BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 7 2008, pages 924–931
doi:10.1093/bioinformatics/btn069

Structural bioinformatics

A comprehensive assessment of sequence-based and template-based methods for protein contact prediction

Sitao Wu and Yang Zhang*

Center for Bioinformatics and Department of Molecular Bioscience, University of Kansas, 2030 Becker Dr, Lawrence, KS 66047, USA

Received on December 13, 2007; revised and accepted on February 16, 2008

Advance Access publication February 22, 2008

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Pair-wise residue-residue contacts in proteins can be predicted from both threading templates and sequence-based machine learning. However, most structure modeling approaches only use the template-based contact predictions in guiding the simulations; this is partly because the sequence-based contact predictions are usually considered to be less accurate than that by threading. With the rapid progress in sequence databases and machine-learning techniques, it is necessary to have a detailed and comprehensive assessment of the contact-prediction methods in different template conditions.

Results: We develop two methods for protein-contact predictions: SVM-SEQ is a sequence-based machine learning approach which trains a variety of sequence-derived features on contact maps; SVM-LOMETS collects consensus contact predictions from multiple threading templates. We test both methods on the same set of 554 proteins which are categorized into 'Easy', 'Medium', 'Hard' and 'Very Hard' targets based on the evolutionary and structural distance between templates and targets. For the Easy and Medium targets, SVM-LOMETS obviously outperforms SVM-SEQ; but for the Hard and Very Hard targets, the accuracy of the SVM-SEQ predictions is higher than that of SVM-LOMETS by 12–25%. If we combine the SVM-SEQ and SVM-LOMETS predictions together, the total number of correctly predicted contacts in the Hard proteins will increase by more than 60% (or 70% for the long-range contact with a sequence separation $\geq$24), compared with SVM-LOMETS alone. The advantage of SVM-SEQ is also shown in the CASP7 free modeling targets where the SVM-SEQ is around four times more accurate than SVM-LOMETS in the long-range contact prediction. These data demonstrate that the state-of-the-art sequence-based contact prediction has reached a level which may be helpful in assisting tertiary structure modeling for the targets which do not have close structure templates. The maximum yield should be obtained by the combination of both sequence- and template-based predictions.

Contact: yzhang@ku.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

*To whom correspondence should be addressed.

1 INTRODUCTION

Ab initio protein structure predictions by folding simulations almost always fail for the medium/large size proteins approximately >120–150 residues (Aloy *et al*., 2003; Jauch *et al*., 2007; Vincent *et al*., 2005; Zhang, 2008). The major reasons include (1) the conformational phase space is too large for the limited computing power to search for and (2) the energy function is not accurate enough to guarantee the minimum at the native state. The pair-wise residue contact information can be used to constrain the simulation search within a smaller phase space which can also improve the minimum of the landscape funnel of the overall energy function (Skolnick *et al*., 1997). It is suggested that just one real contact in every eight residues will be enough to guide the simulations to generate correct folds for the single-domain proteins with up to 200 residues long (Li *et al*., 2004). In reality, the predicted contacts usually include some level of false positive data. Recent results show that contact restraints with an accuracy higher than 22% could have positive effect on the *ab initio* simulations (Zhang *et al*., 2003).

The methods of protein contact prediction can be categorized into 3 classes: (1) statistical methods using correlated mutations (Gobel *et al*., 1994; Halperin *et al*., 2006; Kundrotas and Alexov, 2006; Olmea and Valencia, 1997; Vicatos *et al*., 2005); (2) machine learning (Fariselli and Casadio, 1999; Punta and Rost, 2005; Vullo *et al*., 2006) and (3) threading template-based voting (Misura *et al*., 2006; Shao and Bystroff, 2003; Skolnick *et al*., 2004; Wu and Zhang, 2007a). There are also other combinations of the first two methods (Fariselli *et al*., 2001; Hamilton *et al*., 2004; Shackelford and Karplus, 2007).

For a given target, the correlated mutation methods collect a set of multiple sequence alignments (MSA) and identify the residue pairs mutating cooperatively along the MSA which tend to be in contact spatially. The hypothesis behind this is that during natural evolution if one residue mutates, the residues in contact will mutate as well, to compensate for the changes and keep the stability of the structure (Gobel *et al*., 1994; Vicatos *et al*., 2005). Despite the theoretical attractiveness, the prediction accuracy based on correlated mutation alone is quite poor. The highest accuracy of the correlated mutation based predictions is around 20% (Fariselli *et al*., 2001; Halperin *et al*., 2006). Recently, Shackelford and Karplus (2007) found that a statistical significance of the correlated mutation combined

with other features result in significant improvements of contact prediction.

The machine learning generates contact predictions by training the contact maps of known structures on a variety of sequence-based features including sequence profiles, secondary structure predictions and correlated mutations. The training techniques include neural networks or support vector machines, where the parameters used for prediction can be locally or globally optimized (Cheng and Baldi, 2007; Fariselli and Casadio, 1999; Pollastri and Baldi, 2002; Punta and Rost, 2005; Shackelford and Karplus, 2007; Vullo *et al*., 2006). The best prediction accuracy by the latest machine-learning techniques can achieve an accuracy around 37% (Cheng and Baldi, 2007; Vullo *et al*., 2006).

Finally, the template-based approaches identify the template proteins by threading which may have similar folds as the target. Contact predictions are then collected from the 3D template structures and alignments. This is a method that most of the tertiary-structure predictions exploit the contact restraints from (Chivian *et al*., 2005; Sali and Blundell, 1993; Shao and Bystroff, 2003; Wu and Zhang, 2007a; Zhang and Skolnick, 2004a; Zhang *et al*., 2003).

The accuracy of the template-based contact prediction methods depends on the quality of the template structures, which are sensitive to the level of homologous sequence cutoffs used in threading. Because the training of the sequence-based methods are independent of the solved protein structures, a general hope is that the sequence-based methods could generate better contact predictions than template-based methods especially for the *ab initio*/new-fold targets. Until now, there is no comprehensive examination of the accuracy of these contact predictions with regard to different template conditions. Although the recent CASP experiments provide a stringent and blind test to the sequence-based and structure-based approaches, the latter is collected from a single 3D model (using a random set with a given cutoff or rank based on residue distances) (Grana *et al*., 2005; Izarzugaza *et al*., 2007), which differs from the most popular methods in collecting contacts from multiple templates based on occurrence frequency. The data set in CASP NF category (10–20 targets) is also relatively small.

In this work, we develop two algorithms of SVM-SEQ for sequence-based contact prediction and SVM-LOMETS for template-based contact prediction. We examine the algorithms based on the same set of large-scale benchmark proteins as well as the CASP7 new-fold targets. One of the major purposes is to examine in detail the strength and weakness of the methods in different category of protein targets, which may be used as a guide for the use of contacts in 3D structure modeling.

2 SYSTEMS AND METHODS

2.1 Definition of contacts

There are a number of contact definitions which differ in the interested atoms, sequence separation and distance cutoffs. The most physically meaningful definition is based on the van der Waal's distance of atoms; but it is not commonly used. In this article, we define a pair of residues as contact if their C_α atom distance is <8 Å, which is close to the definition used in 3D structure modeling (Wu *et al*., 2007; Zhang *et al*., 2003). Depending on the sequence separation, we divide the contacts into short-, medium- and long-ranges, which correspond to the sequence separation equal to 6–11, 12–24 and >24, respectively, following the categorization in CASP7 experiments (Izarzugaza *et al*., 2007). However, CASP7 uses the C_β atoms instead of C_αs.

The effect of the contact restraints to 3D structural modeling depends on both accuracy and number of the predictions. For a given range of sequence separation, the accuracy of predictions is defined as $Acc = N_{\text{corr}}/N_{\text{pred}}$, where N_{corr} is the number of correctly predicted contacts that are true contacts in the native structure, N_{pred} is the total number of predicted contacts in the range. We assess the number of contact predictions by the percentage relative to the length of the target sequence L, i.e. $Pct = N_{\text{pred}}/L$.

2.2 SVM-SEQ

For a set of selected PDB proteins, residue pairs in the training structures are categorized into 'contacted' or 'non-contacted' based on the above-mentioned definitions. In principle, the training data should cover as many as possible residue pairs. However, including too many pairs in the training data will request a long training CPU time and large disk spaces. Second, the number of non-contacted pairs is much larger than that of contacted pairs in real structures (>20 : 1). Therefore, a training machine based on the whole set of residue pairs will be biased to the non-contacted pairs, which would result in fewer predictions for contacted pairs. For this purpose, by trial and error, we keep the ratio of non-contacted/contacted residue pairs as 4:1 by randomly selecting a subset of residue pairs.

There are two sets of features exploited.

Local window features The local features include: (1) position-specific scoring matrices (PSSM); (2) secondary structure predictions and (3) solvent accessibility predictions. The PSSM is generated by the PSI-BLAST search of the query against a non-redundant sequence database (Altschul *et al*., 1997), with 20 log-odds scores are taken at each position. The secondary structure is predicted by PSIPRED (Jones, 1999), where three states are defined as alpha-helix (represented by [0 1 0]), beta-strand [0 0 1] and coil [1 0 0]. The solvent accessibility is predicted by the neural network training (Chen and Zhou, 2005; Wu *et al*., 2007), where two states are assigned to the residues based on the solvent accessible surface area <25% (buried, represented by [0 1]) or ≥25% (exposed, by [1 0]). Both features of (2) and (3) are binary features. Since for each residue we count a 15-residue window, the number of total local features for the residue is 375 (=15 * (20 + 3 + 2)). For a residue pair (i and j), the total number of the local window features is 750.

In-between segment feature sets The residue pairs of similar local features sometime have different contact status because of the different structural components between the considered residues. To account for the contribution of these residues, we also train the contact maps on the in-between segment features. These include: (1) the number of the residues between i and j, i.e. $|i\text{-}j|$; (2) the compositional percentage of three secondary structure elements and two burial states for the in-between residues; (3) state distributions of the in-between residues, which are specified by four moments $F_n = \langle (k - \langle k \rangle)^n \rangle$, $n = 1, 2, 3$ and 4, where k (= m−i) is the position of the mth residues relative to i along the chain and each moment is calculated for five specific states of helix, strand, coil, burial and exposure and (4) the local features of five selected in-between residues which are evenly distributed between i and j. Here the window size for the local features is one. The total numbers of in-between features are 31/37/168 for each short/medium/long-range residue pair, respectively.

Following the collection of the real contact maps and the local and in-between feature data from the set of training proteins, the SVM software developed by Joachims (2002) is used to classify the contacted and non-contacted residue pairs. We also tried to train the same input

data based on neural network (NN) with a variety of parameters tuned. Based on the result of 554 testing proteins, the contact prediction accuracy of our best NN is about 30% lower than that by SVM at a given $Pct = 0.5$. One reason is probably that the final solution of SVM is a global optimum while that of NN is local (Burges, 1998).

2.3 LOMETS and SVM-LOMETS

LOMETS (Wu and Zhang, 2007a) is a local meta-threading server which include nine locally installed threading programs, i.e. FUGUE (Shi *et al*., 2001), HHSEARCH (Soding, 2005), PAINT, PPA-I, PPA-II (Wu and Zhang, 2007a), PROSPECT2 (Xu and Xu, 2000), SAM-T02 (Karplus *et al*., 2003), SP3 and SPARKS2 (Zhou and Zhou, 2004; Zhou and Zhou, 2005). For each target, LOMETS first thread a sequence through the PDB library to identify possible templates by the programs. The consensus contact residue pairs, which are ranked based on the occurrence frequency on the top 20/30/50 templates depending on the category of the target, are collected as contact predictions. The threading programs in LOMETS represent a diverse set of the state-of-the-art algorithms using different approaches: Sequence profile alignments (PPA-1, PPA-II, SPARKS2, SP3), structural profile alignments (FUGUE), pair-wise potentials (PROSPECT2, PAINT) and the hidden Markov models (HHsearch, SAM-T02). Template identification and the sequence-template alignment are therefore complementary to each other. A consensus combination of the meta-server algorithms significantly outperforms the individual threading methods (Fischer, 2003; Ginalski *et al*., 2003; Wu and Zhang, 2007a).

One defect of the LOMETS prediction is the coarse-grained distance cutoff (e.g. a distance of 7.9 Å or 8.1 Å results in different assignment of contact or non-contact despite the tiny difference). Also, the alignment quality of the templates has not been considered. In SVM-LOMETS, we use the SVM algorithm to train the distance cutoff parameters and the alignment qualities on the contact map. For each pair of residues (i and j), the training features include: (1) frequency of the contacts occurred on the top N templates where N will be decided below; (2) the average and the SD of the C_α distance (d_{ij}) calculated from the templates which have $d_{ij} < 12$ Å; (3) the continuity of the template fragments near the residues which is defined as the number of continuously aligned residues within a 5-residue window; (4) the burial depth of the residues which is calculated as the distance from the C_α atoms of i and j to the centroid of the template structure divided by the radius of gyration; (5) the average of the normalized Z-scores (Z/Z_0) of the templates with the contact occurred, where $Z = (S - \langle S \rangle)/\sigma_S$ and S is the raw-score of original threading alignment and σ_S the SD. Z_0 is a program specific Z-score cutoff to distinguish 'good' and 'bad' templates, i.e. Z_0(FUGUE) = 6, Z_0(HHsearch) = 11, Z_0(PAINT) = 0.5, Z_0(PPA-I) = 8.2, Z_0(PPA-II) = 7, Z_0(PROSPECT2) = 3.2, Z_0(SAM-T02) = 9.5, Z_0(SP3) = 8 and Z_0(SPARKS2) = 8.8; (6) the structural similarity of the templates with the contacts occurred which is calculated by the average TM-score between the templates (Zhang and Skolnick, 2004b) and (7) the predicted TM-score of the templates respective to the native which is estimated by a separate SVM training based on the Z-score and the alignment length. The average difference of the estimated TM-score and the real TM-score of the templates is 0.09 based on a set of 11 080 testing templates. The first four features correspond to the property of the specific residue pairs; the last three account for the quality of the templates which are related to the reliability of the contacts.

The final result of the SVM-LOMETS predictions is sensitive to N, the number of used templates. In general, we want to include more templates so that we can cover more contacts and have more consensus information. But including more templates will involve more false positive alignments. In SVM-LOMETS, we collect nine different training datasets with 10, 20, 30, 40, 50, 60, 70, 80 and 90 top templates where the rank of the templates is decided based on the average performance of the threading program and the specific Z-scores as described in Wu and Zhang (2007a). The targets are split into three categories of 'Easy', 'Medium' and 'Hard' based on the specific Z-scores of the programs. Therefore, there are 27 SVM classifiers trained in SVM-LOMET. A mapping table from the number of templates of Z-score $> 0.55 * Z_0$ to N will be used to determine which SVM classifier will be finally used to generate the contact predictions for a specific target. This mapping table is decided based on the best contact accuracy in the training proteins.

3 RESULTS AND DISCUSSION

3.1 Training dataset

For the purpose of training SVM-SEQ and SEQ-LOMETS, we select 500 non-homologous proteins from PDBSELECT (Hobohm and Sander, 1994) with a pair-wise sequence identity <25% and sizes ranging from 50 to 559. We remove from the training set the proteins that include broken chains or missing entities or format errors. Overall, the training data include 21 826/26 738/28 000 contacts and 87 304/106 952/112 000 non-contacts which are randomly selected in the short/medium/long-ranges, respectively. The data follows a ratio of 1:4 for the contact/non-contact residues.

3.2 Results of sequence-based methods: SVM-SEQ versus other machine-learning methods

Although there is difference in the tested proteins and the definition of contacts, the overall accuracy of the contact predictions for the same number of predictions in literature gradually increases due to the increase of databases and the improvement of algorithms (see for a summary of Table S1 in Supplementary Material).

For testing SVM-SEQ, we first generate contact predictions for a set of 173 proteins taken from Fariselli *et al*. (2001). These proteins are selected because the authors listed the names of all targets in the article so that we could have a direct comparison of our method using the same protein set. For each target, our contacts are ranked based on the confidence score defined by the SVM-light outputs (Joachims, 2002). At each sequence-separation range and for each target, we collect the top-ranked predictions with a $Pct = 0.5$, which will result in an overall $Pct = 1.5$. The accuracy ($Acc = 0.36$) of the short-range contact prediction is higher than that of the medium- (0.28) and long-range (0.25) ones. This is expected because the short-range contacts are mainly from the regularity of secondary structures which can be predicted with a high accuracy by the sequence-profile based training (Jones, 1999). But even for the hardest long-range contact prediction, the accuracy (0.25) of our method is higher than the all-range average accuracy (0.21) by Fariselli *et al*. (2001) who did not split their data into specific ranges and had an overall $Pct = 0.5$. The average overall accuracy of SVM-SEQ for all proteins is 0.29 with $Pct = 1.5$, compared with $Acc = 0.21$, $Pct = 0.5$ by Fariselli *et al*. (2001) (see Table S1).

The second test set includes 554 non-homologous proteins with a pair-wise sequence identity <25% and with length ranging from 50 to 300 residues (a list of the proteins can be found at http://zhang.bioinformatics.ku.edu/LOMETS/list1).

Table 1. Sequence-based contact predictions on 554 testing proteins

Range	Method	*Acc*	*Pct*
Short	SVM-SEQ	0.362	0.490
	SVMCON	0.320	0.490
Medium	SVM-SEQ	0.297	0.459
	SVMCON	0.279	0.459
Long	SVM-SEQ	0.222	0.457
	SVMCON	0.224	0.457
All	SVM-SEQ	0.291	1.393
	SVMCON	0.271	1.393

Table 2. Template-based contact predictions on 554 test proteins

Range	Method	*Acc*	*Pct*
Short	SVM-LOMETS	0.516	0.299
	LOMETS	0.484	0.299
Medium	SVM-LOMETS	0.499	0.324
	LOMETS	0.470	0.324
Long	SVM-LOMETS	0.444	0.591
	LOMETS	0.417	0.591
All	SVM-LOMETS	0.526	1.185
	LOMETS	0.500	1.185

To minimize the possible memorization of the training proteins, these testing proteins are non-homologous to the 500 training proteins with a sequence identity <25%. In Table 1, we show the prediction results of SVM-SEQ together with SVMCON (Cheng and Baldi, 2007) for the 554 proteins. SVMCON is one of the best sequence-based contact predictors which was ranked as top-5 in the CASP7 (only clearly outperformed by SAM-T06 depending on the way of counting scores) (Izarzugaza *et al.*, 2007). It is also the only software we can obtain and install in our local computer, which allows us to generate the predictions for the same set of test proteins and using the same contact definitions. However, we note that the proteins homologous to the training proteins of SVMCON may not be completely removed from our testing data because the training protein list of SVMCON is unknown to us.

For the convenience of comparison, we select the same number ($L/2$) of top predicted contacts in each sequence separation range and for each target. In some cases where the number of contact predictions by SVMCON is less than $L/2$ at some range, we will use the same amount of contacts for both methods for the same target, which results in the average predictions slightly less than $L/2$. As shown in Table 1, for the short- and medium-range contacts, SVM-SEQ outperforms SVMCON by 13 and 6%, which correspond to a P-value $<1.0 \times 10^{-15}$ and 0.0005 by t–test, respectively. For long-range, SVM-CON is very slightly better but the difference is not statistically significant with a P-value $= 0.79$ in t-test. The average overall accuracy of SVM-SEQ is about 7% higher than that of SVMCON with a P-value $<10 \times 10^{-6}$ by t-test.

These data show some modest (if any) advantage of SVM-SEQ in comparison with current machine-learning methods. The improvement on the short- and medium-ranges may be due to the larger set of training data combining both local and in-between features, more specific training on different sequence separation ranges, and the tuned ratio of contact/non-contact numbers. Consequently, our machine is about 50% slower than SVMCON. In the following, we will refer the sequence-based results to that by SVM-SEQ.

3.3 Results of template-based methods: SVM-LOMETS versus LOMETS

The template-based methods of LOMETS and SVM-LOMETS are tested on the same set of 554 proteins. For the convenience of comparison, we select the number of predicted contacts in SVM-LOMETS equals to that of LOMETS for each target in each sequence-separation range while the latter collects contacts based on a frequency cutoff >0.18 (Wu and Zhang, 2007a). To avoid the contamination of homologous templates, we exclude all templates with the sequence identity >20% to the target sequence from the threading template library or detectable by PSI-BLAST with an E-value <0.05 (run with the option '-j 3 –h 0.001').

As shown in Table 2, for the same number of predictions, the average accuracy of SVM-LOMETS is 0.52/0.50/0.44 for short/medium/long-range contacts, respectively, compared with the accuracy of LOMETS of 0.48/0.47/0.42. The overall accuracy of SVM-LOMETS and LOMETS are 0.53 and 0.50, respectively, a difference of 5.2% with a P-value $<1.0 \times 10^{-10}$ by t-test. This improvement of SVM-LOMETS demonstrates the effect of the detailed tuning of the distance parameters and template quality. In the following, we will refer the template-based results to that by SVM-LOMETS.

3.4 Template-based versus sequence-based methods

3.4.1 Overall result In Table 3, we summarize the contact prediction results from the template-based (SVM-LOMETS) and sequence-based (SVM-SEQ) methods. For each sequence separation range, we select the top $L/2$ predictions, which result in $1.5L$ total predictions for each target.

The results for all the 554 targets are listed in the last two rows of Table 3. For the full-range prediction, the average accuracy (0.39) of SVM-LOMETS is obviously higher than that (0.28) of SVM-SEQ. The average TM-score of the first template is 0.41, which indicates that the LOMETS templates have considerable structural similarity to the native although the homologous templates are excluded, which account for the better overall performance of SVM-LOMETS. If we further divide the predictions into different ranges, SVM-LOMETS mainly outperforms SVM-SEQ in the medium- and long-ranges (by 26 and 118%, respectively). For the short-range contacts, the accuracy of SVM-SEQ is about 3% higher than that of SVM-LOMETS, which is probably because the short-range contacts are correlated with the local secondary structure regularity and the highly-accurate (~80%) secondary structure prediction from PSIPRED has been used as the major input feature in the SVM-SEQ training. This tendency can be clearly seen in Figure 1 (Row 1), a head-to-head comparison of the two

Table 3. Sequence-based and template-based contact predictions on 554 test proteins

Target type	N[a]	TM[b]	Ns[c]	Method	Short-range		Medium-range		Long-range		All-range	
					Acc	*Pct*	*Acc*	*Pct*	*Acc*	*Pct*	*Acc*	*Pct*
Easy	220	0.544	2774	SVM-LOMETS	0.440	0.5	0.491	*0.5*	0.702	0.5	0.542	1.5
				SVM-SEQ	0.404	0.5	0.331	0.5	0.270	0.5	0.334	1.5
Medium	98	0.438	807	SVM-LOMETS	0.352	0.5	0.366	0.5	0.530	0.5	0.404	1.5
				SVM-SEQ	0.348	0.5	0.257	0.5	0.197	0.5	0.263	1.5
Hard	236	0.270	526	SVM-LOMETS	0.256	0.5	0.226	0.5	0.200	0.5	0.220	1.5
				SVM-SEQ	0.325	0.5	0.264	0.5	0.172	0.5	0.247	1.5
Very hard	16	0.190	7	SVM-LOMETS	0.305	0.5	0.215	0.5	0.097	0.5	0.206	1.5
				SVM-SEQ	0.346	0.5	0.222	0.5	0.206	0.5	0.258	1.5
All	554	0.408	1468	SVM-LOMETS	0.351	0.5	0.366	0.5	0.472	0.5	0.388	1.5
				SVM-SEQ	0.361	0.5	0.290	0.5	0.217	0.5	0.284	1.5

[a]Number of proteins in each category.
[b]Average TM-score of the first LOMETS template.
[c]Average number of homologous sequences identified by PSI-BLAST.

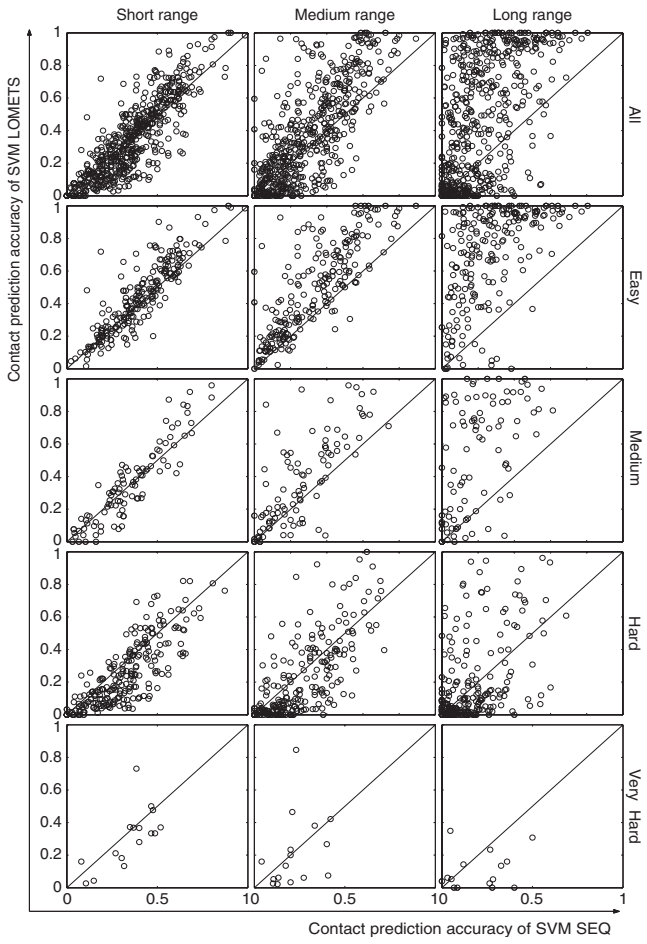

Fig. 1. The accuracy of contact predictions by SVM-LOMETS versus that by SVM-SEQ at different target categories and different sequence separation ranges, where the same $Pct = 0.5$ is taken for both predictions.

methods in different ranges, where SVM-SEQ does a much better job in short-range than in medium- and long-ranges.

3.4.2 Targets in different category We divide the 554 test targets into 'Easy', 'Medium' and 'Hard' targets according to the combined threading significance score of LOMETS, i.e. if there is at least one template with $Z > Z_0$ in each of the threading programs the target is Easy; if there is no good template of $Z > Z_0$ in any of the programs it is Hard and the others are medium target. The number of targets and the contact prediction results in each category are listed in the upper part of Table 3. A head-to-head comparison of the contact predictions for different targets is shown in Figure 1.

As expected, the performance of SVM-LOMETS is strongly correlated with the target category or the quality of the templates. Actually, the average TM-score of the first template in Easy, Medium and Hard are 0.54, 0.44 and 0.27, respectively, which results in an accuracy reduction in the whole range by 2.5 times from 0.54 to 0.22. It is worth noting that for the 'Easy' and 'Medium' targets, the accuracy of the medium- and long-range predictions is higher than that of short-range. This is partially because the total number of the native contacts in the short-range is much smaller than that in the long range. For Easy targets, because of the correctness of global topology of the threading templates, 70.2% contact predictions in long-range are correct with a $Pct = 0.5$. But even if we could predict all native contacts in the first $L/2$ predictions, the average accuracy of the short-range prediction is < 0.6 with a $Pct = 0.5$. Another reason is the extensive alignment gaps in threading which reduces further the upper-limit of the short-range contact predictions.

A somewhat unexpected result is the dependence of SVM-SEQ on the target categories since the SVM-SEQ does not exploit the template information. The accuracy of SVM-SEQ is decreased by about 35% from Easy (0.334) to Hard targets (0.247). In Figure 2, we present the prediction accuracy versus the number of sequence homologues hit by PSI-BLAST with

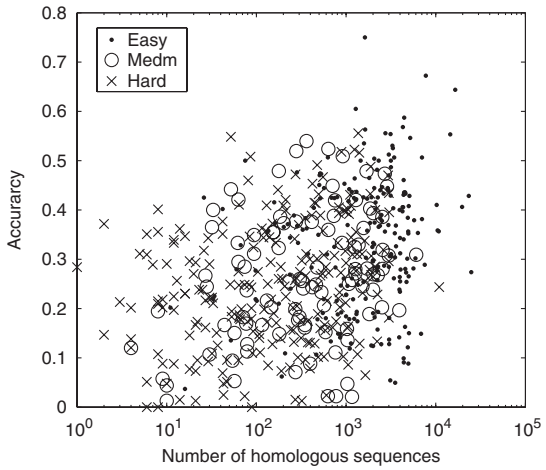

Fig. 2. The prediction accuracy by SVM-SEQ versus the number of homologous sequence by PSI-BLAST with an *E*-value <0.001 for Easy, Medium and Hard targets.

an E-value <0.001, which shows a weak but clear correlation (correlation coefficient = 0.26). Most of the Easy targets have a higher number of homologous sequences (see also column 4 of Table 3). The larger number of homologous sequences helps construct a better PSSM which SVM-SEQ has been mainly trained on. This explains the different performance of SVM-SEQ in different categories, a tendency also noticed by Shackelford and Karplus (2007).

The more interesting data in Table 3 is the performance of contact prediction in the Hard targets, where the average accuracy of SVM-SEQ for all range (0.25) is higher than that of SVM-LOMETS (0.22). This is mainly due to the good prediction of SVM-SEQ in the short- and medium-ranges which are 27% and 17%, respectively, higher than that of SVM-LOMETS. In the more important long-range contacts, SVM-SEQ (0.17) is still slightly lower than SVM-LOMETS (0.20).

In order to completely rule out the template effect, we isolate a set of 16 'Very Hard' proteins which have a TM-score < 0.23 for the first threading template and with <15 PSI-BLAST homologous sequences. This cutoff is subjective and can be replaced by other reasonable cutoffs. The average TM-score of the first templates is 0.19, which is close to a random structure pair of TM-score = 0.17 (Zhang and Skolnick, 2004b). In this new category, the accuracy of the SVM-SEQ predictions is higher than that of SVM-LOMETS in all contact ranges. Especially in the long-range prediction, the accuracy of SVM-SEQ is 0.21, which is double of that (0.10) of SVM-LOMETS.

From the data of the Hard and Very Hard targets, it is obvious that the sequence-based contact prediction can be comparable or even better than that by the template-based prediction. These may be the targets on which the sequence-based contact prediction can help 3D structure modeling. In Figure 3, we specifically count the portion of correctly predicted contacts by SVM-SEQ, which is not predicted by SVM-LOMETS. Both methods have the same number of predictions in each target. Although the percentage of the SVM-SEQ correct predictions new to SVM-LOMETS is small in Easy

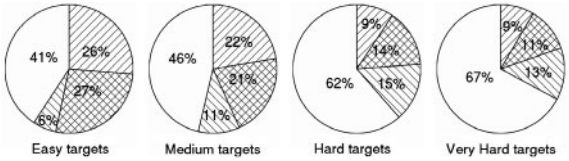

Fig. 3. The portion of the correctly predicted contacts. The forward and backward shadow regions represent, respectively, the prediction by SVM-LOMETS and SVM-SEQ, with the cross-shadow regions by both methods. The numbers indicate the percentage over the whole set of real contacts in the native structures.

(18%) and Medium (34%) targets, for the Hard (/Very Hard targets), 51%(/54%) of the correct contacts in SVM-SEQ are new to the SVM-LOMETS prediction. Although the majority of the native contacts are not predicted by either SVM-SEQ or SVM-LOMETS for Hard (/Very Hard) targets, a combination of SVM-SEQ with SVM-LOMETS can result in an enlargement of the total correct contact predictions by 62% (/67%), where the number of correct long-range contact predictions increases by 70% (/157%), compared with using SVM-LOMETS alone. This gives hope to the employment of the *ab initio* contact prediction in the structure modeling for the Hard targets (Wu and Zhang, 2007b).

3.4.3 New fold targets in CASP7 We examine the sequence- and template-based methods on the 15 new fold (NF) targets as categorized in the CASP7 experiment. Because by definition there is no similar structure solved in the PDB library for these targets, this should represent a stringent test of the template-based contact prediction methods. In fact, when we run LOMETS on a template library with templates elder than April 2006, the average TM-score of the first template is 0.23, close to the TM-score threshold for random structure pairs (~0.17), which confirms the CASP7 categorization for the NF targets.

For the CASP7 contact prediction evaluation, contacts are defined by the distance of C_β atoms. Since SVM-LOMETS and SVM-SEQ have been trained and discussed in C_α, we will still use the C_α distance to define our contact and compare it to the CASP7 C_β contacts. We also retrained our methods based on the C_β distance but there is no observable difference. For each sequence separation range, we only evaluate the top $L/5$ contact predictions, following the cutoff used in CASP7 (Izarzugaza *et al.*, 2007). The results are listed in Columns 3 and 4 of Table 4. Remarkably, for this set of proteins, SVM-SEQ significantly outperforms SVM-LOMETS in all the ranges. Especially, for the long-range contacts, because the templates are almost random, the accuracy by SVM-LOMETS is close to random (0.054) but SVM-SEQ has an average accuracy of 0.202.

We also list in Table 4 the results of SAM-T06_server by Karplus Group (Shackelford and Karplus, 2007), the best server predictor in the CASP7 experiment, with contact prediction data taken from http://predictioncenter.org/casp7/. We have excluded the targets which have less than $L/5$ contact predictions. On average, the accuracy of SVM-SEQ is a little higher than that by SAM-T06 in all sequence separation ranges. Since the number of the evaluated targets with $L/5$ contact

Table 4. Contact predictions on the FM targets of CASP7

		SVM-LOMETS	SVM-SEQ	SAM-T06_server
Short	N[1]	15 (7)	15 (7)	7 (7)
	Acc	0.180 (0.232)	0.340 (0.436)	0.184 (0.184)
	Pct	0.2 (0.2)	0.2 (0.2)	0.2 (0.2)
Medium	N	13 (13)	15 (13)	14 (13)
	Acc	0.078 (0.078)	0.233 (0.226)	0.147 (0.152)
	Pct	0.2 (0.2)	0.2 (0.2)	0.2 (0.2)
Long	N	15 (13)	15 (13)	13 (13)
	Acc	0.054 (0.055)	0.202 (0.228)	0.163 (0.163)
	Pct	0.2 (0.2)	0.2 (0.2)	0.2 (0.2)
All	N	13 (6)	15 (6)	6 (6)
	Acc	0.108 (0.158)	0.258 (0.358)	0.212 (0.212)
	Pct	0.6 (0.6)	0.6 (0.6)	0.6 (0.6)

[1]Number of proteins which have $L/5$ contact predictions in each sequence separation range. The numbers in parenthesis are those proteins having all three methods with $L/5$ contact predictions.

predictions varies among the prediction methods, in the parenthesis of Table 4, we also list the data for the common targets where all three methods have $L/5$ contact predictions. The results are similar as that of all targets. Here we note that SVM-SEQ does not include any of the CASP7 targets in the training set. Although the small set of testing targets prevent us from drawing any solid conclusion here, it seems to be safe to state that the SVM-SEQ is close to the state-of-the-art sequence-based contact prediction methods. We will further examine SVM-SEQ in the forthcoming CASP8 experiment.

4 CONCLUSIONS

We develop two algorithms of SVM-SEQ and SVM-LOMETS for protein-contact predictions. SVM-SEQ generates the predictions only based on sequence information, where secondary structures, solvent accessibility, sequence profile and sequence separations derived from the sequences are trained on contact maps by the SVM technique. Based on the same number of predictions, the accuracy of the contact prediction by SVM-SEQ is comparable to the top sequence-based machine-learning methods published in the literature and in recently CASP7 experiments.

SVM-LOMETS generates the contact predictions based on the threading templates from multiple threading programs. The contact frequency, C_α-distances and template qualities are trained on the contact maps by SVM as well. It generates slightly better contact predictions (by 5.2%) than the original LOMETS method, which collects the predictions purely based on counting the contacts on the multiple templates.

We assess the sequence-based and template-based contact predictions using the same set of 554 non-homologous proteins. First of all, even after removing the homologous templates with sequence identity >20% or detectable by PSI-BLAST, there is still a considerable portion of good templates which can be detected by the sophisticated threading techniques. Therefore, the overall accuracy of contact prediction based on templates is much higher than that by the sequence-based contact predictions. This explains the dominate rule of template-based contact prediction in guiding 3D structural modeling.

For the Hard (or Very Hard) targets where threading does not have significant alignments, SVM-SEQ generates contact predictions with an accuracy comparable or better than the template-based prediction along all sequence-separation ranges. There are more than 50% of the sequence-based correctly predicted contacts that are not generated by the template-based methods. A combination of SVM-SEQ with SVM-LOMETS results in an increase of the total number of correct contact predictions by more than 60% (or 70% if only long-range contact predictions are counted). A preliminary test demonstrates that incorporating the SVM-SEQ contact predictions in the I-TASSER simulation results in an about 5% TM-score increase for the first models of the Hard targets (Wu and Zhang, 2007b).

The most significant advantage of the sequence-based contact predictions is seen for the new fold targets as categorized in the CASP7 experiment, which have no similar template structures solved in the PDB library. While the accuracy of the threading template-based contact prediction is close to random, SVM-SEQ still generates contact predictions with about 20% of them being correct for the top $L/5$ long-range predictions. This may be the most promising category where the sequence-based contact prediction can help the tertiary structure construction.

ACKNOWLEDGEMENTS

The calculation in this work was partially done on the KU ITTC computer cluster. Y.Z. is supported by the Alfred P. Sloan Foundation. The project is supported by KU Start-up Fund 06194.

Conflict of Interest: none declared.

REFERENCES

Aloy,P. *et al.* (2003) Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins*, **53** (Suppl 6), 436–456.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Burges,C.J.C. (1998) A tutorial on Support Vector Machines for pattern recognition. *Data Mining Knowl. Discov.*, **2**, 121–167.

Chen,H. and Zhou,H.X. (2005) Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res.*, **33**, 3193–3199.

Cheng,J. and Baldi,P. (2007) Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **8**, 113.

Chivian,D. *et al.* (2005) Prediction of CASP6 structures using automated Robetta protocols. *Proteins*, **61** (Suppl 7), 157–166.

Fariselli,P. and Casadio,R. (1999) A neural network based predictor of residue contacts in proteins. *Protein Eng.*, **12**, 15–21.

Fariselli,P. *et al.* (2001) Prediction of contact maps with neural networks and correlated mutations. *Protein Eng.*, **14**, 835–843.

Fischer,D. (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins*, **51**, 434–441.

Ginalski,K. *et al.* (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.

Gobel,U. *et al.* (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Grana,O. *et al*. (2005) CASP6 assessment of contact prediction. *Proteins*, **61** (Suppl 7), 214–224.

Halperin,I. *et al*. (2006) Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins*, **63**, 832–845.

Hamilton,N. *et al*. (2004) Protein contact prediction using patterns of correlation. *Proteins*, **56**, 679–684.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci*., **3**, 522–524.

Izarzugaza,J.M. *et al*. (2007) Assessment of intramolecular contact predictions for CASP7. *Proteins*, **69**, 152–158.

Jauch,R. *et al*. (2007) Assessment of CASP7 structure predictions for template free targets. *Proteins*, **69**, 57–67.

Joachims,T. (2002) Learning to classify text using Support Vector Machines. *Dessertation*. Springer, Software available at http://svmlight.joachims.org/.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*., **292**, 195–202.

Karplus,K. *et al*. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins*, **53** (Suppl 6), 491–496.

Kundrotas,P.J. and Alexov,E.G. (2006) Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. *BMC Bioinformatics*, **7**, 503.

Li,W. *et al*. (2004) Application of sparse NMR restraints to large-scale protein structure prediction. *Biophys. J*., **87**, 1241–1248.

Misura,K.M. *et al*. (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proc. Natl Acad. Sci. USA*, **103**, 5361–5366.

Olmea,O. and Valencia,A. (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des*., **2**, S25–32.

Pollastri,G. and Baldi,P. (2002) Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, **18** (Suppl 1), S62–70.

Punta,M. and Rost,B. (2005) PROFcon: novel prediction of long-range contacts. *Bioinformatics*, **21**, 2960–2968.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*., **234**, 779–815.

Shackelford,G. and Karplus,K. (2007) Contact prediction using mutual information and neural nets. *Proteins*, **69**, 159–164.

Shao,Y. and Bystroff,C. (2003) Predicting interresidue contacts using templates and pathways. *Proteins*, **53** (Suppl 6), 497–502.

Shi,J. *et al*. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol*., **310**, 243–257.

Skolnick,J. *et al*. (2004) Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins*, **56**, 502–518.

Skolnick,J. *et al*. (1997) MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol*., **265**, 217–241.

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

Vicatos,S. *et al*. (2005) Prediction of distant residue contacts with the use of evolutionary information. *Proteins Struct. Funct. Bioinform*., **58**, 935–949.

Vincent,J.J. *et al*. (2005) Assessment of CASP6 predictions for new and nearly new fold targets. *Proteins Struct. Funct. Bioinform*., **61**, 67–83.

Vullo,A. *et al*. (2006) A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, **7**, 180.

Wu,S. and Zhang,Y. (2007a) LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res*., **35**, 3375–3382.

Wu,S. and Zhang,Y. (2007b) Could the sequence-based contact predictions be useful for protein tertiary structure modeling? *Invited talk given at MPI Conference 2007, September 30, Lawrence, Kansas*.

Wu,S. *et al*. (2007) Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol*., **5**, 17.

Xu,Y. and Xu,D. (2000) Protein threading using PROSPECT: design and evaluation. *Proteins*, **40**, 343–354.

Zhang,Y. (2008) Progress and Challenges in protein structure prediction. *Curr. Opin. Struct. Boil*., in press.

Zhang,Y. and Skolnick,J. (2004a) Automated structure prediction of weakly homologous proteins on a genomic scale'. *Proc. Natl Acad. Sci. USA*, **101**, 7594–7599.

Zhang,Y. and Skolnick,J. (2004b) Scoring function for automated assessment of protein structure template quality. *Proteins Struct. Funct. Bioinform*., **57**, 702–710.

Zhang,Y. *et al*. (2003) TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys. J*., **85**, 1145–1164.

Zhou,H. and Zhou,Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins*, **55**, 1005–1013.

Zhou,H. and Zhou,Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins*, **58**, 321–328.