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

Jianzhao Gao¹, Yuedong Yang²,* and Yaoqi Zhou²,*

¹School of Mathematical Sciences and LPMC, Nankai University, Tianjin, People’s Republic of China and ²Institute 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: Anna Tramontano

Received on June 8, 2016; revised on July 29, 2016; accepted on August 18, 2016

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 indicators are expected to be useful for prediction, assessment, and refinement of protein structures.

Availability and Implementation: 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 three-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 and Disfani, 2011; Kurgan et al., 2008; 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 (Eichkolt and Cheng, 2012; Heffernan et al., 2015; Wang et al.,...
2 Methods and materials

2.1 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 2 Å, 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 protein structure prediction technique (CASP11, http://www.predic tioncenter.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 \( d_0 = 3 \AA \), \( d_i \) 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).

2.2 Deep neural-network architecture

We employed the deep neural network implemented by 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 shown in Figure 1.

2.3 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 \( \phi \) and \( \psi \) angles and Cz-atom-based angle \( \theta \) and rotational angle \( \tau \) (2*4 = 8 features), contact numbers based on Cz and Cβ atoms (CNz and CNβ, 2 features), respectively, and up and down half-sphere exposures (HSE) based on the Cz-Cβ vector and the Cz-Cα vector (HSEβ-up, HSEβ-down, HSEz-up, and HSEz-down, 4 features), respectively. Here, \( \phi \) and \( \psi \) angles are the rotational angles about the N-Cα bond and the Cα-Cβ bond, respectively. \( \theta \) for residue \( i \) is the angle between \( C_{\alpha(i-1)} - C_{\alpha(i)} - C_{\alpha(i+1)} \) and \( \tau_i \) is the angle rotated about the \( C_{\alpha(i-1)} - C_{\alpha(i)} \) bond. Contact numbers (CNz and CNβ) are number of residues within 13 Å of a residue’s Cz or Cβ atom, respectively. HSE is the residue-residue contact number in upper or down half sphere according to a pre-specified vector (the Cz – Cβ vector or the Cz – 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 Figure 1.

2.4 Outputs

We are aiming to predict absolute errors between predicted and measured structural properties for twelve predicted structural properties (\( \Delta SS, \Delta \phi, \Delta \psi, \Delta \theta, \Delta \tau, ARSA, \Delta CNz, \Delta CN\beta, \Delta HSE/\beta-up, \Delta HSE/z-down, \Delta HSE/z-up and \Delta HSE/z-down \)). \( \Delta SS \) is an error indicator of predicted secondary structure. \( \Delta SS = 1 \), if the predicted secondary structure is the same as the actual secondary structure according to three-state prediction, and 0 otherwise. \( \Delta \theta = |\phi^{\text{Pred}} - \phi^{\text{Expt}}| \). For rotational angles \( \phi, \psi \), and \( \tau \), the smaller value of \( |\text{Angle}^{\text{Pred}} - \text{Angle}^{\text{Expt}}| \) or \( 360^\circ - |\text{Angle}^{\text{Pred}} - \text{Angle}^{\text{Expt}}| \) is employed as the prediction target to account for the angle periodicity.

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.
ΔRSA is the absolute difference between predicted and actual relative solvent accessibility. Similarly, ΔCNx, ΔCN β, 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.

2.5 Training, test and performance evaluation
The neural network model was trained by 10-fold cross validation with TR4590 and independently tested by TS1199 and CASP11. In 10-fold cross validation, the training dataset was randomly divided into 10 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.

2.6 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(ΔV^actu, ΔV^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 \( E_{ij} = \log(P_{ij}/\Sigma N(i,j)) \) where \( P_{ij} = N(i,j)/\Sigma N(i,j) \). Then, for each model structure, we can obtain (1) \( \Delta V^\text{actu} \) from our method for each sequence position \( m \), (2) \( \Delta V^\text{pred} \) 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 to the (i, j) bins that \( \Delta V^\text{actu} \) and \( \Delta V^\text{pred} \) belong to and (4) the average of the neighboring energy scores with a window size of \( 7 \) (\( q_m = \sum E_{ij} \), \( k + m)/7 \), \( k = -3, -2, -1, 0, 1, 2, 3 \). 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.

3 Results
Table 1 summarizes the performance of error prediction by 10-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 0.52–0.56 for PCC and between 0.59–0.66 for SPC with low standard deviation between 10 folds. MAE values range from 4° for Δθ, 12° for Δψ, 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 ψ, 30° for ψ and 32° for τ). A much higher SPC than PCC values for all angle errors indicate non-linear 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 (ΔCNx and ΔCN β) 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 10-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., 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 10-fold cross-validation, 0.818 for TS1199 and 0.799 for CASP11. The ROC curves are shown in Supplementary Figure S1. Again, the performance for 10-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 100% of the sorted residues. As an example, results for Δθ and Δτ are shown in Figure 2. The average actual errors monotonically increase according to sorted predicted errors,

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

<table>
<thead>
<tr>
<th>Prediction Target</th>
<th>TR4590 (Ten-fold cross validation)</th>
<th>TS1199</th>
<th>CASP11</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCC</td>
<td>SPC</td>
<td>MAE</td>
</tr>
<tr>
<td>Δϕ</td>
<td>0.52 ± 0.00</td>
<td>0.59 ± 0.00</td>
<td>12.15 ± 0.14</td>
</tr>
<tr>
<td>Δψ</td>
<td>0.54 ± 0.00</td>
<td>0.61 ± 0.00</td>
<td>20.28 ± 0.31</td>
</tr>
<tr>
<td>Δθ</td>
<td>0.56 ± 0.00</td>
<td>0.62 ± 0.00</td>
<td>4.32 ± 0.05</td>
</tr>
<tr>
<td>Δτ</td>
<td>0.56 ± 0.01</td>
<td>0.66 ± 0.00</td>
<td>20.75 ± 0.32</td>
</tr>
<tr>
<td>ΔRSA</td>
<td>0.46 ± 0.00</td>
<td>0.48 ± 0.00</td>
<td>7.54 ± 0.07</td>
</tr>
<tr>
<td>ΔHSEα-up</td>
<td>0.35 ± 0.00</td>
<td>0.33 ± 0.00</td>
<td>2.68 ± 0.02</td>
</tr>
<tr>
<td>ΔHSEβ-up</td>
<td>0.35 ± 0.00</td>
<td>0.33 ± 0.00</td>
<td>2.61 ± 0.01</td>
</tr>
<tr>
<td>ΔCNx</td>
<td>0.30 ± 0.01</td>
<td>0.26 ± 0.01</td>
<td>3.29 ± 0.02</td>
</tr>
<tr>
<td>ΔCNβ</td>
<td>0.30 ± 0.01</td>
<td>0.26 ± 0.01</td>
<td>3.29 ± 0.03</td>
</tr>
<tr>
<td>ΔHSEα-down</td>
<td>0.29 ± 0.01</td>
<td>0.26 ± 0.01</td>
<td>2.23 ± 0.01</td>
</tr>
<tr>
<td>ΔHSEβ-down</td>
<td>0.26 ± 0.01</td>
<td>0.23 ± 0.01</td>
<td>2.10 ± 0.01</td>
</tr>
</tbody>
</table>

*PCC and SPC are Pearson and Spearman correlation coefficients between predicted and actual absolute errors, respectively. MAE: Mean absolute error between predicted and actual absolute differences.
consistent with strong positive correlation (PCC~0.5 and SPC~0.6) between predicted and actual errors for $\Delta \theta$ and $\Delta \tau$. 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 Figure 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 ($\Delta \theta$, $\Delta \phi$, RSA, $\Delta$CNs, $\Delta$CNb, HSE$b$-up, HSE$b$-down, HSE$x$-up and HSE$x$-down) shown in Supplementary Figure 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 (ASS) or according to predicted secondary structure probability by SPIDER2. As shown in Supplementary Figure S3, the performance of ASS is only marginally better than that of predicted secondary structure probability by SPIDER2 by identifying higher percent of accurately predicted residues.

Figure 3A and B shows 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 $\Delta \theta$ and $\Delta \tau$, respectively. Angle errors in helical residues are accurately predicted regardless the type of amino acid residue. The highest errors in $\Delta \theta$ and $\Delta \tau$ interestingly are glycine (G) in sheet conformation, suggesting the difficulty in pinpointing the flexible glycine (G) in the sheet conformation. The heat maps for $\Delta \phi$ and $\Delta \psi$ (Supplementary Fig. S4), however, show that the highest errors in $\Delta \phi$ and $\Delta \psi$ are glycine (G) and tryptophan (W) in coil conformation, respectively.

As an example, Figure 4 compared predicted and actual $C_\alpha$-based torsion angle $\Delta \tau$ for zinc protease from actinobacteria Streptomyces caesitius (PDB ID 1c7kA in TS1199). The structures are color-coded according to predicted and actual $\Delta \tau$ in Figure 4A and B, respectively. These $\Delta \tau$ values are also shown as a function of residue index in Figure 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 $\Delta \tau$ captures highly accurate (blue) areas reasonably well but with under-predicted $\Delta \tau$ 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 $\Delta \tau$ and between 0.33 and 0.44 for $\Delta$CNs (0.33), $\Delta$CNb (0.33), $\Delta$HSE$\downarrow$-down (0.35), $\Delta$HSE$\downarrow$-down (0.35), $\Delta \psi$ (0.37), $\Delta$HSE$b$-up (0.40), $\Delta$HSE$x$-up (0.40), $\Delta \theta$ (0.40), $\Delta \psi$ (0.44). The weakest correlation was observed for $\Delta$RSA (0.19). The statistically significant correlation ($P$-value $<$ $2.2 \times 10^{-16}$) for all predicted errors confirmed the usefulness of these variables as novel features for quality assessment.

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, $C_\alpha$-based angles and torsion angles,
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 Figure 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, Figure 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 Δr. 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

This work was supported by the Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP) grant 20130031120001 (to J.G.). 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


