski, D.; Rost, B.; Baldi, P. Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. *Proteins: Struct., Funct., Genet.* **2002**, *47* (2), 228−235.

(27) King, R. D.; Sternberg, M. Identification and Application of the Concepts Important for Accurate and Reliable Protein Secondary Structure Prediction. *Protein Sci.* **1996**, *5* (11), 2298−2310.

(28) Ouali, M.; King, R. D. Cascaded Multiple Classifiers for Secondary Structure Prediction. *Protein Sci.* **2000**, *9* (6), 1162−1176.

(29) Frishman, D.; Argos, P. Incorporation of Non-Local Interactions in Protein Secondary Structure Prediction From the Amino Acid Sequence. *Protein Eng., Des. Sel.* **1996**, *9* (2), 133−142.

(30) Garnier, J.; Gibrat, J. F.; Robson, B. GOR Method for Predicting Protein Secondary Structure From Amino Acid Sequence. *Methods Enzymol.* **1996**, *266*, 540−553.

(31) Griep, S.; Hobohm, U. PDBselect 1992−2009 and PDBfilter-Select. *Nucleic Acids Res.* **2010**, *38* (Database issue), D318−D319.

(32) Wang, G. L.; Dunbrack, R. L. PISCES: a Protein Sequence Culling Server. *Bioinformatics* **2003**, *19* (12), 1589−1591.

(33) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577−2637.

(34) R Core Team. *R: a Language and Environment for Statistical Computing*; Vienna, Austria, 2014.

(35) Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F. *E1071: Misc Functions of the Department of Statistics*; TU Wien; 2014.

(36) Chang, C.-C.; Lin, C.-J. LIBSVM: a Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (3), 1.

(37) Zemla, A.; Venclovas, C.; Fidelis, K.; Rost, B. A Modified Definition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment. *Proteins: Struct., Funct., Genet.* **1999**, *34* (2), 220−223.

(38) Kieslich, C. A.; Tamamis, P.; Guzman, Y. A.; Onel, M.; Floudas, C. A. Highly Accurate Structure-Based Prediction or HIV-1 Coreceptor Usage Suggests Intermolecular Interactions Driving Tropism. *PLoS One* **2016**, *11*, e0148974.

(39) Subramani, A.; Floudas, C. A. Beta-Sheet Topology Prediction with High Precision and Recall for Beta and Mixed Alpha/Beta Proteins. *PLoS One* **2012**, *7* (3), e32461.

(40) Klepeis, J. L.; Floudas, C. A. Prediction of Beta-Sheet Topology and Disulfide Bridges in Polypeptides. *J. Comput. Chem.* **2003**, *24* (2), 191−208.

(41) Cheng, J. L.; Baldi, P. Three-Stage Prediction of Protein Beta-Sheets by Neural Networks, Alignments and Graph Algorithms. *Bioinformatics* **2005**, *21*, I75−I84.

DOI: 10.1021/acs.jcim.5b00566
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX