Bioinformatics, 32(6), 2016, 843–849
doi: 10.1093/bioinformatics/btv665
Advance Access Publication Date: 14 November 2015
Original Paper

Structural bioinformatics

Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins

Rhys Heffernan[1], Abdollah Dehzangi[1,2,3], James Lyons[1], Kuldip Paliwal[1], Alok Sharma[2,4], Jihua Wang[5], Abdul Sattar[2,6], Yaoqi Zhou[5,7,]* and Yuedong Yang[7,]*

[1]Signal Processing Laboratory, School of Engineering, Griffith University, Brisbane, Australia, [2]Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, Australia, [3]Medical Research Center (MRC), Department of Psychiatry, University of Iowa, Iowa City, USA, [4]School of Engineering and Physics, University of the South Pacific, Private Mail Bag, Laucala Campus, Suva, Fiji, [5]Shandong Provincial Key Laboratory of Biophysics, Institute of Biophysics, Dezhou University, Dezhou, Shandong 253023, China, [6]National ICT Australia (NICTA), Brisbane, Australia and [7]Institute for Glycomics and School of Information and Communication Technique, Griffith University, Parklands Dr. Southport, QLD 4222, Australia

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on June 30, 2015; revised on August 26, 2015; accepted on November 7, 2015

Abstract

Motivation: Solvent exposure of amino acid residues of proteins plays an important role in understanding and predicting protein structure, function and interactions. Solvent exposure can be characterized by several measures including solvent accessible surface area (ASA), residue depth (RD) and contact numbers (CN). More recently, an orientation-dependent contact number called half-sphere exposure (HSE) was introduced by separating the contacts within upper and down half spheres defined according to the Cα-Cβ (HSEβ) vector or neighboring Cα-Cα vectors (HSEα). HSEα calculated from protein structures was found to better describe the solvent exposure over ASA, CN and RD in many applications. Thus, a sequence-based prediction is desirable, as most proteins do not have experimentally determined structures. To our best knowledge, there is no method to predict HSEα and only one method to predict HSEβ.

Results: This study developed a novel method for predicting both HSEα and HSEβ (SPIDER-HSE) that achieved a consistent performance for 10-fold cross validation and two independent tests. The correlation coefficients between predicted and measured HSEβ (0.73 for upper sphere, 0.69 for down sphere and 0.76 for contact numbers) for the independent test set of 1199 proteins are significantly higher than existing methods. Moreover, predicted HSEα has a higher correlation coefficient (0.46) to the stability change by residue mutants than predicted HSEβ (0.37) and ASA (0.43). The results, together with its easy Cα-atom-based calculation, highlight the potential usefulness of predicted HSEα for protein structure prediction and refinement as well as function prediction.

Availability and implementation: The method is available at http://sparks-lab.org.

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

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

1 Introduction

Measuring exposure of amino acid residues of proteins to solvent is important for understanding and predicting protein structure, function and interactions (Gilis and Rooman, 1997; Lee and Richards, 1971; Rost and Sander, 1994; Tuncbag *et al.*, 2009). The most common measure is solvent accessible surface area (ASA) (Connolly, 1983) with applications ranging from protein structure prediction (Bennett-Lovsey *et al.*, 2008; Wu and Zhang, 2008; Yang *et al.*, 2011) to protein function prediction (Lou *et al.*, 2014; Zhang *et al.*, 2010; Zhang *et al.*, 2008; Zhao *et al.*, 2013; Zhao *et al.*, 2011). The inability of ASA to discriminate deeply buried residues from those buried just beneath the surface leads to development of residue depth (Chakravarty and Varadarajan, 1999) by averaging the distance to the nearest atom on the protein surface. Both ASA and residue depth, however, require precise evaluation of the protein surface in full atomic details that is time consuming. Having a time-consuming calculation and requiring an all-atomic model limited the usefulness of ASA and residue depth in the *ab initio* prediction of protein structure where protein conformation is often represented by main-chain atoms (Bradley *et al.*, 2003; Faraggi *et al.*, 2009) or only Cα atoms (Kihara *et al.*, 2001; Yang and Liu, 2006) in initial conformational sampling.

Another measure of residue exposure to solvent is contact number (CN), which counts the number of residues within specific distance cut offs. The distances are based on the positions of Cα or Cβ atoms. Thus, unlike ASA and residue depth, only a coarse-grained model is needed for evaluating CNs. Different distance cut offs have been used in earlier studies (Pollastri *et al.*, 2002; Yuan, 2005). It was shown that CN with a distance cut off between 12 and 14 Å is the best for protein fold recognition (Karchin *et al.*, 2004).

All above-mentioned solvent-exposure measures, however, do not contain explicit information regarding the orientation of side chains that are important for functional and structural studies. Hamelryck designed a new measure by splitting the sphere around the Cα atom into two half spheres along the vector of Cα-Cβ atoms (Hamelryck, 2005). The half sphere containing the Cβ atom was defined as upper and the other as down half spheres. The numbers of Cα atoms enclosed in these two half-spheres were named as Half-Sphere Exposure (HSE)-up and HSE-down, respectively. In addition, he has substituted the vector Cα-Cβ with a pseudo vector generated from the sum of vectors $C\alpha_{i-1}$-$C\alpha_i$ and $C\alpha_{i+1}$-$C\alpha_i$. This HSE is denoted as HSEα (-up and -down) to distinguish from the HSE calculated based on the Cα-Cβ vector (here and hereafter, it will be annotated as HSEβ). HSEα does not require the positions of Cβ-atoms. One advantage of HSE is that its value is independent of amino acid type because it describes a residue's coordination environment rather than a quantity related to its size such as ASA and RD. Interestingly, HSEα-up outperforms other solvent exposure measures including CN, ASA, relative ASA, residue depth and the other three HSE definitions (HSEβ-up, HSEβ-down and HSEα-down) in correlation to changes in protein stability due to mutations and to conservation of amino acid residues (Hamelryck, 2005). More recently, HSE was found to be better than ASA for prediction of B cell epitopes of proteins from their three dimensional structures (Kringelum *et al.*, 2012). HSE has also been found useful in many other applications (Franzosa and Xia, 2009; Kringelum *et al.*, 2012; Sweredoski and Baldi, 2008). Most of these applications obtained HSE based on known protein structures. Because the structures for most proteins are not known experimentally, a sequence-based prediction is desirable.

Many sequence-based methods were developed for predicting ASA (Adamczak *et al.*, 2004; Ahmad *et al.*, 2003; Cheng *et al.*, 2005; Dor and Zhou, 2007; Garg *et al.*, 2005; Yuan and Huang, 2004) and CN (Kinjo and Nishikawa, 2006; Pollastri *et al.*, 2002; Yuan, 2005). However, there is no method available for prediction of HSEα and only one (HSEpred) for the prediction of HSEβ (Song *et al.*, 2008). We found that the correlation coefficients between actual HSEβ and those predicted by HSEpred (up and down, respectively) are less than 0.43 for our dataset of 1199 proteins. One possible factor is that HSEpred was trained on a small dataset of 632 proteins and was not tested on independent datasets. Therefore, a more accurate method is clearly needed.

Recently, we developed a method called SPIDER 2 (Heffernan *et al.*, 2015) that utilized predicted secondary structures, backbone torsion angles and ASA, iteratively, in order to improve their accuracies. The method achieved 82% accuracy for secondary structure prediction, 0.76 for the correlation coefficient between predicted and actual solvent accessible surface area, 19° and 30° for mean absolute errors of backbone φ and ψ angles, respectively, and 8° and 32° for mean absolute errors of Cα-based θ and τ angles, respectively, for an independent test dataset of 1199 proteins. Here, we expand the iterative technique to the prediction of HSEα (-up and -down), HSEβ (-up and -down) and CN by employing a large dataset containing 4590 protein chains. The method was independently tested in a dataset of 1199 proteins and a dataset of 69 proteins from the Critical Assessment of Structure Prediction technique (CASP 11, 2014). As a result, highly accurate and robust prediction was obtained (e.g. a correlation coefficient of 0.73 for HSEβ-up, 0.69 for HSEβ-down and 0.76 for contact number on the independent test set). Comparison to two previous methods for contact prediction and HSEβ confirmed the superior performance of the method obtained. The HSEα and HSEβ predictors are incorporated as a package in SPIDER 2 available at http://sparks-lab.org or downloadable as a standalone package.

2 Methods

2.1 Datasets

We have employed the same dataset as used in our previous study (Lyons *et al.*, 2014), which consists of 5789 proteins (1 246 420 residues). This dataset was generated by the sequence culling server PISCES (< 25% pairwise sequence identity and <2.0 Å resolution) (Wang and Dunbrack, 2003). From this dataset, 4590 proteins were randomly selected as training set (TR4590) and the remaining 1199 proteins were utilized as an independent test set (TS1199). In addition, we downloaded the targets from critical assessment of structure prediction technique (CASP 11, 2014, http://www.predictioncenter.org/casp11/). After removing redundant sequences (30% in between or to the training set), we obtained a set of 69 proteins (CASP11) out of original 99 proteins. This set contains 18 617 residues.

2.2 Input features

For each amino acid, we have extracted 69 input features. The first 20 features are substitution probabilities of amino acids from the PSSM matrix. The PSSM is generated by PSI-BLAST (Altschul *et al.*, 1997) with three iterations of searching against 90% non-redundant protein database (downloaded from ftp://toolkit.genzentrum.lmu.de/pub/HH-suite/databases/). The next 30 features are extracted from the HMM-profile generated by HHblits with default parameters against Uniprot20 protein database (Remmert *et al.*, 2011). The HMM profile includes 10 transition probabilities between matching, insertion and deletion states in addition to 20 substitution probabilities. The additional 12 features represent predicted structural properties of amino acids by SPIDER 2 that include predicted probabilities in three secondary structure states (Helix, Sheet and Coil), ASA, and sine and cosine values of the main chain torsional angles (φ and ψ) and main chain angles between Cα atoms (θ and τ). Here, sine and cosine of angles were employed to remove the effect of angle periodicity (Lyons *et al.*, 2014). The last seven features are physicochemical representatives (PP7) of twenty standard amino acids, namely, a steric parameter (graph shape index), hydrophobicity, volume, polarizability, isoelectric point, helix probability and sheet probability. Additionally, we utilized a window size of 17 amino acids (8 amino acids at each side of the target amino acid). Together with a global feature of the protein length, this led to a total of 1174 input features ($17 \times 69 + 1 = 1174$). The window size was taken from SPIDER 2 (Heffernan *et al.*, 2015) without further optimization.

2.3 Parallel multiple-step iterative deep neural-network learning

Here, we implemented the same learning scheme as previously used in SPIDER 2. As shown in Figure 1, the features were input into the deep learning neural network to obtain an initial prediction of HSEs (HSE-up, HSE-down and CN). The deep artificial neural network consists of three hidden layers, each with 150 nodes. The input data was normalized to the range of 0–1. The weights were initialized by unsupervised learning from stacked sparse auto-encoders, and then refined by using standard back propagation. In this study, we employed the deep neural network from the MATLAB toolbox, implemented by Palm (2012). Deep neural networks have been widely implemented in the prediction of protein structure (Nguyen *et al.*, 2014). In the second iteration, the predicted values from the first iteration were added into the input features to predict HSEs. This process iterated with updated predicted values until convergence. We found that the prediction accuracy did not increase after two iterations. The training process was performed separately for HSEα (HSEα-up, HSEα-down and CN) and HSEβ (HSEβ-up, HSEβ-down and CN). All contacts are defined with 13 Å distance cut off.

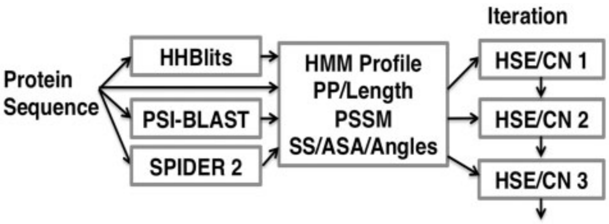

Fig. 1. The general flowchart for the multiple-step iterative algorithm implemented for training of SPIDER-HSE

2.4 Evaluation method

The method was first examined by 10-fold cross validation where the training set TR4590 was randomly divided into 10-folds. Nine folds were used in turn for training and the remaining one for testing until all 10-folds were tested. As SPIDER 2 has been trained in the same training set, we avoided over-training by following the same 10-folds and utilizing the 10-fold cross validation results during the training of SPIDER 2 as input features. Moreover, we tested our method in the independent test sets TS1199 and CASP11 dataset by using TR4590 as the training set. The prediction performance of CN, HSEα and HSEβ was measured by Pearson correlation coefficient (PCC) as used in previous studies (Song *et al.*, 2008; Yuan, 2005).

3 Results and discussions

3.1 Overall prediction performance

Figure 2 and Table 1 show the results of 10-fold cross validation and independent test in the first four iterations for prediction of CN, HSEβ-up and HSEβ-down. For the independent test, PCCs achieve the highest value at the second iteration with 0.734, 0.693 and 0.756 for HSEβ-up, HSEβ-down and CN, respectively. The same is true for 10-fold cross valuation. Thus additional iterations are unnecessary. We also noted that the correlation coefficients from the 10-fold cross validation and from the independent test set are essentially the same. For example, PCCs for CN at the second iteration are 0.757 and 0.756 for 10-fold cross validation and independent test, respectively. PCC for HSEβ up at the second iteration are 0.733 and 0.734 for 10-fold cross validation and independent test, respectively. High consistency between 10-fold cross validation and independent test indicates the robustness of our trained method. The small standard deviations between ten subsets from 10-fold cross validation further confirm a stable performance.

We further found that the accuracy for HSEα is similar to that of HSEβ. For example, The PCCs of HSEα-up and HSEβ-up at the second iteration for independent test are 0.729 and 0.734, respectively. The PCCs of HSEα-down and HSEβ-down at the second iteration for independent test are 0.717 and 0.693, respectively. Because of similarity between HSEα and HSEβ, here and hereafter, we will only present the results for HSEβ from the second iteration based on the independent test unless specifically mentioned.

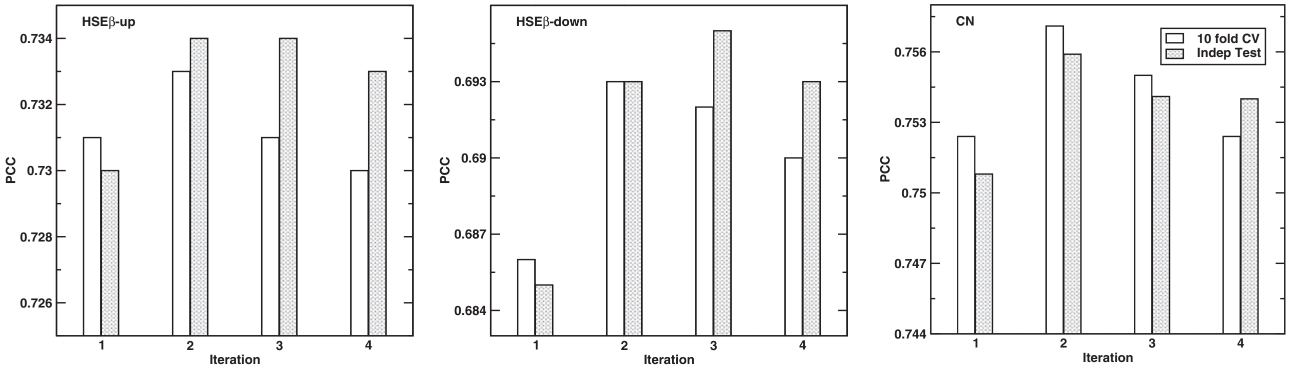

Fig. 2. The Pearson Correlation Coefficients achieved for HSEβ-up, HSEβ-down and contact number by 10-fold cross validation (open bar) and on the independent dataset TS1199 (filled bar) in different iterations

3.2 The contribution from each feature group

It is of interest to know the contribution of each feature group to the overall prediction performance of SPIDER-HSE. Table 2 compares PCCs by utilizing a single feature group or by removing them individually from SPIDER-HSE. In the first test, predicted ASA is the single best feature as it achieved the highest PCC of 0.72, 0.69 and 0.64 for CN, HSE-up and HSE-down, respectively. This is understandable because ASA and CN both describe the level of solvent exposure. Such high correlation confirms the suitability to substitute ASA by the simpler representation of solvent exposure by HSE. The performance of two conservation profiles HMM and PSSM is only slightly worse than ASA with HMM marginally better than PSSM. The physical parameters alone without sequence conservation information can make a decent prediction with PCC around 0.5 for CN and HSE-up.

Table 1. Pearson Correlation Coefficients of CN, HSEβ-up and HSEβ-down in iterations for 10-fold cross validation and independent test set TS1199

Parameter	Dataset	1	2	3	4
HSEβ-up	10-fold[a]	0.731	0.733	0.731	0.730
	SD[b]	0.009	0.008	0.008	0.008
	Test	0.730	0.734	0.734	0.733
HSEβ-down	10-fold[a]	0.686	0.693	0.692	0.690
	SD[b]	0.007	0.008	0.008	0.008
	Test	0.685	0.693	0.695	0.694
CN	10-fold[a]	0.752	0.757	0.755	0.752
	SD[b]	0.010	0.011	0.011	0.011
	Test	0.751	0.756	0.754	0.754

[a] The average and [b] the standard deviation of PCCs from the 10-fold cross validation.

One interesting observation is that predicted angles are more useful than predicted secondary structure in predicting HSE-up (0.58 versus 0.50) and HSE-down (0.54 versus 0.48) although they contribute similarly in predicting CN (0.584 versus 0.586). This suggests that continuous main-chain torsion angles have more orientation information than discrete states of secondary structure.

Another interesting observation is that protein length itself as a single feature is not very useful in predicting HSE. Its combination with other features, however, proved useful. Removing protein length will reduce the correlation coefficient in the independent test from 0.756 for CN to 0.740. This is the largest reduction, compared to removing other feature groups. The relatively small change by removing other feature groups is partly due to high redundancies between these feature groups.

For example, HHM and PSSM describe residue conservation during evolution, whereas secondary structure and main-chain angles both represent the main-chain conformations. In addition, ASA, SS and main-chain angles have been obtained from PSSM and PP7 feature groups for their prediction. Nonetheless, a drop in performance by removing any single feature group indicates usefulness of all these features for the overall performance.

3.3 Comparison to previous methods

To compare our results with previously reported methods, we reproduced the results of HSEpred by both its locally running version and online server (http://sunflower.kuicr.kyoto-u.ac.jp/∼sjn/hse/links.html) on test sets TS1199 and CASP 11. In addition, we compared with a contact number prediction method CRNpred (Kinjo and Nishikawa, 2006) (http://ipr.pdbj.org/crnpred/). As shown in Table 3, the PCC values for HSE are less than 0.5, compared to 0.7 from our method for both test sets. Our predicted CN (PCC = 0.76 for the independent test set) are also significantly more accurate than either HSEpred (PCC = 0.56) or CRNpred (PCC = 0.70). A similar

Table 2. Pearson correlation coefficients (PCCs) of predicted CN, HSEβ-up and HSEβ-down by using single feature group or by removing it from SPIDER-HSE for independent test set TS1199

	CN	HSEβ-up	HSEβ-down		CN	HSEβ-up	HSEβ-down
–	–	–	–	SP-HSE[b]	0.756	0.734	0.693
ASA[a]	0.721	0.689	0.637	−ASA[c]	0.746	0.728	0.680
HMM	0.709	0.682	0.620	−HHM	0.742	0.724	0.676
PSSM	0.694	0.683	0.607	−PSSM	0.751	0.730	0.684
SS	0.586	0.495	0.477	−SS	0.751	0.729	0.685
Angles	0.584	0.584	0.536	−Angles	0.752	0.729	0.681
PP7	0.523	0.520	0.436	−PP7	0.748	0.729	0.683
Length	0.013	0.006	0.006	−Length	0.740	0.725	0.679

[a]Predicting by using only individual feature group.
[b]SPIDER2-HSE by using full features.
[c]Predicting by excluding one feature group.

Table 3. Comparison to HSEpred and CRNpred in the independent dataset TS1199 and CASP11

Methods	TS1199			CASP11		
	CN	HSEβ-up	HSEβ-down	CN	HSEβ-up	HSEβ-down
CRNpred	0.697 [a]	–	–	0.691 [a]	–	–
HSEpred (Local)	0.624	0.490	0.398	0.590	0.467	0.394
HSEpred (Online)	0.555	0.429	0.326	0.527	0.427	0.331
1st Iteration (This work)	0.751	0.7301	0.685	0.766	0.749	0.692
2nd Iteration (This work)	0.756	0.7343	0.693	0.770	0.751	0.699

[a] The contact number based on 12 Å.

result was obtained for the CASP 11 set. CRNpred was trained with a contact defined by a cut off 12 Å, whereas we have used 13 Å as a cut off. To be consistent with CRNpred, we specially trained our method based on 12 Å cut off. The accuracy of our method is unchanged. This is not surprising as the correlation coefficient between the contact number from the cut off of 12 Å generated from protein structures is highly correlated to the contact number from the cut off of 13 Å (PCC = 0.976).

3.4 Correlation of predicted HSE to the stabilities of mutants

To examine the usefulness of sequence-based prediction of HSE, we compared predicted HSE to the stabilities of mutants. Hamelryck found that HSE from protein structures strongly correlates with experimentally measured stability changes due to mutation (Hamelryck, 2005). Here, we expanded the stability dataset by using protherm database recently updated in 2013 (Kumar *et al.*, 2006). As with Hamelryck, we limited point mutants from VAL/ILE/LEU to ALA. These three residues were selected because mutations from small hydrophobic residues to ALA will not cause significant changes in polar interaction or in global conformations. Therefore, the protein stability change is dominated by the change of solvent accessibility, and thus eligible for evaluating solvent exposure measures. A total of 220 mutants were found after removing two outliers with ΔΔG above 3.0 kJ/mol (details listed in Supplementary Materials).

As shown in Table 4, HSEα-up, HSEβ-up, CN and ASA calculated from experimental structures consistently have strong correlation to ΔΔG (negative correlation for HSEα-up, HSEβ-up and CN with PCC = −0.595, −0.541 and −0.494 respectively, positive correlation for ASA with PCC = 0.538). Negative correlation for CN and HSE-up is because CN is negatively correlated to ASA (i.e. more contacts mean less solvent accessible). Thus, HSEα-up has the best correlation, which confirmed the result of the previous study (Hamelryck, 2005) with a larger dataset. For predicted CN, HSE and ASA, the correlations to experimental ΔΔG become weaker. Predicted HSEα-up has the best correlation with PCC = −0.461.

Table 4. Pearson correlation coefficients of HSE and ASA with (-ΔΔG) for 220 ILE/LEU/VAL to ALA mutants

	CN	HSEβ-up	HSEβ-down	HSEα-up	HSEα-down	ASA[a]
Experimental	0.494	0.541	0.088	0.595	−0.044	0.538
Predicted	0.322	0.373	0.039	0.461	−0.040	0.438

[a]The PCC was calculated with (-ΔΔG) for all measures except ASA in order to have a positive value.

One interesting observation is that HSE-down has no correlation to ΔΔG. Similar results were obtained earlier (Hamelryck, 2005). This is consistent with our physical intuition that the contacts with the front of the side-chain of an amino acid residue contribute most to the interaction of this residue to other amino acid residues. Although the correlation, as shown in Figure 3, between HSEα-up and ΔΔG is only slightly stronger (PCC = −0.461) than that between ASA and ΔΔG (PCC = 0.438), the much simpler calculation of HSEα-up than that of ASA will make it more useful in structure prediction. The fact that HSEα-up (PCC = −0.46) correlates better than HSEβ-up to ΔΔG (PCC = −0.37) indicates the importance of a separate predictor for HSEα developed here.

4 Conclusions

This work has developed the first sequence-based method for predicting HSEα, in addition to prediction of HSEβ and contact number. Trained by a large dataset of >4000 proteins and independently tested by two separate datasets, we showed that our predictions of HSEβ and contact numbers are significantly more accurate than existing methods (HSEpred and CRNpred) with correlation coefficients between predicted and actual number at about 0.7 for 10-fold cross validation and independent tests. This highly accurate prediction was built on highly accurate prediction of secondary structure, backbone angles and solvent accessible surface area by our previous method SPIDER 2. Another contribution was from the established iterative deep learning scheme. The deep neural network allows us to train the server on a dataset that is >7 times larger than the previous method HSEpred, and thus avoids potential over-training of the predictor. In addition, DNN was found to be much faster to

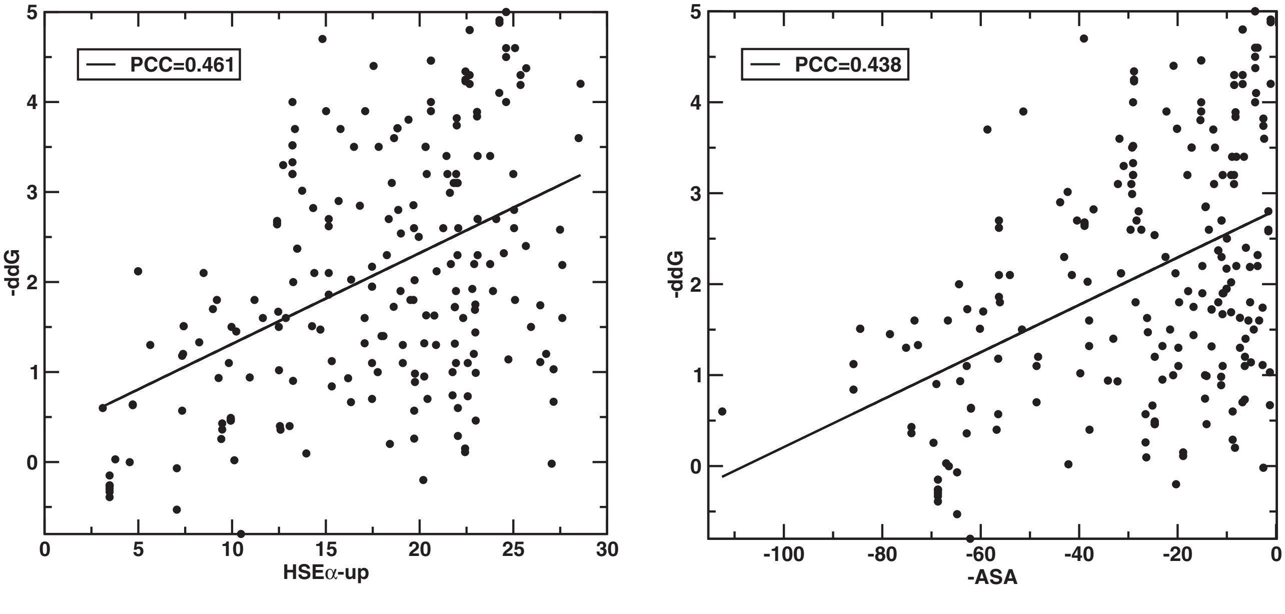

Fig. 3. The predicted HSEα-up and ASA versus ΔΔG. The line is the linear least-square fitting to the data

converge with better performance than SVM and random forest methods in training big datasets with a large number of features (Bengio and LeCun, 2007; Schmidhuber, 2015). The usefulness of predicted HSEα is demonstrated by its improved correlation to experimentally measured stability change due to mutation, over predicted ASA.

Fast calculation of HSEα that requires the positions of Cα atoms only makes it an ideal sequence-specific restraint for coarse-grained modeling, protein structure prediction and refinement. The HSEα and HSEβ predictors are incorporated as a package in SPIDER 2 available at http://sparks-lab.org. To speed up calculations, we provide another version by using all features except HMM profile. This version leads to a slight reduction in PCCs but cuts the total running time by half because it requires to prepare only one of the two sequence profiles that are the most time-consuming.

Acknowledgements

We also 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.

Funding

This work was supported in part by National Health and Medical Research Council (1059775 and 1083450) of Australia and Australian Research Council's Linkage Infrastructure, Equipment and Facilities funding scheme (project number LE150100161) to Y.Z., National Natural Science Foundation of China (61271378) to J.W. and Y.Y., the Taishan Scholar Program of Shandong to Y.Z., and the Microsoft Azure for Research Awarded to Y. Y. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

Conflict of Interest: none declared.

References

Adamczak,R. *et al.* (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*, **56**, 753–767.

Ahmad,S. *et al.* (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, **50**, 629–635.

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.

Bengio,Y. and LeCun,Y. (2007) Scaling learning algorithms towards AI. *Large-Scale Kernel Mach.*, **34**, 321–359.

Bennett-Lovsey,R.M. *et al.* (2008) Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins*, **70**, 611–625.

Bradley,P. *et al.* (2003) Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins*, **53**, 457–468.

Chakravarty,S. and Varadarajan,R. (1999) Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure*, **7**, 723–732.

Cheng,J. *et al.* (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **33**, W72–W76.

Connolly,M.L. (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science*, **221**, 709–713.

Dor,O. and Zhou,Y. (2007) Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. *Proteins: Struct. Funct. Bioinf.*, **68**, 76–81.

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.

Franzosa,E.A. and Xia,Y. (2009) Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.*, **26**, 2387–2395.

Garg,A. *et al.* (2005) Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins*, **61**, 318–324.

Gilis,D. and Rooman,M. (1997) Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.*, **272**, 276–290.

Hamelryck,T. (2005) An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins: Struct. Funct. Bioinf.*, **59**, 38–48.

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.*, **5**, 11476.

Karchin,R. *et al.* (2004) Evaluation of local structure alphabets based on residue burial. *Proteins: Struct. Funct. Bioinf.*, **55**, 508–518.

Kihara,D. *et al.* (2001) TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 10125–10130.

Kinjo,A.R. and Nishikawa,K. (2006) CRNPRED: highly accurate prediction of one-dimensional protein structures by large-scale critical random networks. *BMC Bioinformatics*, **7**, 401.

Kringelum,J.V. *et al.* (2012) Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput. Biol.*, **8**, e1002829.

Kumar,M.S. *et al.* (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.

Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.

Lou,W. *et al.* (2014) Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naïve Bayes. *PloS One*, **9**, e86703.

Lyons,J. *et al.* (2014) Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *J. Comput. Chem.*, **35**, 2040–2046.

Nguyen,S.P. *et al.* (2014) DL-PRO: A novel deep learning method for protein model quality assessment. In: Neural Networks (IJCNN), 2014 International Joint Conference *on*. IEEE, pp. 2071–2078.

Palm,R.B. (2012) Prediction as a candidate for learning deep hierarchical models of data. Master Thesis, *Technical University of Denmark*, *Palm*, 24–25.

Pollastri,G. *et al.* (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins: Struct. Funct. Bioinf.*, **47**, 142–153.

Remmert,M. *et al.* (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins: Struct. Funct. Bioinf.*, **20**, 216–226.

Schmidhuber,J. (2015) Deep learning in neural networks: an overview. *Neural Networks*, **61**, 85–117.

Song,J. *et al.* (2008) HSEpred: predict half-sphere exposure from protein sequences. *Bioinformatics*, **24**, 1489–1497.

Sweredoski,M.J. and Baldi,P. (2008) PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. *Bioinformatics*, **24**, 1459–1460.

Tuncbag,N. *et al.* (2009) Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, **25**, 1513–1520.

Wang,G. and Dunbrack,R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wu,S. and Zhang,Y. (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*, **72**, 547–556.

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. and Liu,H. (2006) Genetic algorithms for protein conformation sampling and optimization in a discrete backbone dihedral angle space. *J. Comput. Chem.*, **27**, 1593–1602.

Yuan,Z. (2005) Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics*, **6**, 248.

Yuan,Z. and Huang,B. (2004) Prediction of protein accessible surface areas by support vector regression. *Proteins*, **57**, 558–564.

Zhang,T. *et al*. (2010) Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr. Protein Peptide Sci*., **11**, 609–628.

Zhang,T. *et al*. (2008) Accurate sequence-based prediction of catalytic residues. *Bioinformatics*, **24**, 2329–2338.

Zhao,H. *et al*. (2013) DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biol*., **14**, R23.

Zhao,H. *et al*. (2011) Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA Biol*., **8**, 988–996.