Bioinformatics Advance Access published August 22, 2016

Structural Bioinformatics

Predicting the errors of predicted local backbone angles and nonlocal solvent-accessibilities of proteins by deep neural networks

Jianzhao Gao[1] , Yuedong Yang[2,*] and Yaoqi Zhou[2,*]

[1]School of Mathematical Sciences and LPMC, Nankai University, Tianjin, People's Republic of China,

[2]Institute for Glycomics and School of Information and Communication Technology, Griffith University, Parklands Dr., Southport, QLD 4222, Australia.

*To whom correspondence should be addressed.

Associate Editor: Prof. Anna Tramontano

Abstract

Motivation: Backbone structures and solvent accessible surface area of proteins are benefited from continuous real value prediction because it removes the arbitrariness of defining boundary between different secondary-structure and solvent-accessibility states. However, lacking the confidence score for predicted values has limited their applications. Here we investigated whether or not we can make a reasonable prediction of absolute errors for predicted backbone torsion angles, Cα-atom-based angles and torsion angles, solvent accessibility, contact numbers and half-sphere exposures by employing deep neural networks.

Results: We found that angle-based errors can be predicted most accurately with Spearman correlation coefficient (SPC) between predicted and actual errors at about 0.6. This is followed by solvent accessibility (SPC~0.5). The errors on contact-based structural properties are most difficult to predict (SPC between 0.2 and 0.3). We showed that predicted errors are significantly better error indicators than the average errors based on secondary-structure and amino-acid residue types. We further demonstrated the usefulness of predicted errors in model quality assessment. These error or confidence indictors are expected to be useful for prediction, assessment, and refinement of protein structures.

Availability: The method is available at http://sparks-lab.org as a part of SPIDER2 package.

Contact: yuedong.yang@griffith.edu.au or yaoqi.zhou@griffith.edu.au

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

1 Introduction

Predicting 3-dimensional structures directly from protein sequences without employing homologous structures as templates remains an unsolved problem although substantial progresses have been made(Dill and MacCallum, 2012; Tai, et al., 2014; Zhou, et al., 2011). Meanwhile, many machine-learning-based methods have been developed for the easier problems: sequence-based prediction of one-dimensional or two-dimensional structural properties such as secondary structures, backbone torsion angles, solvent accessible surface area, and contact maps(Ali, et al., 2014; Kurgan, et al., 2008; Kurgan and Disfani, 2011; Singh, et al., 2014; Zhou and Faraggi, 2010). These predicted structural properties have been employed as restraints for three-dimensional structure prediction (Dill and MacCallum, 2012; Tai, et al., 2014; Zhou, et al., 2011). As more protein structures are experimentally determined and machine-learning techniques become increasingly more powerful, the prediction

accuracy for these structural properties continues to improve(Eickholt and Cheng, 2012; Heffernan, et al., 2015; Wang, et al., 2016). One noticeable trend in these predictions is the move from classification of discrete states such as solvent accessibility states (buried or accessible) and three-state secondary structure to continuous, real values of backbone torsion angle and solvent accessible surface area (Zhou and Faraggi, 2010).

One advantage of real-value prediction is the removal of arbitrariness in defining the boundary between discrete states. For example, a pre-set 25% accessibility was employed to define an accessible or buried state (Kim and Park, 2004; Nguyen and Rajapakse, 2005). Different secondary structure assignment techniques can disagree with each other as much as 15-25% (Colloc'h, et al., 1993; Zhang, et al., 2008) because there are no well-defined boundaries to separate a non-ideal helical or strand conformation from a coil conformation. Real-value prediction has one disadvantage, however. For multi-state classifications, predicted actual value for one class can be employed to compute the confidence or probability score (Jones, 1999). However, predicted real values do not have corresponding confidence scores. Without confidence scores, the usefulness of predicted structural properties as restrains for three-dimensional structure prediction is limited (Faraggi, et al., 2009).

The objective of this paper is to examine whether or not confidence scores for predicted real values of protein structural properties are predictable in a reasonable accuracy. Recently, we have developed a method called SPIDER2 that provides accurate real-value prediction of backbone torsion angles, Cα-atom-based angles and torsion angles, solvent accessibility, contact numbers and half-sphere exposure (Heffernan, et al., 2016; Lyons, et al., 2014). Here we developed a corresponding method called SPIDER-Delta to predict the confidence of those predicted structural properties. We found that predicted errors for angles are highly correlated with actual angle errors while errors for contact-based structural properties are more difficult to predict.

2 Methods and Materials

Datasets: To avoid the over-training, we employed exactly the same datasets employed for training and test of SPIDER2 (Heffernan, et al., 2015; Lyons, et al., 2014). This dataset contains 5789 proteins with sequence identity less than 25% and X-ray resolution better than 2Å, in which 4590 proteins are employed as training and cross validation (TR4590) and 1199 proteins as an independent test dataset (TS1199). SPIDER was also tested by the targets from critical assessment of structure prediction technique (CASP11, http://www.predictioncenter.org/casp11/). This independent test set (CASP11) contains 72 proteins after removing redundancy within CASP targets and to TR4590 and TS1199 with a sequence identity cutoff of 30%.

To demonstrate the usefulness of predicted errors, we downloaded all top 1 models predicted by servers for 72 proteins in CASP11 (a total of 3017 models, CASP11MOD). The local structural quality of each model is evaluated by sequence-position-dependent S-score (Ray, et al., 2012). $S_i=1/(1+(d_i/d_0)^2$, where $d_0=3Å$, d_i is the distance between the residue i in the model structure and the same residue in the native structure after pairwise structural alignment by SPalign(Yang, et al., 2012).

Deep neural-network architecture: We employed the deep neural network implemented by Palm (Palm, 2012) to build the model for predicting the confidence of predicted structural properties. The unsupervised weights were initialized by stacked sparse auto-encoder with learning rate of 0.05. Then the weights were further refined by standard backward propagation. The neural networks consist of three hidden layers, with 150 hidden neurons in each layer. The learning rates for different layers are 1.0, 0.5, 0.2, and 0.05, respectively. The flowchart of deep neural networks is show in **Fig. 1.**

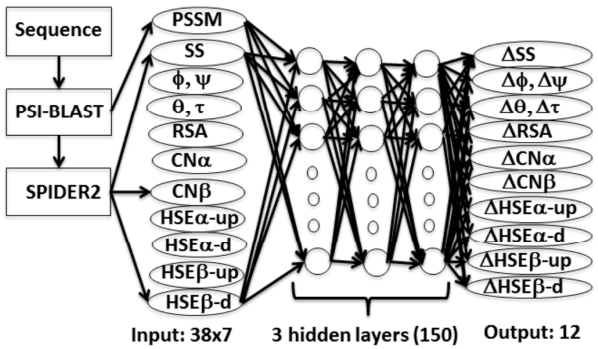

Fig. 1. Flowchart of deep neural networks. The input layer consists of 266 features (sequence profiles from PSI-BLAST and predicted structural properties by SPIDER2). There are three hidden layers, each of which has 150 neurons. The output layer contains predicted absolute deviations of 12 predicted structural properties from their actual values.

Input features: A total of 38 input features for a given amino acid residue are made of predicted structural properties from SPIDER2 (18 features) and Position Specific Scoring Matrix (PSSM) generated by PSI-BLAST(Altschul, et al., 1997) with three iterations of searching against NR database with an E-value of 0.001 (20 features). Predicted structural properties from SPIDER2 include probabilities for three types of secondary structure (3 features), relative solvent accessibility (RSA) (1 feature), cosine/sine functions of backbone φ and ψ angles and Cα-atom-based angle θ and rotational angle τ (2*4=8 features), contact numbers based on Cα and Cβ atoms (CNα and CNβ, 2 features), respectively, and up and down half-sphere exposures (HSE) based on the Cα-Cβ vector and the Cα-Cα vector (HSEβ-up, HSEβ-down, HSEα-up, and HSEα-down, 4 features), respectively. Here, φ and ψ angles are the rotational angles about the N-Cα bond and the Cα-C bond, respectively. θ_i for residue i is the angle between $C\alpha_{i-1}$-$C\alpha_i$-$C\alpha_{i+1}$ and τ_i is the angle rotated about the $C\alpha_i$-$C\alpha_{i+1}$ bond. Contact numbers (CNα and CNβ) are number of residues within 13Å of a residue's Cα or Cβ atom, respectively. HSE is the residue-residue contact number in upper or down half sphere according to a pre-specified vector (the Cα-Cβ vector or the Cα-Cα vector) (Hamelryck, 2005). We also used a sliding window size of 7 (3 amino acids at each side of the query amino acid residue) to represent each residue. This leads to 266 input features for per residue as shown in **Fig. 1.**

Outputs: We are aiming to predict absolute errors between predicted and measured structural properties for twelve predicted structural properties (ΔSS, Δφ, Δψ, Δθ, Δτ, ΔRSA, ΔCNα, ΔCNβ, ΔHSEβ-up, ΔHSEβ-down, ΔHSEα-up, and ΔHSEα-down). ΔSS is an error indicator of predicted secondary structure. ΔSS=1, if the predicted secondary structure is the same as the actual secondary structure according to three-state prediction, and 0 otherwise. $\Delta\theta=|\theta^{Pred}-\theta^{Expt}|$. For rotational angles φ, ψ, and τ, the smaller value of $|Angle^{Pred}-Angle^{Expt}|$ or $360°-|Angle^{Pred}-Angle^{Expt}|$ is employed as the prediction target to account for the angle periodicity. ΔRSA is the absolute difference between predicted and actual relative solvent accessibility. Similarly, ΔCNα, ΔCNβ, ΔHSEβ-up, ΔHSEβ-down, ΔHSEα-up, and ΔHSEα-down are absolute errors in contact numbers and half sphere exposures, respectively. There are a total of 12 outputs.

Training, test and performance evaluation. The neural network model was trained by ten-fold cross validation with TR4590 and independently tested by TS1199 and CASP11. In ten-fold cross validation, the training dataset was randomly divided into ten subsets. Nine subsets were employed for training and the remaining one subset was for test. This process repeated ten times so that all subsets were employed for test. The performance for ΔSS was evaluated by the area under the receiver operating characteristic curve (AUC). All other predicted errors were evaluated by the Pearson correlation coefficient (PCC), Spearman correlation coefficient (SPC) and the mean absolute error (MAE) between predicted and actual errors.

Model Quality Assessment. To evaluate model quality using predicted errors, we obtained the number of residues with actual (ΔV^{actu}) and predicted errors (ΔV^{pred}) for each variable V, $N(\Delta V^{actu}, \Delta V^{pred})$, based on the results from 10-fold cross validation on the TR4590 set. ΔV were divided into 18 bins from minimal to maximal values (i, j=1,…,18) for four torsion angles (N_{bin}=18) and 20 bins (N_{bin}=20) for other structural properties. An energy score is calculated by $E_{ij}=log(P_{ij} * N_{bin})$ where $P_{ij}=N(i,j)/\sum_i N(i,j)$. Then, for each model structure, we can obtain 1) ΔV^{pred} from our method for each sequence position m, 2) ΔV^{actu} by using predicted V from SPIDER2 and assuming actual V from the model structure for each sequence position, 3) a sequence-position dependent energy score $E_{ij}(m)$ according the (i, j) bins that ΔV^{actu} and ΔV^{pred} belong to and 4) the average of the neighboring energy scores with a window size of 7 $(q_m=\sum_k E(k+m)/7, k=-3,-2,-1,0,1,2,3)$. This window-based q-score is used to calculate the correlation to the actual local structural quality S-score defined above. Here q-score can be evaluated for 11 predicted errors, separately, except discrete secondary structures.

Table 1. Results of error prediction by ten-fold cross validation and independent tests

Prediction	TR4590 (Ten-fold cross validation)			TS1199			CASP11		
Target	PCC[a]	SPC	MAE	PCC	SPC	MAE	PCC	SPC	MAE
Δφ	0.52±0.00	0.59±0.00	12.15±0.14	0.52	0.60	12.89	0.45	0.53	13.35
Δψ	0.54±0.00	0.61±0.00	20.28±0.31	0.53	0.62	22.26	0.50	0.56	22.17
Δθ	0.56±0.00	0.62±0.00	4.32±0.05	0.56	0.63	4.52	0.53	0.60	4.52
Δτ	0.56±0.01	0.66±0.00	20.75±0.32	0.56	0.67	22.32	0.53	0.64	22.23
ΔRSA	0.46±0.00	0.48±0.00	7.54±0.07	0.46	0.48	7.80	0.42	0.44	8.12
ΔHSEα-up	0.35±0.00	0.33±0.00	2.68±0.02	0.35	0.33	2.70	0.28	0.28	2.80
ΔHSEβ-up	0.35±0.00	0.33±0.00	2.61±0.01	0.35	0.33	2.63	0.28	0.28	2.73
ΔCNα	0.30±0.01	0.26±0.01	3.29±0.02	0.30	0.27	3.35	0.19	0.18	3.43
ΔCNβ	0.30±0.01	0.26±0.01	3.29±0.03	0.30	0.27	3.34	0.19	0.18	3.43
ΔHSEβ-down	0.29±0.01	0.26±0.01	2.23±0.01	0.29	0.27	2.26	0.21	0.2	2.32
ΔHSEα-down	0.26±0.01	0.23±0.01	2.10±0.01	0.26	0.23	2.12	0.19	0.17	2.15

[a]PCC and SPC are Pearson and Spearman correlation coefficients between predicted and actual absolution errors, respectively. MAE: Mean absolute error between predicted and actual absolute differences.

3 Results

Table 1 summarizes the performance of error prediction by ten-fold cross validation and independent tests. Predicted absolute errors for all angles (Δφ, Δψ, Δθ, and Δτ) have a strong correlation with actual angle errors. Correlation coefficients in ten-fold cross validation are between 0.52 to 0.56 for PCC and between 0.59 to 0.66 for SPC with low standard deviation between 10 folds. MAE values range from 4° for Δθ, 12° for Δφ, 20° for Δψ to 21° for Δτ. These MAE values for different angle errors follow the same trend as the MAE values for angles given by SPIDER (8° for θ, 19° for φ, 30° for ψ and 32° for τ). A much higher SPC than PCC values for all angle errors indicate nonlinear relations between predicted and actual errors.

The performance of error prediction for RSA is lower than those of angle errors with PCC at 0.46 and SPC at 0.48. This is followed by the upper half-sphere contacts (ΔHSEα-up and ΔHSEβ-up) with PCC at 0.35, contact numbers (ΔCNα and ΔCNβ) with PCC at 0.30, and down half-sphere contacts (ΔHSEβ-down and ΔHSEα-down) with PCC at 0.29 and 0.26, respectively. That is, errors in contact numbers are the most difficult to predict.

Table 1 further shows that there is essentially no difference in performance between ten-fold cross-validation by TR4590 and independent test by TS1199. This highlights the robustness of the method developed. Slightly worse performance was observed for the CASP 11 set (CASP11), confirming that the CASP targets are more challenging to predict as shown previously(Heffernan, et al., 2016; Heffernan, et al., 2015; Lyons, et al., 2014).

We also predicted the probability that the predicted secondary structure is the same as the actual secondary structure (ΔSS). We found that the area under the curve is 0.832 for ten-fold cross-validation, 0.818 for TS1199, and 0.799 for CASP11. The ROC curves are show in supplement **Fig. S1**. Again, the performance for ten-fold cross validation on TR4590 is nearly identical to that for independent test on TS1199, highlighting the robustness of the method obtained. We employed predicted secondary structure probabilities given by SPIDER2 directly. The resulting AUC values are 0.807 for TS1199 and 0.800 for CASP11, respectively. This suggests that ΔSS provide marginal improvement from the original secondary structure probability from SPIDER2 for the large test set TS1199.

To evaluate the usefulness of predicted errors as a confidence score, we sort all amino acid residues in TS1199 according to predicted errors in increasing order along with their corresponding actual error values. Then we calculate average actual errors for top 1%, 2%, …, 99% and

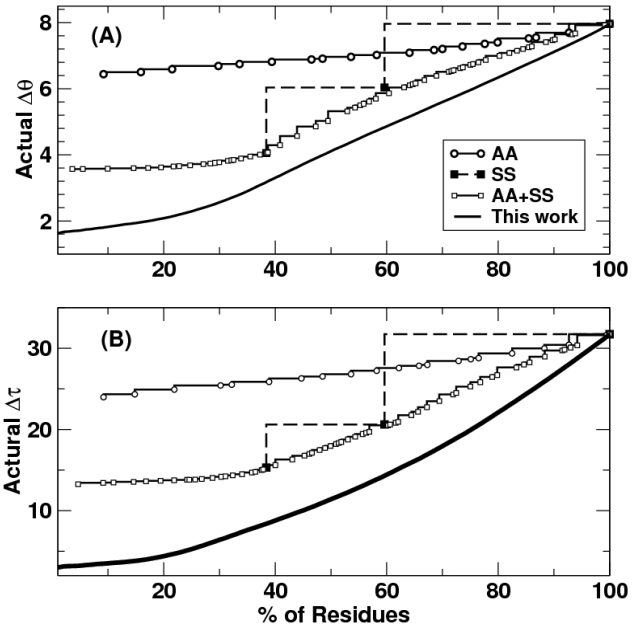

Fig. 3. The average actual error as a function of percentage of residues (1% to 100%) that are sorted according to their predicted errors in the independent test set TS1199. This curve is compared to percentage of residues (1% to 100%) sorted according to the average errors based on predicted secondary structure (SS) types, amino acid residue types (AA), and both as labeled. (A) Δθ and (B) Δτ angle.

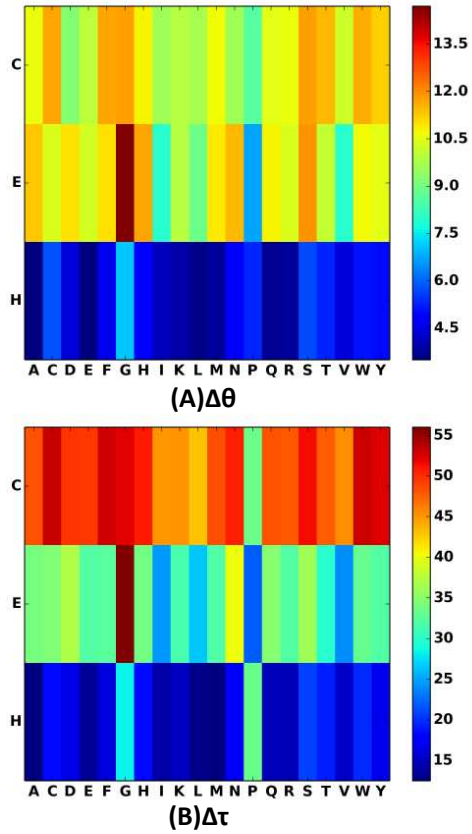

Fig. 2. Heat map for Δθ (A) and Δτ (B) based on the test dataset TS1199 in secondary structure types (coil (C), sheet (E) and helix (H)) and in 20 amino acid residue types.

100% of the sorted residues. As an example, results for Δθ and Δτ are shown in **Fig. 2**. The average actual errors monotonically increase according to sorted predicted errors, consistent with strong positive correlation (PCC~0.5 and SPC~0.6) between predicted and actual errors for Δθ and Δτ. One can also estimate the prediction errors according to residue type, secondary structure type or both. This can be done by obtaining the average actual errors for 20 amino acid types, predicted three secondary structure types and secondary-structure-dependent amino acid types (60 values) along with the cumulative percentages of residues covered by these categories sorted according to average actual errors. As **Fig. 2** shows, predicted errors (black line) are significantly better in separating those residues with highly accurate angles (low actual errors) from poorly predicted angles (high actual errors) and thus are more reliable confidence indicators than residue and secondary-structure-based classifications. Similar results are observed for other structural properties (Δφ, Δψ, ΔRSA, ΔCNα, ΔCNβ, ΔHSEβ-up, ΔHSEβ-down, ΔHSEα-up, and ΔHSEα-down) shown in Supplement **Fig. S2A-H**.

Similarly, the average accuracy of predicted secondary structures can be plotted as the cumulative percentage from 1% to 100% of residues sorted according to predicted probabilities of errors in secondary structures (ΔSS) or according to predicted secondary structure probability by SPIDER2. As shown in supplement **Fig. S3**, the performance of ΔSS is only marginally better than that of predicted secondary structure probability by SPIDER2 by identifying higher percent of accurately predicted residues.

Fig. 3A and 3B show two-dimensional heatmaps in secondary structure type (C for coil, E for sheet, H for helix) and amino acid type in single letter code for Δθ and Δτ, respectively. Angle errors in helical residues are accurately predicted regardless the type of amino acid residue. The highest errors in Δθ and Δτ interestingly are glycine (G) in sheet conformation, suggesting the difficulty in pinpointing the flexible glycine (G) in the sheet conformation. The heat maps for Δφ and Δψ (**Fig. S4**), however, show that the highest errors in Δφ and Δψ are glycine (G) and tryptophan (W) in coil conformation, respectively.

As an example, **Fig. 4** compared predicted and actual Cα-based torsion angle Δτ for zinc protease from actinobacteria *Streptomyces caespitosus* (PDB ID 1c7kA in TS1199). The structures are color-coded according to predicted and actual Δτ in **Fig. 4A** and **4B**, respectively. These Δτ values are also shown as a function of residue index in **Fig. 4C**. The Pearson correlation coefficient between predicted and actual values is 0.56, which is the same as the overall PCC value on TS1199. Consistent with high correlations, our predicted Δτ captures highly accurate (blue) areas reasonably well but with under-predicted Δτ for largest errors in coil regions, in particular. This is largely due to inability of our neural network methods to predict largest errors (extreme values).

To examine the usefulness of predicted errors in assessing model quality, we calculated q-scores for 11 predicted errors and their correlations with actual model quality S-scores on the CASP11MOD dataset. We found that PCC is the strongest (0.46) for Δτ and between 0.33 and 0.44 for ΔCNα (0.33), ΔCNβ (0.33), ΔHSEα-down (0.35), ΔHSEβ-down (0.35), Δφ (0.37), ΔHSEβ-up (0.40), ΔHSEα-up (0.40), Δθ (0.40), Δψ (0.44). The weakest correlation was observed for ΔRSA (0.19). The statistically significant correlation (p-value<2.2×10^{-16}) for all predicted errors confirmed the usefulness of these variables as novel features for quality assessment.

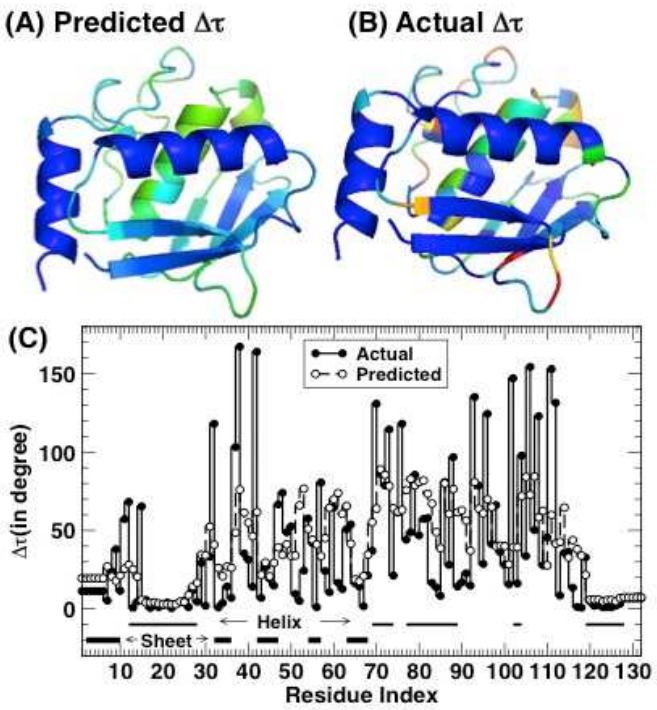

Fig. 4. Protein structure (PDB #1c7kA) in TS1199 color-coded in the same scheme according to predicted (A) and actual (B) Δτ from blue (low errors) to red (high errors). (C) Δτ as a function of residue index. Helical and sheet regions are as labeled. The Pearson correlation coefficient of actual error and predicted error is 0.56.

4 Discussion

We have developed a method called SPIDER-Delta dedicated to prediction of the errors of predicted real-value structural properties include backbone torsion angles, Cα-based angles and torsion angles, solvent accessible surface area and contact numbers. The method has the best performance for angles with SPC about 0.6 between predicted and actual values, followed by relative ASA (SPC about 0.5) for ten-fold cross validation (TR4590) and independent test (TS1199). However, the errors for contact numbers are not as accurately predicted (SPC at 0.2-0.3). Predicted errors in secondary structures only provide marginal improvement over the direct use of predicted secondary structure probability. The performance of the method is robust as the results of ten-fold cross validation are essentially the same as the results of the large independent test (TS1199).

It is of interest to know the importance of various features in error prediction. There are a total of 266 features based on 20 PSSM and 18 predicted structural properties and a sliding window size of 7 residues. However, it is computationally infeasible to examine 266 features individually with nearly one million amino acid residues in the training set for deep neutral networks. Thus, we investigated the performance of features according to different groups (secondary structure probability (SS), relative solvent accessibility (RSA), angles (Angles), contact number and half sphere exposure (HSE), and position specific scoring matrix (PSSM)). We evaluated the impact of removing one group at a time on the average of mean absolute errors of 12 target values (Δφ, Δψ, Δθ, Δτ, ΔRSA, ΔSS, ΔHSEα-up, ΔHSEβ-up, ΔCNα, ΔCNβ, ΔHSEβ-down, ΔHSEα-down) based on 10-fold cross validation. We found that removing torsion angles leads to the largest increase in error (3%) but only minor increases (0.2-0.4%) for other group features, including PSSM. This is likely due to the fact that all structural properties were predicted by PSSM. Nevertheless, all features have made statistically significant (p-values$<5\times10^{-6}$), positive contributions in improving prediction of errors.

We also examined the usefulness of other features in error prediction by employing the consistency score between various predicted structural properties. For example, the consistency between predicted Cα-based angles/torsion angles and those angles calculated from predicted backbone torsion angles and the root-mean-squared distance between fragment structures generated from Cα-based angles/torsion angles and those from backbone torsion angles were found useful as an indicator of accuracy of local structures (Lyons, et al., 2014). However, we found that addition of these two consistency-based features did not lead to further significant improvement in predicted confidence scores.

The above studies were based on deep learning neutral networks. The large training data (~1 million residues) and 12 outputs prevented us to test other machine-learning techniques. For example, SVM would be computationally too slow to train and test. Moreover, many previous studies have concluded that deep neural networks are superior to SVM, regular neural network (NN) and other models in learning from large data sets(Bengio, et al., 2013).

The predicted errors, however, should be employed as a confidence or probability score for a predicted structural property, rather than directly utilized as the absolute error of the given structural property. This is true even for most accurately predicted angle errors. As illustrated in **Fig. 4**, although predicted errors are highly correlated to actual errors, predicted values are systematically smaller than actual values for largest errors in particular. This is because of the inherent difficulty of machine-learning methods to predict extreme values (Faraggi, et al., 2009). Our future study will investigate if building a dedicated predictor for predicted coil residues will enhance the ability to predict large errors because most large errors are belong to coil residues. Separate treatment of helix, sheet, and coil in an ensemble learning was shown important for improving prediction of mutation-induced changes in protein stability (Folkman, et al., 2016).

Despite this limitation, **Fig. 2** shows that predicted errors are a much better indicator than the error indicated based on the average errors according to predicted secondary structures and amino-acid residue types. The latter was employed to demonstrate that real-value predicted torsion angles are more useful as restraints than three-state secondary structure types in fragment-free protein structure prediction (Faraggi, et al., 2009). Thus, we expect that predicted errors are more useful as restraints for protein structure prediction. Similarly, in template-based techniques such as SPARKS X (Yang, et al., 2011), matching template structural properties with predicted properties based on estimated errors has improved the ability of recognizing correct structural folds. More accurately predicted errors will be likely useful to further improve fold recognition.

To directly test the usefulness of predicted errors, we employed them as model quality assessment scores by calculating q-scores from predicted errors. For most predicted errors, predicted q-scores have a statistically significant correlation to actual local structure-quality S-scores with the highest PCC value at 0.46 for a single feature of Δτ. This suggests

that predicted errors are potentially useful new features for further improving existing methods for model assessment.

SPIDER-Delta obtained here is available at http://sparks-lab.org as a part of the SPIDER2 structure-property-prediction package. Because we are predicting the error bound for a predicted value, it is inevitable for our method to be tied with a specific predictor to obtain the predicted value (in our case SPIDER2). If one wants to avoid this dependence on a specific predictor, it is necessary to employ multiple predictors and make prediction of the average errors of these multiple predictors. A method like this would reveal the regions whose structural properties such as backbone angles that are most difficult to predict for any methods. This is an interesting subject deserving further studies.

Acknowledgements

We gratefully acknowledge the support of the Griffith University eResearch Services Team and the use of the High Performance Computing Cluster "Gowonda" to complete this research. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

Funding

J.G. was supported by Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP) grant 20130031120001. This work is also supported by National Health and Medical Research Council of Australia (Contract grant numbers: 1059775 and 1083450); Australian Research Council's Linkage Infrastructure, Equipment and Facilities funding scheme (Contract grant number: LE150100161) to Y.Z.

Conflict of Interest: none declared.

References

Ali, S.A., *et al.* (2014) A Review of Methods Available to Estimate Solvent-Accessible Surface Areas of Soluble Proteins in the Folded and Unfolded States, *Curr Protein Pept Sc*, **15**, 456-476.

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

Bengio, Y., Courville, A. and Vincent, P. (2013) Representation Learning: A Review and New Perspectives, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1798-1828.

Colloc'h, N., *et al.* (1993) Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment, *Protein engineering*, **6**, 377-382.

Dill, K.A. and MacCallum, J.L. (2012) The protein-folding problem, 50 years on, *Science*, **338**, 1042-1046.

Eickholt, J. and Cheng, J.L. (2012) Predicting protein residue-residue contacts using deep networks and boosting, *Bioinformatics*, **28**, 3066-3072.

Faraggi, E., *et al.* (2009) Predicting Continuous Local Structure and the Effect of Its Substitution for Secondary Structure in Fragment-Free Protein Structure Prediction, *Structure*, **17**, 1515-1527.

Folkman, L., *et al.* (2016) EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models, *J Mol Biol*, **428**, 1394-1405.

Hamelryck, T. (2005) An amino acid has two sides: a new 2D measure provides a different view of solvent exposure, *Proteins: Structure, Function, and Bioinformatics*, **59**, 38-48.

Heffernan, R., *et al.* (2016) Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins, *Bioinformatics*, **32**, 843-849.

Heffernan, R., *et al.* (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, *Sci Rep-Uk*, **5**, 11476.

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

Kim, H. and Park, H. (2004) Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor, *Proteins*, **54**, 557-562.

Kurgan, L., *et al.* (2008) Sequence-Based Methods for Real Value Predictions of Protein Structure, *Curr Bioinform*, **3**, 183-196.

Kurgan, L. and Disfani, F.M. (2011) Structural Protein Descriptors in 1-Dimension and their Sequence-Based Predictions, *Curr Protein Pept Sc*, **12**, 470-489.

Lyons, J., *et al.* (2014) Predicting backbone Calpha angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network, *Journal of computational chemistry*, **35**, 2040-2046.

Nguyen, M.N. and Rajapakse, J.C. (2005) Prediction of protein relative solvent accessibility with a two-stage SVM approach, *Proteins*, **59**, 30-37.

Palm, R.B. (2012) Prediction as a candidate for learning deep hierarchical models of data, *Technical University of Denmark, Palm*, **25**.

Ray, A., Lindahl, E. and Wallner, B. (2012) Improved model quality assessment using ProQ2, *BMC Bioinformatics*, **13**, 224.

Singh, H., Singh, S. and Raghava, G.P. (2014) Evaluation of protein dihedral angle prediction methods, *PloS one*, **9**, e105667.

Tai, C.H., *et al.* (2014) Assessment of template-free modeling in CASP10 and ROLL, *Proteins*, **82 Suppl 2**, 57-83.

Wang, S., *et al.* (2016) Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields, *Sci Rep-Uk*, **6**.

Yang, Y., *et al.* (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates, *Bioinformatics*, **27**, 2076-2082.

Yang, Y., *et al.* (2012) A new size - independent score for pairwise protein structure alignment and its application to structure classification and nucleic - acid binding prediction, *Proteins*, **80**, 2080-2088.

Zhang, W., Dunker, A.K. and Zhou, Y.Q. (2008) Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks, *Proteins-Structure Function and Bioinformatics*, **71**, 61-67.

Zhou, Y. and Faraggi, E. (2010) Prediction of one-dimensional structural properties of proteins by integrated neural network. In Rangwala, H. and Karypis, G. (eds), *Protein Structure Prediction: Method and Algorithms (*Wiley, Hoboken, NJ, pp. 44-74.

Zhou, Y.Q., *et al.* (2011) Trends in template/fragment-free protein structure prediction, *Theor Chem Acc*, **128**, 3-16.