Protein domain-domain interaction prediction via deep neural networks.

Tugce Oruc, Christopher Thomas, Peter Winn

University of Birmingham, School of Biosciences

ABSTRACT

Proteins are one of the building blocks of life, and their structures and functions maintain most of the cellular processes. Determination of protein structures is not only critical for understanding its working mechanism but also vital for protein engineering and drug design. Although many experimental approaches exist to reveal the structures of proteins, limitations of experimental methods led researchers to develop computational approaches to determine structures. Implementation of machine learning algorithms provided great improvements in the protein structure prediction area for small and medium-sized proteins. On the other hand, for large proteins, determination of the structure of the overall protein complex remains a big challenge. One common approach for determination of large protein complexes is to determine the structures of protein subunits (i.e. domains) individually and arrange their positions and orientations correctly. For this purpose, we used convolutional neural networks to predict the distance potentials between the monomers of the target domain pairs. Successful distance potential predictions allowed us to generate correct interfaces between the domain pairs. This method will help to determine the structures of large, multi-domain protein complexes that can result in understanding their function better and lead to design successful experiments for protein engineering.

METHODS

For the generation of features, sequences of two domains were concatenated with respect to their order in the chain. Based on the concatenated target sequence, homologous sequences were searched to generate multiple sequence alignments (MSAs).

The following features were fed into the neural network: CCMpred results as coevolution prediction (DCA), the dot product of CCMpred predictions (DCAdot), mutual information, normalized mutual information, statistical potential, secondary structure predictions, predicted accessible surface area, statistical coupling analysis (SCA) matrix.

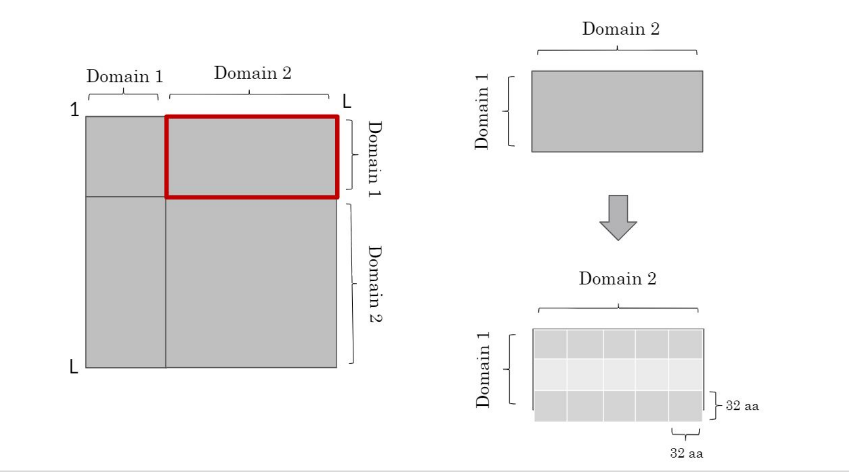


Figure 1. Feature matrix generation for neural networks. From the overall matrix, the intersection region of the first domain and the second domain is extracted and further divided into square matrices with a length of 32 amino acids (aa).

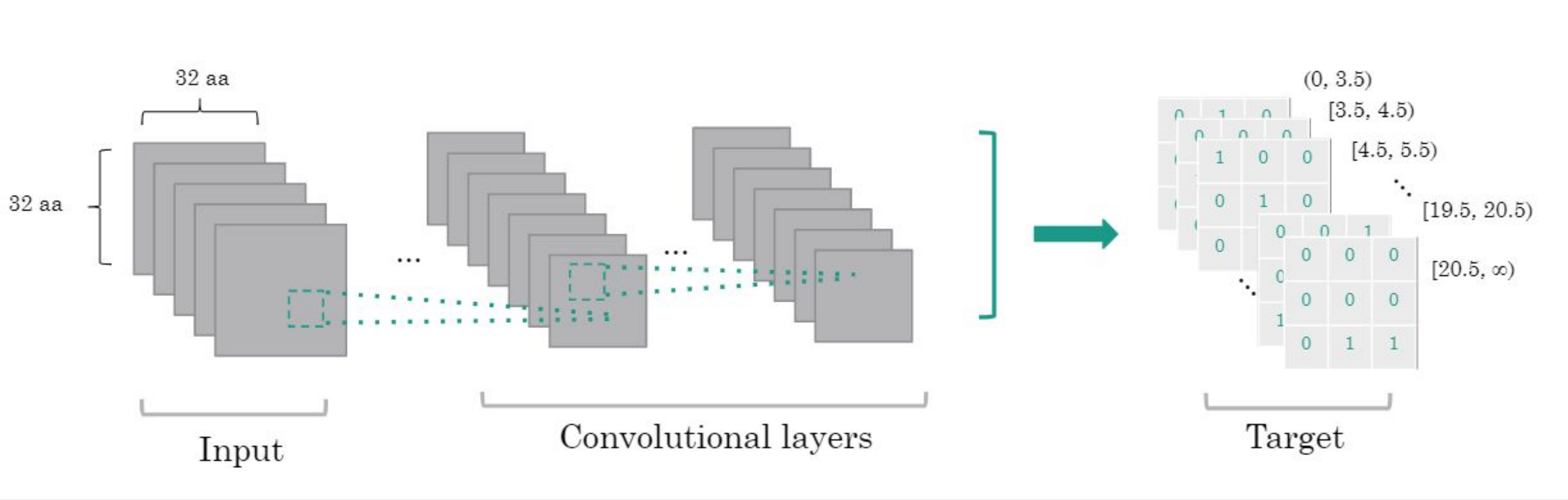


Figure 2. Neural network architecture for residue pair distance prediction. Convolutional neural networks were used to train models. As output matrix, 1 Å interval bins were used. For all residue pairs in the target matrix, the real distance bin is marked with 1 and the other layers are marked as 0. For one feature set, network was trained for six (or ten) times.

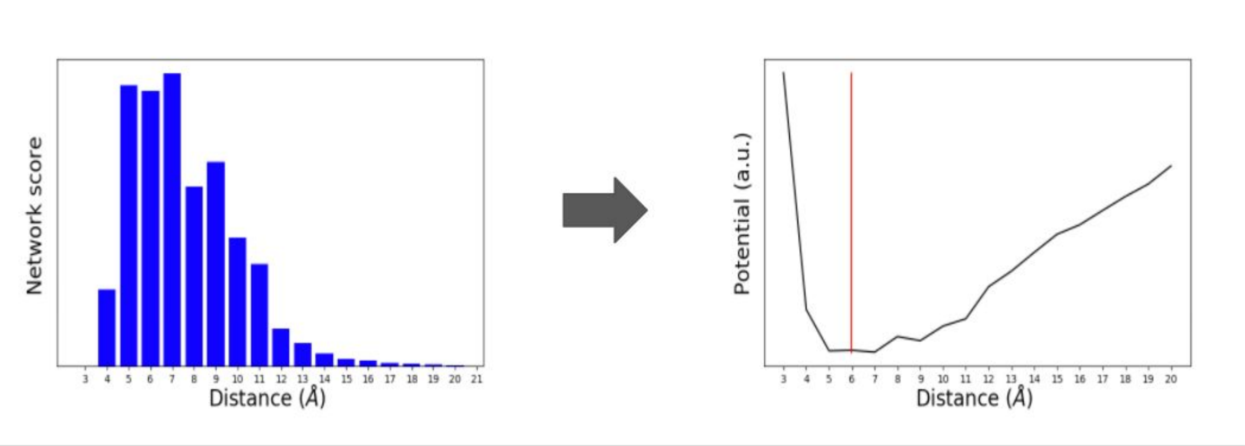
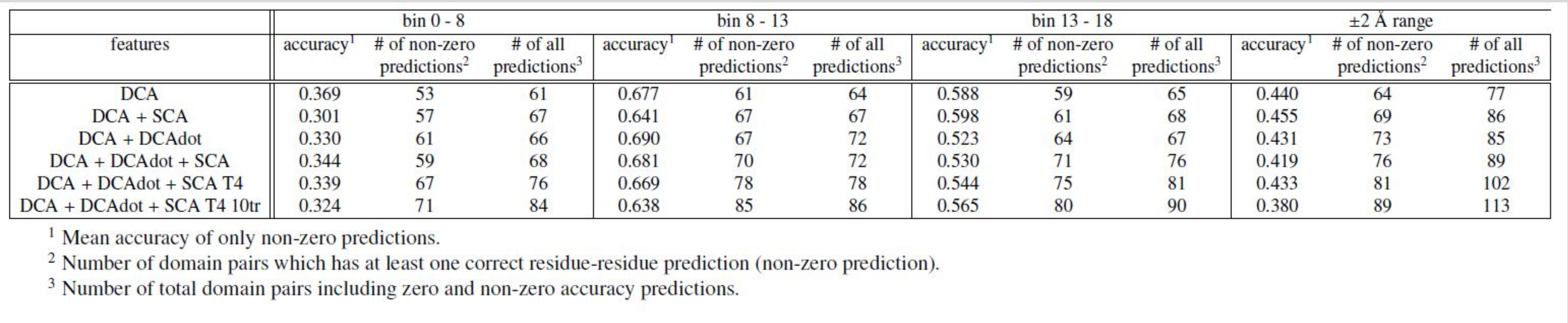


Figure 3. Distance potentials were calculated from network score distribution. For a predicted residue pair, network score distribution was converted into a distance potential by calculation of negative log-likelihood.

Distance potentials were applied as constraints on domain pairs as spline function to predict the correct interface with Rosetta docking program.

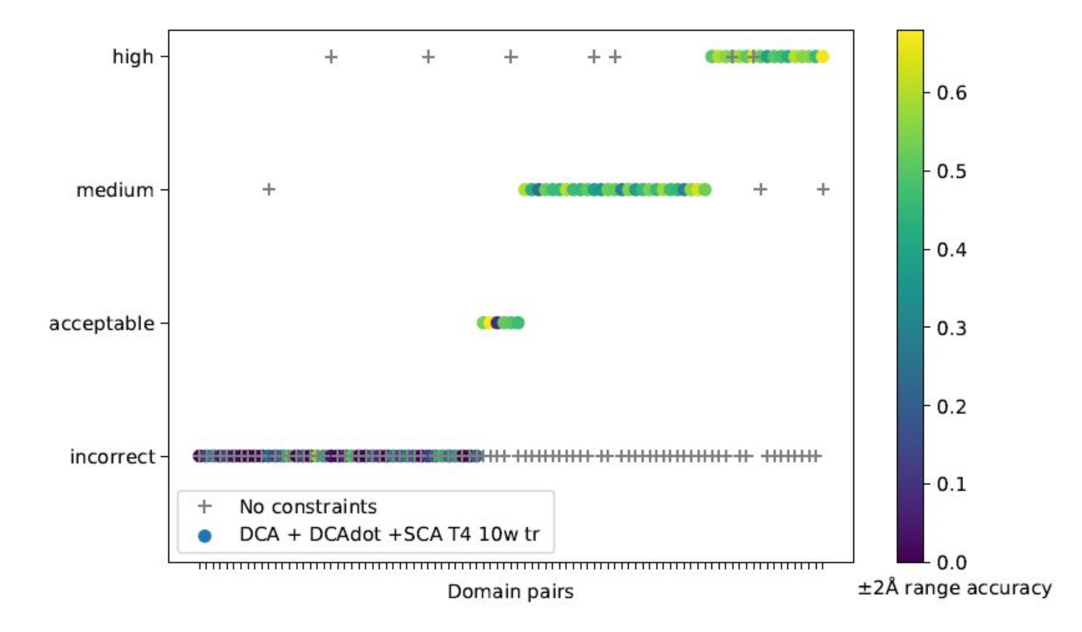
RESULTS AND DISCUSSION



features	bin 0 - 8 accuracy[1]	bin 0 - 8 # of non-zero predictions[2]	bin 0 - 8 # of all predictions[3]	bin 8 - 13 accuracy[1]	bin 8 - 13 # of non-zero predictions[2]	bin 8 - 13 # of all predictions[3]	bin 13 - 18 accuracy[1]	bin 13 - 18 # of non-zero predictions[2]	bin 13 - 18 # of all predictions[3]	±2 Å range accuracy[1]	±2 Å range # of non-zero predictions[2]	±2 Å range # of all predictions[3]
DCA	0.369	53	61	0.677	61	64	0.588	59	65	0.440	64	77
DCA + SCA	0.301	57	67	0.641	67	67	0.598	61	68	0.455	69	86
DCA + DCAdot	0.330	61	66	0.690	67	72	0.523	64	67	0.431	73	85
DCA + DCAdot + SCA	0.344	59	68	0.681	70	72	0.530	71	76	0.419	76	89
DCA + DCAdot + SCA T4	0.339	67	76	0.669	78	78	0.544	75	81	0.433	81	102
DCA + DCAdot + SCA T4 10tr	0.324	71	84	0.638	85	86	0.565	80	90	0.380	89	113

[1] Mean accuracy of only non-zero predictions.
[2] Number of domain pairs which has at least one correct residue-residue prediction (non-zero prediction).
[3] Number of total domain pairs including zero and non-zero accuracy predictions.

Table 1. Accuracies of predictions on the validation domain pairs with different feature sets. Different feature sets has been tested on validation set to determine the best input features. Among them DCA + DCAdot + SCA T4 10tr set provided good accuracy with better generalization. T4: chopping the matrix into 32 by 32 matrices from three additional different boundaries. 10tr: 10 trainings.



Models trained with DCA + DCAdot + SCA T4 10tr feature set was used to predict distance potentials and predicted potentials were introduced as constraints on Rosetta. For each domain pair, 4500 structures were generated and the structure with the lowest Rosetta energy was selected as the final model

Figure 4. Quality evaluation of predicted domain interfaces of the test set proteins with and without constraints. When constraints were not applied. only ten interfaces could be predicted correctly; whereas, 50 interfaces were predicted correctly when predicted distance potentials were implemented as constraints. For the domain pairs with higher ±2 Å range accuracies, interfaces were predicted with at least acceptable quality. Domain interface quality was determined based on CAPRI protein docking competition criteria.

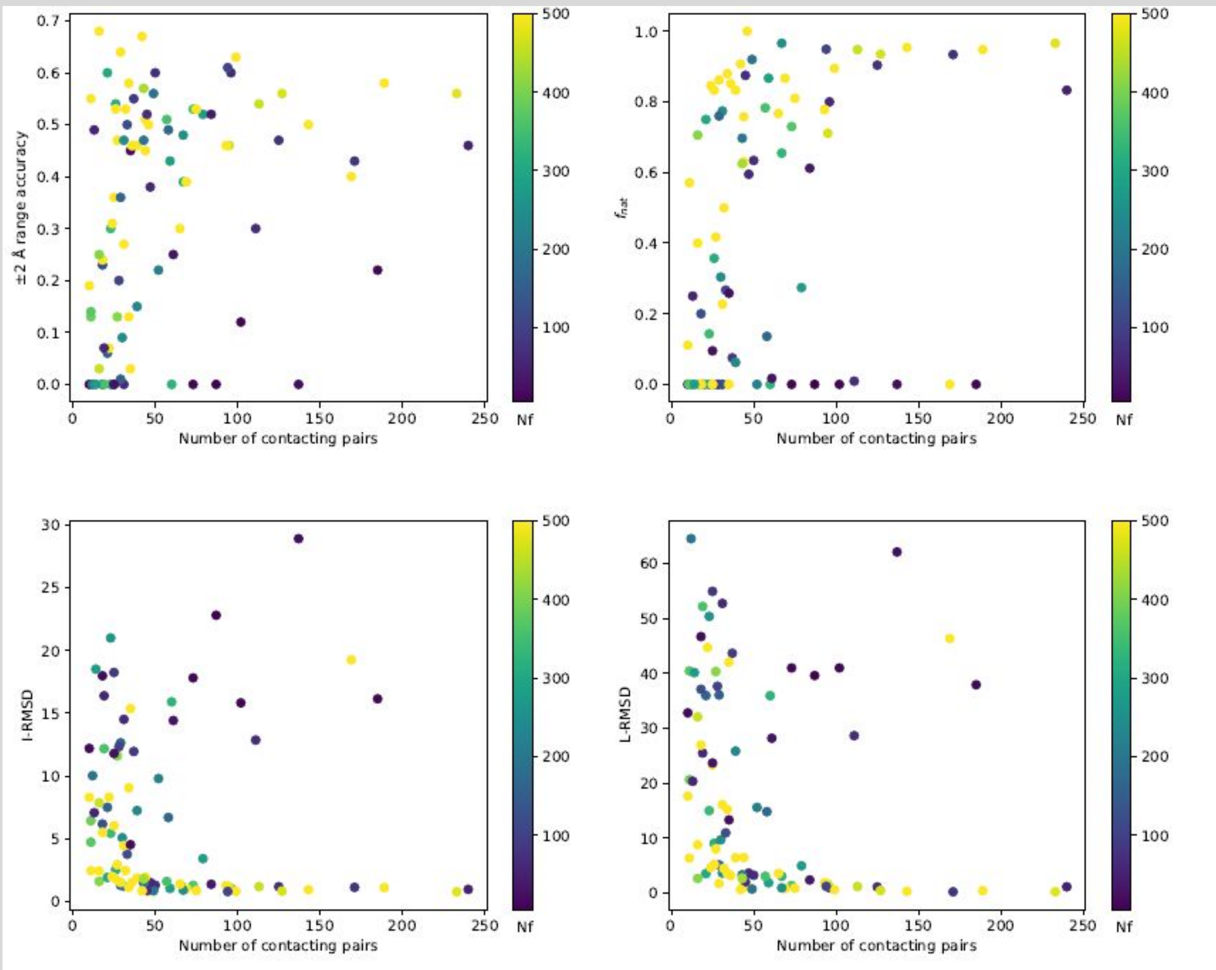


Figure 5. Prediction success increases as the number of the real contacts (domain interface surface) and Nf values (alignment depth) increase. Interface of the domain pairs with higher number of real contacts can be predicted successfully unless it has low Nf value. Fraction of native contacts (f_{nat}), interface root mean square deviation (I-RMSD), ligand root mean square deviation (L-RMSD) are measures to determine the quality of the predicted domain interface. Higher f_{nat} indicates better predictions, lower I-RMSD and L-RMSD values indicate better predictions. Nf value: number of sequences in the MSA divided by the sequence length, representing how deep the alignment is.

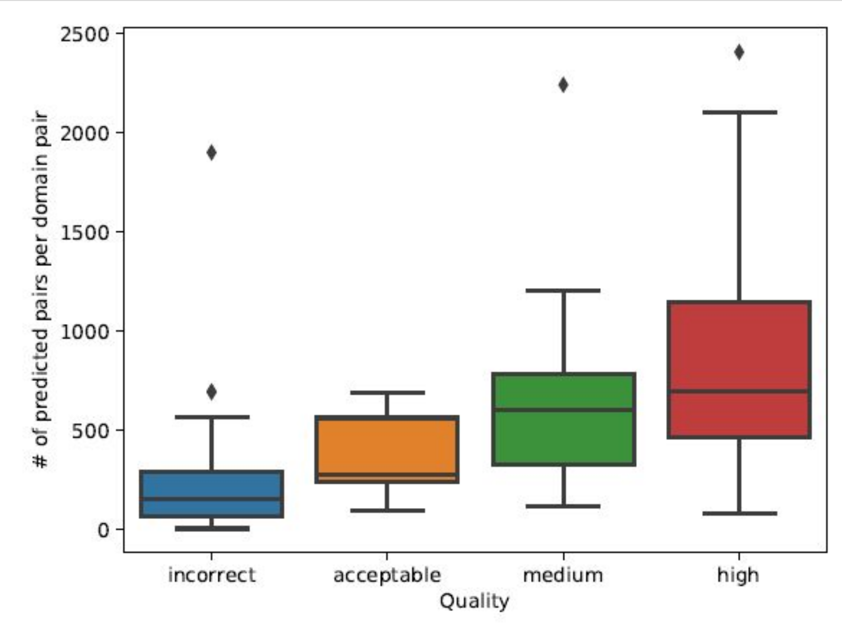


Figure 6. Comparison of the predicted residue pair numbers per domain pair between the quality groups. The number of predictions per domain pair is higher in the domains pairs that have more successful interfaces.

CONCLUSION

- Distance potentials can be predicted between residue pairs on domain pairs via deep neural networks.
- Implementation of the predicted distance potentials improves predicted domain interface quality.
- The size of the interface surface (number of contacting pairs) and alignment depth (Nf values) are limiting factors causing incorrect interface predictions.

ACKNOWLEDGEMENTS

I am thankful to the Darwin Trust of Edinburgh for the financial support and the Alan Turing Institute for the enrichment placement.