LOCUSTRA: Accurate Prediction of Local Protein Structure Using a Two-Layer Support Vector Machine Approach

Olav Zimmermann*,† and Ulrich H. E. Hansmann†,‡

John von Neumann Institut für Computing, Research Centre Jülich, 52425 Jülich, Germany, and Department of Physics, Michigan Technological University, Houghton, Michigan 49931

Received May 25, 2008

Constraint generation for 3d structure prediction and structure-based database searches benefit from fine-grained prediction of local structure. In this work, we present LOCUSTRA, a novel scheme for the multiclass prediction of local structure that uses two layers of support vector machines (SVM). Using a 16-letter structural alphabet from de Brevern et al. (*Proteins: Struct., Funct., Bioinf.* **2000**, *41*, 271−287), we assess its prediction ability for an independent test set of 222 proteins and compare our method to three-class secondary structure prediction and direct prediction of dihedral angles. The prediction accuracy is $Q_{16} = 61.0\%$ for the 16 classes of the structural alphabet and $Q_3 = 79.2\%$ for a simple mapping to the three secondary classes helix, sheet, and coil. We achieve a mean $\phi(\psi)$ error of 24.74°(38.35°) and a median RMSDA (root-mean-square deviation of the (dihedral) angles) per protein chain of 52.1°. These results compare favorably with related approaches. The LOCUSTRA web server is freely available to researchers at http://www.fz-juelich.de/nic/cbb/service/service.php.

1. INTRODUCTION

Although the surging number of available protein structures in the Protein Data Bank (PDB)[2] has increased the probability that one can predict the structure of a protein via homology modeling, for the majority of sequences this remains impossible. More distantly related templates can be identified using fold recognition techniques where local structure predictions provide a means to reduce the number of possible folds. Structure constraints have recently been reported to accelerate the search for low-energy conformations of proteins by Monte Carlo simulations.[3] In many cases the constraints originate from three-class secondary structure predictions, for example, from the program PSIPRED,[4] but algorithms have been developed also for the prediction of other frequent structural motifs.[5,6] For instance, several groups have clustered local protein structures and calculated their cluster centers.[1,7] These sets of local structure representatives, for which the term "structural alphabets" has been coined, are capable of approximating a given protein structure. Strategies have been developed to predict them from the amino acid sequence using recurrent neural networks,[8] hidden Markov models,[9] logistic regression,[10] decision trees, random forests, or support vector machines (SVM).[7] In this work, we introduce LOCUSTRA, a method based on SVMs to predict the local structure of a protein chain from its sequence. Predicting the letters of a structural alphabet has the benefit to provide narrow structure constraints for the entire protein chain and to explicitly take into account local structure correlations.

Using a mapping to dihedral angles, we compare the quality of the 16-class LOCUSTRA predictions to three-class secondary structure predictions and methods for the prediction of dihedral angles. The discussion focuses on performance determinants and applications.

* Corresponding author. Phone: +49-2461-61-1520. Fax: +49-2461-61-2430. E-mail: olav.zimmermann@fz-juelich.de.
† Research Centre Jülich.
‡ Michigan Technological University.

2. METHODS

2.1. Data Set. We use nonredundant subsets of the Protein Data Bank (PDB).[2] For our sequence and structure data sets, we use the PDB25 set (pairwise sequence identity ≤25%) and select X-ray structures with resolution better than 2.0 Å. Our set contains 1112 protein chains with 193 208 residues. 890 chains are used for training and cross validation while a holdout of 222 chains serves as an independent test set.

2.2. Structural Alphabet. We use the structural alphabet from de Brevern and co-workers.[1] It contains 16 structure representatives, *a*−*p*, each five residues in length. This set has been reported to approximate protein structures reasonably well, having more than 50% of the dihedral angles within 21° and 97% within 90°.[1] Each structural alphabet letter has been defined by eight backbone dihedral angles (Table 1). The most frequent letters, *d* and *m*, correspond to the repetitive dihedral patterns found in the center of β-sheets and α-helices, respectively. The structural alphabet letters *a*, *b*, *c* are often found at the N-termini of β-sheets, *e* and *f* at their C-termini. In α-helices, N- and C-termini are characterized by the letters *k*, *l* and *n*, *o*, *p*, respectively, while *g*, *h*, *i*, *j* are predominantly found in coil regions.

In order to measure the conformational distance and plasticity, we calculate the root-mean-square deviation over the dihedral angles (RMSDA) of a fragment pair.[1] For comparison of two fragments *x*, *y* each having *L* residues and $2(L - 1)$ interior backbone dihedral angles, the RMSDA is calculated by

10.1021/ci800178a CCC: $40.75
Published on Web 09/03/2008

$$\mathrm{RMSDA} = \left[\frac{\sum_{f=1}^{L-1} ((\psi_{f,x} - \psi_{f,y})^2 + (\phi_{f+1,x} - \phi_{f+1,y})^2)}{2(L-1)} \right]^{1/2} \quad (1)$$

For each of the overlapping 5-residue fragments, we assign the structural alphabet letter with the smallest RMSDA to the central residue. The last two columns of Table 1 show the distribution of the letters in the training and test sets.

2.3. Support Vector Machines. Our prediction algorithm employs support vector machines (SVM), a supervised machine learning algorithm; that is, it requires positive and negative examples for training. For a comprehensive introduction to SVMs see ref 11. Throughout this study we use the C-SVM and ν-SVM algorithm implementations from the LIBSVM software[12] with a radial-basis-function kernel. Input data for SVM training are vectors composed of a class label and several numerical input values called features. The resulting model is a nonlinear decision hyperplane that maximizes the distance to the closest datapoints in the high-dimensional feature space. This model is then used to classify feature vectors derived from new, unknown data.

2.4. Prediction Schemes. Structural alphabets have many classes while SVMs provide binary classifiers. In order to map multiclass problems onto binary classification, one can use one-per-class: the positive class contains samples of one structural alphabet letter while the negative class contains samples from all other letters. For k classes this scheme requires k classifiers; that is, 16 one-per-class classifiers in the case of the de Brevern structural alphabet. The alternative is pairwise coupling: The positive class contains samples of one structural alphabet letter while the negative class contains samples from one other letter. The number of classifiers is $k(k-1)/2$; that is, for the de Brevern alphabet 120 pairwise coupling classifiers are required. In this work, we suggest a combined two-layer scheme (LOCUSTRA) where the first layer is represented by the pairwise-coupling classifiers described above, and the second layer consists of one-per-class classifiers that use the prediction output of the pairwise-coupling classifiers as input. The LOCUSTRA scheme also employs heuristics for disambiguation and derivation of a consensus prediction for the dihedral angles of each residue.

2.5. Encoding. In order to allow the algorithms to harness information from homologous and analogous proteins, we obtain a profile of the amino acid propensities for each sequence position using the software PSI-BLAST.[13] To encode each amino acid residue together with its local neighborhood, we use a sliding window of length 15 over the position specific scoring matrix calculated from the PSI-BLAST alignment. With 20 different amino acids, and one additional column per position indicating whether it is beyond the chain, we obtain feature vectors of length $(20 + 1) \times 15 = 315$ for each residue. We use the cross-validation-based method of LIBSVM to estimate class probabilities from the decision values.[14]

The one-per-class classifiers for the second layer take those probabilities obtained from the 120 pairwise-coupling classifier predictions per sequence position as features. In order to keep the size of the input vectors manageable, we use a reduced sliding window of 7 residues. Including a chain extension indicator column, as described above, each input vectors of this classifier contains $(120 + 1) \times 7 = 847$ features.

2.6. Training. Due to the large differences in class size (cf. Table 1) we prepare training sets for the classifiers based on position specific scoring matrices that ensure the same number of positive and negative examples. These sets thus contain between 2000 and 80 000 labeled vectors. The classifier set was trained using ν-SVM with standard parameters ($\nu = 0.5$). The second layer SVM was trained with seven disjunct training sets of each 20 000 vectors per structural alphabet letter. Here we used C-SVM and performed a coarse grid-based optimization of the regularization parameter C, the hyperparameter γ of the radial-basis-function kernel, and also the regularization weights w for the positive and negative class because we kept the natural class bias.

2.7. Assessment. Prediction of the test samples follows the same scheme as the training. For the LOCUSTRA scheme we apply the following heuristics after the two SVM classification steps to derive a consensus prediction per residue:

1. The structural alphabet letter with the highest number of votes from the one-per-class classifiers of the second layer is selected as prediction.

2. If there is more than one letter with the highest number of votes, we select d = central beta sheet if it is among the alternatives.

3. Else we select m = central helix.

Table 1. Structural Alphabet from de Brevern and Co-Workers:[1] Dihedral Angles and Occurrence

letter	ψ_{i-2}	ϕ_{i-1}	ψ_{i-1}	ϕ_i	ψ_i	ϕ_{i+1}	ψ_{i+1}	ϕ_{i+2}	train	test
a	41.14	75.53	13.92	−99.80	131.88	−96.27	122.08	−99.68	5676	1383
b	108.24	−90.12	119.54	−92.21	−18.06	−128.93	147.04	−99.90	6616	1630
c	−11.61	−105.66	94.81	−106.09	133.56	−106.93	135.97	−100.63	12371	3008
d	141.98	−112.79	132.20	−114.79	140.11	−111.05	139.54	−103.16	29362	7411
e	133.25	−112.37	137.64	−108.13	133.00	−87.30	120.54	77.40	3428	866
f	116.40	−105.53	129.32	−96.68	140.72	−74.19	−26.65	−94.51	9735	2354
g	0.40	−81.83	4.91	−100.59	85.50	−71.65	130.78	84.98	1663	395
h	119.14	−102.58	130.83	−67.91	121.55	76.25	−2.95	−90.88	3195	816
i	130.68	−56.92	119.26	77.85	10.42	−99.43	141.40	−98.01	2425	616
j	114.32	−121.47	118.14	82.88	−150.05	−83.81	23.35	−85.82	1144	249
k	117.16	−95.41	140.40	−59.35	−29.23	−72.39	−25.08	−76.16	7900	1938
l	139.20	−55.96	−32.70	−68.51	−26.09	−74.44	−22.60	−71.74	7125	1701
m	−39.62	−64.73	−39.52	−65.54	−38.88	−66.89	−37.76	−70.19	41099	10294
n	−35.34	−65.03	−38.12	−66.34	−29.51	−89.10	−2.91	77.90	2507	601
o	−45.29	−67.44	−27.72	−87.27	5.13	77.49	30.71	−93.23	3733	854
p	−27.09	−86.14	0.30	59.85	21.51	−96.30	132.67	−92.91	4884	1199

Table 2. Definition of Prediction Categories for Calculation of the Matthews Correlation Coefficient (MCC) and the Accuracy

		observation	
		+1	−1
prediction	+1	*tp*	*fp*
	−1	*fn*	*tn*

We measure the quality of our predictions using Matthews correlation coefficient (MCC) instead of accuracy as the MCC corrects for class bias.[15]

$$\mathrm{MCC} = \frac{tp \cdot tn - fp \cdot fn}{[(tp + fp) \cdot (tp + fn) \cdot (tn + fp) \cdot (tn + fn)]^{1/2}} \quad (2)$$

using the definitions in Table 2. However, accuracy will be given for the main results to allow for comparison to published data. The accuracy Q_K for a K-class prediction is defined as

$$Q_K = \frac{tp + tn}{tp + fp + tn + fn} \quad (3)$$

For the discussion we also use the average confusion rate Cf between two structural alphabet letters x and y:

$$Cf = \frac{1}{2}(P(pr = y|ob = x) + P(pr = x|ob = y)) \quad (4)$$

The structural distance of predictions are measured using the RMSDA described in eq 1. To compare the 16-class predictions of LOCUSTRA and the three-class predictions of PSIPRED, we map each structural alphabet letter to a dihedral angle pair. For PSIPRED we take the average value for ϕ and ψ of those residues in the training set predicted as helix, sheet, and coil, respectively. For LOCUSTRA we use the ϕ and ψ value of the central residue of the respective structural alphabet letter (cf. Table 1). As the fragments corresponding to the letters predicted by LOCUSTRA are overlapping and we have an ensemble of seven classifiers for each class, we collect the ϕ and ψ values corresponding to each vote for a letter and assign the median of the ϕ and ψ lists to the residue.

3. RESULTS

3.1. Performance of the LOCUSTRA Prediction Scheme. The first layer pairwise-coupling classifiers show a wide spread in classification performance. The lowest MCC in the test set is obtained for pairs with very high class bias ((*d*, *g*),(*d*, *j*),(*m*, *g*),(*m*, *j*) in Table 3). In order to avoid the parametrization of class weightings, the pairwise-coupling classifier training sets have equal numbers of positive and negative examples. The minimal separation power for the bias-free training sets is observed between classes *i* and *j* (MCC = 0.43, see Supporting Information). These two classes have <16° difference for five of the eight dihedrals (cf. Table 1).

The second layer one-per-class classifiers show very good classification performance for classes *m* (α-helix), *d* (*β*-sheet) and for those structural alphabet letters that mainly occur at the N-termini (*k*, *l*) and C-termini (*n*, *o*) of helices (Table 4). This indicates that structure motifs of several consecutive letters can be learned by the second level one-per-class classifiers. The second layer has also learned the output behavior of all pairwise-coupling classifiers trained on pairs of structural alphabet letters (*x*, *y*), with $x, y \in \{a...p\} \wedge x \neq y$ on input vectors from structural alphabet letters *z* with $z \notin \{x, y\}$. The poorest performance is obtained for the rare classes *g* and *j*.

Applying the heuristics to derive the consensus prediction does not impair the performance, although the additional decision is a source of noise (Table 5). We suspect that this effect is outweighed by the suppression of structurally incompatible overlaps between the predicted fragments and the use of ensembles of seven classifiers per class. The mean accuracy for all residues in the test set measured as Q_{16} is 61.0%, and the median value per protein chain is 62.7%. The confusion matrix between all classes is provided as Supporting Information. To illustrate the average performance, Figure 1 shows a mutant of the K15-Transpeptidase from *Streptomyces sp.*[16] (PDB-code 1es5) for which LOCUSTRA achieves a Q_{16} (defined by eq 3) of 60.5%. It is noteworthy that even the mispredictions often predict structural alphabet letters that correspond to structures similar to the true one (green and yellow). Mispredictions are more frequent in exposed regions. The mispredicted residues in the long *β*-strand belong to the active site of the enzyme. Active sites often have strained conformations and therefore can be more difficult to predict.

3.2. Comparison to Related Approaches. Comparison of the LOCUSTRA performance to secondary structure prediction is not straightforward. Secondary structure predictions are usually three-class predictions whereas our study uses a structural alphabet with 16 classes. A simple mapping of the 16 structural alphabet letters to three classes (*m*, helix; *d*, sheet; all other, coil) obtains a Q_3 (defined by eq 3) of 79.2%. While this is in the same range as the best secondary structure prediction programs, the values are not directly comparable due to different class definitions. For a meaningful comparison with secondary structure prediction programs, we convert the 16-class and three-class predictions of LOCUSTRA and PSIPRED, respectively, to dihedral angles. The RMSDA between the resulting vectors of dihedral angles to those from the experimental structures in the PDB is our performance criterion. Figure 2 shows that the performance of both algorithms is strongly correlated which indicates that proteins that are difficult to predict for PSIPRED pose similar difficulties for LOCUSTRA. This is not surprising since both programs use the same position specific scoring matrices (calculated from PSI-BLAST) as their primary input. The median error to the experimental structure per protein chain in terms of RMSDA is 5.3% larger for PSIPRED. From the plot one can deduce that PSIPRED performs better than LOCUSTRA for some proteins with small absolute RMSDA, which in most cases corresponds to predominantly helical proteins while LOCUSTRA has an advantage for the majority of the other examples.

While finishing this study, Dong et al. introduced another method to predict the structural alphabet of de Brevern.[17] As it also uses a 2-layer approach, it is interesting to compare it to the method developed in this study. Dong et al. tested both SVM and neural networks. As they used a very large training set size (200 000 residues), neural networks provided the better cost/performance ratio. In the paper, they report an average accuracy of 58.5%. From their predictions on the independent test set (supplementary material to Dong et al.),[17] we calculated a "real world" performance of 44% accuracy. As their test set contains 24 chains common to

Table 3. Test Performance (MCC) of Pairwise-Coupling Classifiers Based on Position Specific Scoring Matrices

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*
b	0.68	X													
c	0.62	0.54	X												
d	0.65	0.48	0.54	X											
e	0.83	0.71	0.73	0.57	X										
f	0.73	0.61	0.62	0.56	0.58	X									
g	0.63	0.49	0.44	0.38	0.47	0.45	X								
h	0.78	0.57	0.67	0.57	0.74	0.67	0.67	X							
i	0.76	0.64	0.63	0.52	0.82	0.65	0.69	0.76	X						
j	0.68	0.58	0.54	0.42	0.75	0.56	0.69	0.65	0.39	X					
k	0.74	0.60	0.69	0.64	0.70	0.65	0.52	0.62	0.72	0.58	X				
l	0.77	0.61	0.67	0.62	0.78	0.65	0.53	0.72	0.64	0.55	0.65	X			
m	0.70	0.62	0.66	0.78	0.63	0.67	0.35	0.61	0.58	0.42	0.65	0.60	X		
n	0.83	0.77	0.74	0.64	0.73	0.71	0.58	0.81	0.86	0.79	0.73	0.78	0.54	X	
o	0.84	0.70	0.71	0.65	0.85	0.79	0.72	0.60	0.79	0.76	0.76	0.76	0.59	0.81	X
p	0.75	0.67	0.63	0.59	0.84	0.74	0.57	0.79	0.52	0.50	0.75	0.72	0.55	0.80	0.77

Table 4. Test Performance of Second Level One-Per-Class Classifier

	one-per-class (rounded av of 7 classifiers)				
letter	tp	tn	fp	fn	MCC
a	717	33441	491	666	0.538
b	428	32973	712	1202	0.287
c	1332	30617	1690	1676	0.390
d	5097	25422	2482	2314	0.594
e	476	33606	843	390	0.428
f	1007	31978	983	1347	0.431
g	127	33952	968	268	0.178
h	290	34089	410	526	0.371
i	267	34225	474	349	0.384
j	139	33870	1196	110	0.230
k	872	32828	549	1066	0.502
l	651	33287	327	1050	0.487
m	8280	23769	1252	2014	0.772
n	315	34502	212	286	0.552
o	417	34248	213	437	0.560
p	495	33621	495	704	0.438

Table 5. Test Performance of LOCUSTRA after Applying Consensus Heuristics

	LOCUSTRA consensus prediction					one-per-class
letter	tp	tn	fp	fn	MCC	MCC
a	802	33104	674	577	0.544	0.538
b	424	32699	836	1198	0.267	0.287
c	1339	30502	1667	1649	0.395	0.390
d	5259	25320	2490	2088	0.614	0.594
e	387	33679	613	478	0.400	0.428
f	969	31866	953	1369	0.422	0.431
g	106	33870	892	289	0.154	0.178
h	313	33825	518	501	0.366	0.371
i	226	34170	374	387	0.362	0.384
j	120	33968	940	129	0.223	0.230
k	898	32540	684	1032	0.488	0.502
l	722	32973	489	971	0.484	0.487
m	8601	23344	1544	1668	0.778	0.772
n	313	34344	212	288	0.550	0.552
o	470	33944	360	383	0.548	0.560
p	488	33489	472	708	0.438	0.438

our own test set, we performed a direct comparison of both methods. LOCUSTRA reaches an accuracy of 60% on these 24 chains compared to 45% by Dong et al.[17](Table 6). A detailed comparison is available as Supporting Information.

A second publication relevant to this study has been published recently by Zhou and co-workers.[18] For their dihedral angle prediction method based on neural networks, they report a mean absolute error of 38.2° for ψ angles and 24.8° for ϕ angles in a 10-fold cross validation, while our algorithm achieves 38.35° and 24.74°, respectively, on an independent test set. This result is remarkable as we did not optimize our algorithm for minimum angular deviation within

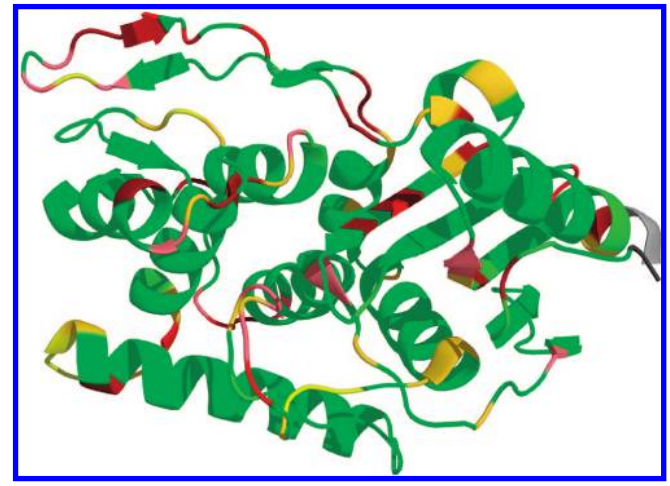

Figure 1. LOCUSTRA predictions for *Streptomyces sp.* K15-Transpeptidase[16] (PDB 1es5) colored by RMSDA error e per structural alphabet letter: green, $e < 60°$; yellow, $60° < e < 70°$; yellow-orange, $70° < e < 80°$; salmon, $80° < e < 90°$; dark-red, $90° < e < 100°$; red, $e > 100°$.

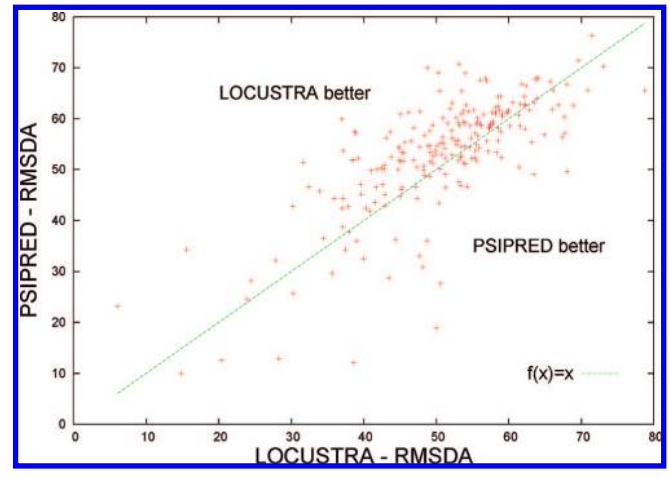

Figure 2. Comparison of RMSDA per chain between LOCUSTRA and PSIPRED on the test set.

Table 6. Comparison between LOCUSTRA and Dong et al. for the 24 Chains Common to Both Test Sets

PDB chain	length N_{res}	Dong		LOCUSTRA	
		N_{corr}	f_{corr}	N_{corr}	f_{corr}
1a6m_	147	92	0.626	114	0.776
1bdo_	76	25	0.329	44	0.579
1czpA	94	44	0.468	48	0.511
1es5A	256	104	0.406	155	0.605
1h97A	143	84	0.587	108	0.755
1i8aA	185	57	0.308	81	0.438
1jy2N	39	23	0.590	20	0.513
1jy2P	40	20	0.500	23	0.575
1kngA	140	68	0.486	97	0.693
1kqpA	267	136	0.509	193	0.723
1kyfA	243	88	0.362	148	0.609
1mfmA	149	53	0.356	68	0.456
1pfbA	51	20	0.392	34	0.667
1qlwA	314	121	0.385	136	0.433
1te5A	249	106	0.426	132	0.530
1tt8A	160	68	0.425	72	0.450
1tzvA	137	75	0.547	100	0.730
1useA	36	33	0.917	33	0.917
1v70A	101	35	0.347	76	0.752
1vhh_	153	62	0.405	80	0.523
1whi_	118	45	0.381	55	0.466
1xkrA	201	99	0.493	150	0.746
1xzoA	168	95	0.565	110	0.655
2bf9A	31	21	0.677	22	0.710
all	3498	1574	0.450	2099	0.600

N_{res} = total number of structural alphabet letters evaluated, N_{corr} = number of letters correctly predicted, f_{corr} = fraction of letters correctly predicted.

the respective fragment classes and used a much smaller training set (890 chains vs 2640).

4. DISCUSSION

4.1. Structural Alphabets as Prediction Targets. Structural alphabets have been optimized to approximate all proteins in a given set as measured by the local Cartesian or dihedral root-mean-square deviation. Unlike previous work that aimed at assessing their ability to describe three-dimensional structures, this work focuses on the development of a prediction algorithm and the methods suitable to assess the prediction performance for a given structural alphabet. From the results of our study, we conclude that the LOCUSTRA prediction scheme with the 16-letter structural alphabet of de Brevern et al. yields at least as much structural information from a protein sequence as state-of-the-art secondary structure prediction programs. The mapping of the PSIPRED predictions to dihedral angles is rather simple as we use the same dihedrals regardless whether residues are in the middle or at the end of a secondary structure element. On the other hand, we use neither training set statistics nor any motif-based error correction in the mapping scheme for LOCUSTRA.

4.2. Determinants of Classification Performance. While overall the performance is comparable to state of the art secondary structure prediction, the capability to distinguish between two structural alphabet letters is highly nonuniform. Discrimination between letters in the LOCUSTRA approach is based on detectable correlation between the sequence profile differences and structural differences of protein chain fragments. In the following we discuss factors that influence the performance.

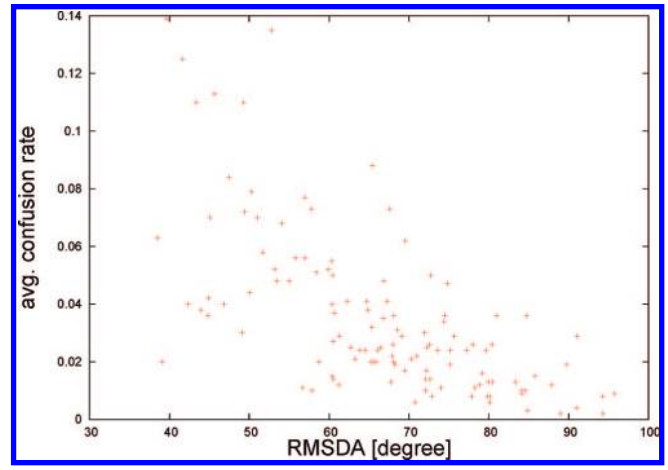

Figure 3. Average confusion rate between all 120 nonidentical pairs of structural alphabet letters as a function of the structural difference between those letters.

4.2.1. Structural Distance of Structural Alphabet Letters. It is generally assumed that the structural distance of the different fragments in the structural alphabet is large enough to imply different sequence preferences. Plotting the confusion rate between two letters against their structural distance reveals a pronounced correlation (Figure 3), indicating that the larger the distance between two structural alphabet letters is, the better our algorithm works. The correlation suggests that pairwise RMSDAs larger than 70° are required to ensure a confusion rate below 5%.

4.2.2. Quality of the Primary SVM Input. The prediction scheme relies on sequence profile similarity to select sequences that are similar in both sequence and three-dimensional environment and therefore are assumed to fold into a similar structure as the target sequence. We use PSI-BLAST, a de facto standard for iterative sequence profile search, to create the position specific scoring matrices that form the input to the first SVM layer. Errors in the PSI-BLAST multiple sequence alignment and bias in the underlying nonredundant NR database can result in profile errors that lower classification performance. Although we are aware that more accurate multiple alignment programs[19,20] and more sensitive profile-profile search methods[21,22] are available, we have chosen PSI-BLAST for this study as it is also the base for PSIPRED which we use as a performance reference for LOCUSTRA.

4.2.3. Class Bias. As we saw in the results for the pairwise-coupling classifiers on the test set (Table 3), the rare classes *g* and *j* could not be well separated from the abundant classes *d* and *m*. The much higher performance of these pairs in the symmetric training set indicates that it is rather the large difference in the number of positive and negative samples (up to 25:1 for (*m*, *g*)) than the absolute number of samples that causes the poor performance. We conjecture that the high sequence entropy tolerated in the middle of helices (*m*) and β sheets (*d*) requires a higher minimum number of samples than those for other classes to describe the class boundary in sequence space. One way to circumvent this problem is to use nonsymmetrical training sets where the natural class bias is maintained. As libSVM does not support class weighting in the ν-SVM mode, this would require a costly parameter optimization in C-SVM mode for the regularization parameters *C*, the RBF kernel

width γ, and the class weights w as we did for the second layer one-per-class classifiers.

5. CONCLUSION

Structural alphabet prediction is a straightforward extension of secondary structure prediction and accordingly can be used for the same purposes. Most approaches for global protein structure search, successful in blind predictions like the CASP competition, use secondary structure information as constraints or to generate initial structures.[23−27] Together with a distance measure,[28] they can also be used to search structures in a database.[29]

We have shown that a similar amount of local structure information is contained in the 16-class prediction of structural alphabet fragments. Therefore, structural alphabets have the potential to advance local structure prediction and supersede three-class secondary structure prediction. However, as PSIPRED is about 1 order of magnitude faster than LOCUSTRA, it will remain the method of choice in cases where the available computer power is limiting—for example, genome-wide structure predictions. Due to the enhanced resolution of coil regions, we expect LOCUSTRA to show higher performance in fold recognition applications. We are working on structural alphabets optimized for prediction and improved prediction methods which harness motif-based error correction codes.

Supporting Information Available: Pairwise-coupling classifier performance on the symmetric training set, confusion table for LOCUSTRA, comparison between LOCUSTRA, and the method of Dong et al. for 24 protein chains. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) de Brevern, A.; Etchebest, C.; Hazout, S. Bayesian Probabilistic Apporach for Predicting Backbone Structures in Terms of Protein Blocks. *Proteins: Struct., Funct., Bioinf.* **2000**, *41*, 271–287.

(2) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(3) Chen, W. W.; Yang, J. S.; Shakhnovich, E. I. A Knowledge-based Move Set for Protein Folding. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 682–688.

(4) Jones, D. T. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *J. Mol. Biol.* **1999**, *292*, 195–202.

(5) Kumar, M.; Bhasin, M.; Natt, N. K.; Raghava, G. P. S. BhairPred: prediction of β-hairpins in a protein from multiple alignment information using ANN and SVM techniques. *Nucleic Acids Res.* **2005**, *33*, W154–W159, Web Server Issue.

(6) Fuchs, P. F.; Alix, A. J. High Accuracy Prediction of β-Turns and Their Types Using Propensities and Multiple Alignments. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 828–839.

(7) Sander, O.; Sommer, I.; Lengauer, T. Local protein structure prediction using discriminative models. *BMC Bioinf.* **2006**, *7*, 14.

(8) Mooney, C.; Vullo, A.; Pollastri, G. Protein Structural Motif Prediction in Multidimensional $\Phi-\Psi$ Space Leads to Improved Secondary Structure Prediction. *J. Comput. Biol.* **2006**, *13*, 1489–1502.

(9) Bystroff, C.; Thorsson, V.; Baker, D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* **2000**, *301*, 173–90.

(10) Benros, C.; de Brevern, A. G.; Etchebest, C.; Hazout, S. Assessing a Novel Approach for Predicting Local 3D Protein Structures from Sequence. *Bioinformatics* **2006**, *62*, 865–880.

(11) Schölkopf, B.; Smola, A. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, 2002.

(12) Chang, C. C.; Lin, C. J. LIBSVM: a library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed Jul 15, 2006).

(13) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; 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*, 3389–402.

(14) Wu, T.; Lin, C.; Weng, R. Probability Estimates for Multi-class Classification by Pairwise Coupling. *J. Mach. Learn. Res.* **2004**, *5*, 975–1005.

(15) Matthews, B. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.

(16) Fonze, E.; Vermeire, M.; Nguyen-Distèche, M.; Brasseur, R.; Charlier, P. The Crystal Structure of a Penicilloyl-serine Transferase of Intermediate Penicillin Sensitivity. *J. Biol. Chem.* **1999**, *274*, 21853–21860.

(17) Dong, Q.; Wang, X.; Lin, L.; Wang, Y. Analysis amd prediction of protein local structure based on structural alphabets. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 163–172.

(18) Xue, B.; Dor, O.; Faraggi, E.; Zhou, Y. Real-value prediction of backbone torsion angles. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 427–433.

(19) Edgar, R. C.; Batzoglou, S. Multiple sequence alignment. *Curr. Opin. Struct. Biol.* **2006**, *16*, 368–373.

(20) Notredame, C. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput. Biol.* **2007**, *3*, e123.

(21) Sadreyev, R.; Grishin, N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* **2003**, *326*, 317–36.

(22) Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **2005**, *21*, 951–960.

(23) Oldziej, S.; Czaplewski, C.; Liwo, A.; Chinchio, M.; Nanias, M.; Vila, J. A.; Khalili, M.; Arnautova, Y. A.; Jagielska, A.; Makowski, M.; Schafroth, H. D.; Kazmierkiewicz, R.; Ripoll, D. R.; Pillardy, J.; Saunders, J. A.; Kang, Y. K.; Gibson, K. D.; Scheraga, H. A. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7547–7552.

(24) Verma, A.; Wenzel, W. Protein structure prediction by all-atom free-energy refinement. *BMC Struct. Biol.* **2007**, *7*, 12.

(25) Kolinski, A.; Bujnicki, J. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins: Struct., Funct., Bioinf.* **2005**, *61 (Suppl. 7)*, 84–90.

(26) Zhou, H.; Pandit, S. B.; Lee, S. Y.; Borreguero, J.; Chen, H.; Wroblewska, L.; Skolnick, J. Analysis of TASSER-based CASP7 protein structure prediction results. *Proteins: Struct., Funct., Bioinf.* **2007**, *69, S8*, 90–97.

(27) Zhang, Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins: Struct., Funct., Bioinf.* **2007**, *69, S8*, 108–117.

(28) Tyagi, M.; Venkataraman, S.; Srinivasan, N.; de Brevern, A.; Offmann, B. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 32–39.

(29) Tyagi, M.; de Brevern, A. G.; Srinivasan, N.; Offmann, B. Protein structure mining using a structural alphabet. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 920–937.

CI800178A