Bioinformatics, 34(23), 2018, 4039–4045
doi: 10.1093/bioinformatics/bty481
Advance Access Publication Date: 19 June 2018
Original Paper

Structural bioinformatics

Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks

Jack Hanson[1,*], Kuldip Paliwal[1], Thomas Litfin[2], Yuedong Yang[2,3] and Yaoqi Zhou[2,*]

[1]Signal Processing Laboratory, Griffith University, Brisbane, QLD 4122, Australia, [2]Institute for Glycomics and School of Information and Communication Technology, Griffith University, Southport, QLD 4222, Australia and [3]School of Data and Computer Science, Sun-Yat Sen University, Guangzhou, Guangdong 510006, China

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on February 9, 2018; revised on May 11, 2018; editorial decision on June 8, 2018; accepted on June 13, 2018

Abstract

Motivation: Accurate prediction of a protein contact map depends greatly on capturing as much contextual information as possible from surrounding residues for a target residue pair. Recently, ultra-deep residual convolutional networks were found to be state-of-the-art in the latest Critical Assessment of Structure Prediction techniques (CASP12) for protein contact map prediction by attempting to provide a protein-wide context at each residue pair. Recurrent neural networks have seen great success in recent protein residue classification problems due to their ability to propagate information through long protein sequences, especially Long Short-Term Memory (LSTM) cells. Here, we propose a novel protein contact map prediction method by stacking residual convolutional networks with two-dimensional residual bidirectional recurrent LSTM networks, and using both one-dimensional sequence-based and two-dimensional evolutionary coupling-based information.

Results: We show that the proposed method achieves a robust performance over validation and independent test sets with the Area Under the receiver operating characteristic Curve (AUC) $>$ 0.95 in all tests. When compared to several state-of-the-art methods for independent testing of 228 proteins, the method yields an AUC value of 0.958, whereas the next-best method obtains an AUC of 0.909. More importantly, the improvement is over contacts at all sequence-position separations. Specifically, a 8.95%, 5.65% and 2.84% increase in precision were observed for the top $L/10$ predictions over the next best for short, medium and long-range contacts, respectively. This confirms the usefulness of ResNets to congregate the short-range relations and 2D-BRLSTM to propagate the long-range dependencies throughout the entire protein contact map 'image'.

Availability and implementation: SPOT-Contact server url: http://sparks-lab.org/jack/server/SPOT-Contact/.

Contact: jack.hanson@griffithuni.edu.au or yaoqi.zhou@griffith.edu.au

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

1 Introduction

Proteins are one of the most biologically important macromolecules with a wide variety of functions. Because the functions of most proteins rely on their uniquely-folded three-dimensional structures, determining protein structures is of great importance to understand functional mechanisms. Due to the high cost and low efficiency of experimental techniques, *ab initio* prediction of protein structures by computational methods have been actively pursued in the past 50 years, but are still yet to be solved. To simplify the problem, breaking the structure-prediction problem into more feasible sub-problems has been the forefront of bioinformatics studies for decades.

One such sub-problem is the prediction of residue-residue contacts. By analyzing whether or not a residue pair in a protein sequence is in contact (i.e. close in 3D space), we are able to form a protein contact map, which provides key structural restraints towards the modeling of a protein's three-dimensional structure. The current methods for predicting protein contact maps can be sorted into two distinct groups: Evolutionary Coupling Analysis (ECA) and machine learning techniques.

ECA methods utilize Multiple Sequence Alignments (MSAs; Göbel *et al.*, 1994) to identify correlation in changing (co-evolving) residue pairs, using the belief that residues in close proximity mutate in sync with the evolutionary functional and structural requirements of a protein. These methods have benefited from the explosion of available protein sequences in the past decade, as these methods perform particularly well when a target protein sequence has a high number of homologues in a protein database (Ovchinnikov *et al.*, 2017). Popular ECA methods include: CCMPred (Seemayer *et al.*, 2014), FreeContact (Kaján *et al.*, 2014), GREMLIN (Kamisetty *et al.*, 2013), PlmDCA (Ekeberg *et al.*, 2013) and PSICOV (Jones *et al.*, 2012). While these methods are useful for predicting long-range contacts in proteins with a high number of sequence homologues, their accuracy is poor if the number of homologues is low (Wang and Xu, 2013).

The other, increasingly accurate methods are based on machine learning techniques. These methods have seen success due to their ability to learn underlying relationships present in sequence-based features given a set of labelled data. They have been found especially effective when predicting on proteins with few homologues. Early machine learning papers utilized Support Vector Machines (SVM; Vapnik, 1998) due to their ability to model complex relationships despite a lack of processing power and extensive data, such as SVMCon (Cheng and Baldi, 2007), SVMSEQ (Wu and Zhang, 2008) and the recently-released R2C (Yang *et al.*, 2016). Other papers have found success in exploiting the ever-increasing amount of available training data by the application of Deep artificial Neural Networks (DNN's; Rumelhart *et al.*, 1985) in various forms, such as two-dimensional (2D) Recursive NNs (Rec-NN's; Baldi and Pollastri, 2003) and Deep Belief Network's (Hinton *et al.*, 2006). Such predictors include Betacon (Cheng and Baldi, 2005), CMAPPro (Di Lena *et al.*, 2012), DeepConPred (Xiong *et al.*, 2017) and NNCon (Tegge *et al.*, 2009).

Complementary methods can be combined in the form of metapredictors, where a single network combines the outputs of several other classifiers. Examples of this architecture are MetaPSICOV (Jones *et al.*, 2015) and NeBCON (He *et al.*, 2017). MetaPSICOV received the best prediction results in a recent review by Wuyun *et al.* (2016).

The recently-released RaptorX-Contact (Wang *et al.*, 2017) and DNCON2 (Adhikari *et al.*, 2017) predictors are the first approaches to attempt the incorporation of the entire protein 'image' as context for prediction. The architecture utilized in these models are Convolutional NN's (CNN; LeCun *et al.*, 1989), with RaptorX-Contact utilizing a Residual CNN, or ResNet (He *et al.*, 2016a). ResNets achieve identity mappings between several layers by employing shortcut connection between the output of a previous layer and the current output. This allows these models to have ultra-deep architectures due to their ease of propagating the error gradient through many layers, and have been shown to benefit from having over 100 convolutional layers. RaptorX-Contact is currently the state-of-the-art predictor for the latest Critical Assessment of Structure in Proteins (CASP) round, demonstrating that access to the entire protein as context is greatly beneficial to learning contact maps (Schaarschmidt *et al.*, 2018; Wang *et al.*, 2018). It also is one of the techniques to benefit from combining sequence-derived features with the information from evolution coupling (Betancourt and Thirumalai, 1999; Miyazawa and Jernigan, 1985). DNCON2, on the other hand, can be split into two sections. The first uses a set of intermediate CNNs to predict contact maps at several distances (from 6–10 Å). It then combines these separate predictions in a secondary CNN to provide the final contact map at 8 Å.

The neural network architecture utilized in this paper was inspired by RaptorX-Contact and the Multi-Dimensional Recurrent Neural Network (MD-RNN) in Graves *et al.* (2007), in which stacked 2D RNNs were proven able to progress information throughout entire 2D images. The advantages here were that the model was able to generalize along all input spatio-temporal dimensions, making the model robust to distortions in any mixture of the input dimensions, with the architecture performing particularly well on warped data in comparison to CNNs. The MD-RNN was simplified in ReNets (not to be confused with ResNets), where the *x* and *y* dimensions' forward and backward RNNs were separated between consecutive layers (Visin *et al.*, 2015). Min *et al.* (2017) described the MD-RNN as an emerging architecture in bioinformatics in a recent review.

In combination with RNNs, Long Short-Term Memory (LSTM) cells (Hochreiter and Schmidhuber, 1997) are commonly used to model long-range context which is vital to modeling complex relationships between non-local datapoints. Bidirectional LSTM networks (Schuster and Paliwal, 1997) have already seen success in bioinformatics applications, where their effective propagation of deep residue structural interdependencies have provided state-of-the-art results in protein secondary structure prediction (Heffernan *et al.*, 2017) and protein disorder prediction (Hanson *et al.*, 2017), the latter of which demonstrating that LSTM cells are able to accurately predict sparse data, an aspect shared between protein disorder and contact map prediction.

In this paper, we aim to capture these deep, underlying relationships between non-local residue pairs in both spatial dimensions for protein contact map prediction by the use of an ultra-deep hybrid network, consisting of a ResNet coupled with 2D Bidirectional-ResLSTMs (2D-BRLSTM). Using this technique, the proposed method, called SPOT-Contact (Sequence-based Prediction Online Tools for Contact map prediction), is able to capture contextual information from the whole protein 'image' at each layer. SPOT-contact has been found to be highly accurate for predicting contacts at all sequence-position separations, significantly outperforming all methods compared.

2 The machine learning approach

2.1 Ensemble of two-dimensional bidirectional recurrent neural networks and ResNets

Our approach to the problem is built up of an ensemble of models, all based on slight variations of the network architecture shown in

Figure 1. The base model can be broken down into four separate segments: input preparation, ResNet, 2D-BRLSTM and Fully-Connected (FC).

The data preparation segment involves the transformation of our sequence-based one-dimensional features into a two-dimensional representation. This is achieved through the outer concatenation function, as described in Wang *et al.* (2017). However, rather than concatenating the features of residues i, j and $\frac{i+j}{2}$ at position (i, j), we omit the concatenation of the midpoint residue. The product of the concatenation stage is then depth concatenated (i.e. concatenated along the feature axis) with the two-dimensional features.

The ResNets used in our model utilize the pre-activation order of operations as proposed in He *et al.* (2016b). As such, an initial convolutional layer was placed before the first residual block. The residual block architecture is shown in the ResNet Block section of Figure 1. The output of the entire ResNet was also activated and normalized prior to the 2D-BRLSTM section. The size of the convolutions alternated between a kernel size of 3 x 3 and 5 x 5, both with 60 filters and had a Exponential Linear Unit (ELU) activation (Clevert *et al.*, 2015). ELU activations have been seen to be more effective than the standard ReLU activation function in learning for ResNets (Shah *et al.*, 2016).

The 2D-RNNs used in this model differ to Graves *et al.* (2007) by the fact that each of the four directions' RNNs in each layer is completely independent of all of the other directions' due to limitations in the coding environment, as is similar to Visin *et al.* (2015). Although each directions' outputs are calculated separately, they are concatenated at the output of each layer to provide information from the entire 2D image plane to the next layer. However, He *et al.* (2016a) discussed that the identity mapping function is less effective at error propagation when the residual connection is connected over only a single layer's activation. Therefore, we add a bottleneck layer before the LSTM layers to increase the depth of the residual connection. This also has the added benefit of reducing the parameter count of the model. The bottleneck and LSTM layers form our BRLSTM blocks (Kim *et al.*, 2017) as shown in Figure 1. This bottleneck connection is established by a 1×1 convolution with an ELU activation.

The default 2D-BRLSTM layer's LSTM cells consist of 200 one-cell memory blocks, culminating in 800 inputs at the succeeding layer. The FC layers consist of 400 nodes plus a bias node with an ELU activation, except at the output layer which has a single output neuron and a sigmoid activation to convert the output into a likelihood of a residue pair being in contact. The network layout from Figure 1 is only changed when omitting the bottleneck layer, and when the 2D-BRLSTM is placed first in the network (see the model variants listed in Supplementary Table S1). All of the parameters discussed above were chosen after thorough experimentation, during which this architecture was found to obtain consistently high accuracies through short, medium and long distance contacts on a validation set, yet fitting into the strict computational memory constraints of a 2D-BRLSTM.

Each model was trained with the ADAM optimization algorithm (Kingma and Ba, 2014), which has been shown to converge more quickly than the traditionally-used Stochastic Gradient Descent with the additional benefit of having standard hyperparameters which require little to no tuning. Regularization was applied to the network through the use of layer normalization (Lei Ba *et al.*, 2016) at each normalization block in Figure 1, and a 50% dropout rate at the output of the FC layer during training (Srivastava *et al.*, 2014).

Using an ensemble predictor allowed us to minimize the effects of generalizations on the data. Because the tuned weights of a neural network learn slightly different representations of the data (due to various weight initializations and training data feeding), these lead to various errors at the output which are dependent on the generalizations made. Assuming that the correct outputs should be more common between the individual predictors, the collective decision between all of the predictors should be less likely to contain the errors pertaining to an individual predictor's generalization (Hansen and Salamon, 1990). All six models (base network, base without bottleneck, base without FC, 2D-BRLSTM prior to ResNet in the base model and the 2D-BRLSTM only model) used in SPOT-Contact are shown along with their network parameters in Supplementary Table S1. The results of SPOT-Contact are provided by the mean of all six networks' outputs. An ensemble of models was also utilized in RaptorX-Contact.

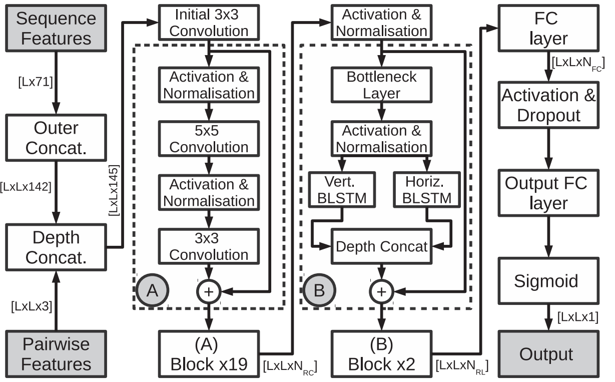

Fig. 1. The network layout of the SPOT-Contact. The ResNet (Residual Convolutional Neural Network) and 2D-BRLSTM (2-Dimensional Bidirectional Residual LSTM Network) functions are provided in boxes 'A' and 'B', respectively. N_{RC}, N_{RL}, and N_{FC} are the depth of the CNN filters in the ResNet block, four times the depth of the LSTM layers in the 2D-BRLSTM block and the depth of the FC layer, respectively, and L is the length of the input protein. Depth concatenation means concatenation along the last dimension

Training of the model was executed in the framework of Google's Tensorflow library (v1.4; Abadi *et al.*, 2016), enabling us to accelerate the training of the model by training the model on an Nvidia GTX TITAN X Graphics Processing Unit (GPU). Oh and Jung (2004) showed that the use of a GPU in neural network training can speed up the total training time by up to a factor ≈20. Total training time was mostly influenced by the depth of the 2D-BRLSTM layers, with each network taking roughly 50 h for 15 epochs over our whole training set. The size of the 2D-BRLSTM layers dictated the memory consumption during training. Thus, the model hyperparameters, such as LSTM cell size, were chosen as a compromise between the training time, memory usage and performance of the model. Deeper and larger architectures were tested by spreading the model over multiple GPU's, but this was not found to improve performance significantly.

2.2 Input features

The inputs to our model included both one-dimensional (i.e. along the primary sequence) and two-dimensional features (i.e. pairwise, or per residue pair). One-dimensional features consisted of the Position-Specific Scoring Matrix (PSSM) profile, the HMM profile from HHblits (Remmert *et al.*, 2012) and several predicted structural probabilities from SPIDER3 (Heffernan *et al.*, 2017). The PSSM profile was generated by three iterations of PSI-BLAST (Altschul *et al.*, 1997) against the UniRef90 sequence database updated in April 2018. The HMM profile was generated by hhblits v3.0.3 with default parameters by searching the UniClust30 profile HMM database updated in October 2017 (Mirdita *et al.*, 2016). The predicted values obtained from SPIDER3 were: 1 relative solvent-Accessible Surface Area, 2 Half-Sphere Exposures based on the Cα atom, the 8 sine/cosines of the backbone torsion angles (theta, tau, phi and psi) and the three probabilities of the predicted secondary structure. Finally, we also employed a set of seven physicochemical properties provided by Meiler *et al.* (2001). This gives a total of 71 1D features for the initial section of the network.

The 2D features consist of the output from CMMPred (Seemayer *et al.*, 2014), and 2 outputs (mutual and direct-coupling information) from DCA (Morcos *et al.*, 2011), resulting in three pairwise features for concatenation with the output of the first section of the network. The data was standardized to have zero mean and unit variance (according to the training data) before being input into the network.

2.3 Datasets

We downloaded 30% non-redundant sequences with resolution <2.5 Å and R-factor < 1.0 from the cullpdb website, which contained 14 541 chains on February 2017. After removing obsolete sequences, sequences containing less than 30 residues and proteins greater than 25% sequence similarity according to BlastClust (Altschul *et al.*, 1997), we kept 12 450 chains. In order for a good comparison with other methods, we have kept all 1250 chains deposited after June 2015 as our independent test set (Test) and the remaining 11 200 as training set. We produced a difficult ('hard') subset of Test (Test-Hard), by removing any proteins in Test that have a PSI-Blast E-value of 0.1 or less to any proteins in the training set (i.e. further removing any proteins with potential homologous relations to the training set). This 'hard' dataset contains 280 chains.

Due to limitations of the coding environment imposed by the large memory usage by the 2D-BRLSTM model, training and testing input proteins are limited to proteins of length ≤ 300 and ≤ 700 residues, respectively. While this restriction excludes many proteins, it still incorporates upwards of 90% of single domain sequences for testing with our model (Islam *et al.*, 1995). This restriction left us with 7557 training proteins, and a validation, Test and Test-Hard sets of 983, 1213 and 277 proteins, respectively. We also obtained 22 template-free modeling (TFM) CASP12 targets as an additional test set. Their sequence similarity to our training set is also <25% according to BlastClust. All of these datasets can be obtained from www.sparks-lab.org.

To gauge the performance increase by training on new sequence profiles, we also trained an exact replica of SPOT-Contact using the UniRef and UniProt datasets from March 2017 and February 2016, respectively. This model, named SPOT-Contact-2016 will serve as a baseline to compare the other predictors to, to illustrate that the performance improvement reported here is not solely caused by a simple update of sequence libraries.

2.4 Performance evaluation

Protein residues in these experiments are considered to be in contact when the inter-residue distance between the two C_β atoms is ≤ 8.0 Å, following the standard CASP definition (Ezkurdia *et al.*, 2009). For a protein of length L, we first separate it into short $(12 > |i-j| \geq 6)$, medium $(24 > |i-j| \geq 12)$ and long $(|i-j| \geq 24)$ sequence-position-separated residues. From these groups, we take the top L/k highest-ranked predictions (where $k \in \{10, 5, 2, 1\}$) from the predicted contact map and calculate the precision, recall and F1 scores of these values, where:

$$\overline{\text{Prec}} = \frac{1}{N}\sum_{n=0}^{N}\frac{\text{True Positives}_n}{\text{True Positives}_n + \text{False Positives}_n}, \tag{1}$$

$$\overline{\text{Recall}} = \frac{1}{N}\sum_{n=0}^{N}\frac{\text{True Positives}_n}{\text{True Positives}_n + \text{False Negatives}_n}, \tag{2}$$

and

$$\text{F1} = 2 \cdot \frac{\overline{\text{Prec}} \cdot \overline{\text{Recall}}}{\overline{\text{Prec}} + \overline{\text{Recall}}}, \tag{3}$$

where N is the number of proteins in the test set.

These three metrics are chosen for consistency with the metrics used in the CASP12 rankings. However, these CASP metrics provide an evaluation on the positive predictions of a predictor, but do not provide any information on the predictions not in the top L/k predictions, especially those negative predictions. Thus, they are biased towards predictors which weight positive predictions. It should be noted that SPOT-Contact was not trained with a scaling factor for the positive samples. To analyze all the predictions from our model, we utilize the Area Under the receiver operating characteristic Curve (AUC) metric as an overall performance evaluator. This metric provides the probability that the predictor will rank a random positive sample higher than a random negative sample (Fawcett, 2006). We can compare the AUC scores of other predictors with respect to SPOT-Contact using a P-value, which indicates the statistical significance of the difference between the two predictors' results (Hanley and McNeil, 1982). The smaller the P-value is, the more significant the difference between the two predictors. The AUC values for short-, medium- and long-range contacts were also obtained separately to examine the improvements in more detail.

The latest CASP publication (Schaarschmidt *et al.*, 2018) noted that the area under the Precision-Recall (PR) curve provides a balanced metric. Thus we also analyze the AUC of the PR curve, to analyze the positive predictions of the predictor.

2.5 Method comparison

In order to gauge the performance of SPOT-Contact on our independent test sets, we chose several other readily-available, recently-developed contact-map predictors. The standalone versions of MetaPSICOV (Download: http://bioinfadmin.cs.ucl.ac.uk/downloads/MetaPSICOV/), Gremlin (Download: http://gremlin.bakerlab.org), SVMCon (Download: http://scratch.proteomics.ics.uci.edu), SVMSeq (Download: http://zhanglab.ccmb.med.umich.edu/SVMSEQ), Deep ConPred and DeepRCon (obtained from Xiong *et al.*, 2017) and PlmDCA (Download: https://github.com/pagnani/PlmDCA) were used in these experiments. The standalone version of MetaPSICOV also provided the results of EVFold, PSICOV and CCMPred. We also submitted jobs to the online servers of DNCON2 (server URL: http://sysbio.rnet.missouri.edu/dncon2/), RaptorX-Contact (Server URL: http://raptorx.uchicago.edu/ContactMap/), R2C (Server URL: http://www.csbio.sjtu.edu.cn/bioinf/R2C/), NeBcon (Server: https://zhanglab.ccmb.med.umich.edu) and CMapPro (Server URL: http://scratch.proteomics.ics.uci.edu/). Other predictors were considered, but were ultimately far too time-consuming to do large-scale predictions or were unavailable (Li *et al.*, 2016).

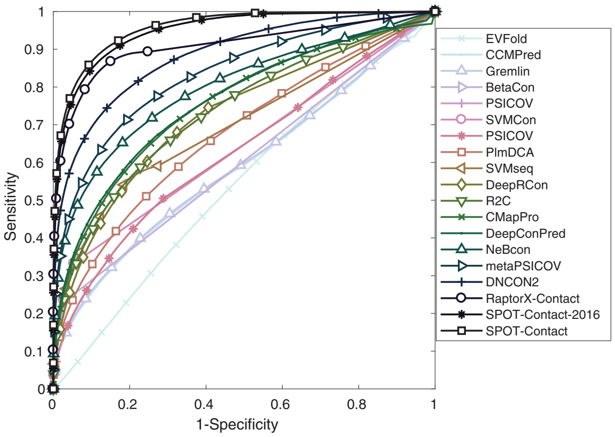

Fig. 2. Receiver Operating Curves for 19 predictors on the Test-Hard subset

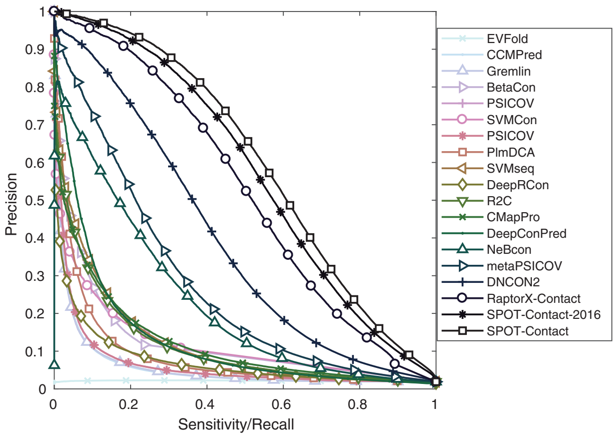

Fig. 3. PR curve for 19 predictors on the Test-Hard subset

3 Results

The results of SPOT-Contact on each of the validation and independent testing datasets are shown in Supplementary Table S2 for AUC and the mean precision of the results when separated into both sequence separation cutoffs (short, medium and long), and top-ranking prediction cutoffs ($L/10$, $L/5$, $L/2$ and L). The F1 results are provided in Supplementary Table S3. Here, we did not perform multi-fold cross-validation in training because of the time-consuming nature of every training run. Nevertheless, the comparable performance in AUC (0.976 versus 0.973) for the validation and the independent Test set indicate the robustness of the ensemble predictor. Similar precisions are observed regardless of whether it is a short, medium, or long-range contact (88%, 88% and 92% for top $L/10$ predictions for short, medium and long-range contacts in the test set, respectively). The predictions of SPOT-Contact on the harder subset (Test-Hard) obtained somewhat lower AUC, precisions and F1 scores. This was expected, as more proteins in this subset have fewer homologous sequences and thus are harder to predict. The average number of effective homologous sequences from HHblits is 7.93 for the Test set but only 6.19 for the Test-Hard set.

3.1 Ensemble model analysis

To illustrate the contribution of each individual predictor, the individual and cumulative precision values for the Test-Hard are shown in Supplementary Tables S4 and S5, respectively. Because it is not possible to directly compare the ResNets used in RaptorX-Contact, we trained a pure ResNet model to compare to other ensemble component models. As the results in Supplementary Table S4 show, while the ResNet model is somewhat effective at short-range prediction, it lacks the long-range modeling capabilities to predict the long-range contacts as effectively as the pure 2D-BRLSTM model. However, ResNets and 2D-BRLSTMs can be enhanced by combining the two in hybrid models, showing the benefits of using the ResNets to congregate the short-term relationships and then using the 2D-BRLSTM to propagate this information throughout the protein image.

Supplementary Table S5 shows how much of a boost to performance each network adds to the original base network, culminating in an increase in the long-range predictions' precisions by 5.31, 4.79, 5.24 and 4.54% for Test-Hard at the cutoffs of top $L/10$, $L/5$, $L/2$, L prediction, respectively. While further gains could be obtained from adding more models, the cumulative performance gains become incrementally smaller as the number of networks in the ensemble increases, indicating that it would not be worthwhile to increase the number of networks further.

3.2 Feature importance

It is of interest to see the effect of the individual feature groups on the prediction accuracy. Much research has been conducted regarding the effectiveness of evolutionary profiles, sequence structure and physicochemical properties on protein structure prediction (Hanson *et al.*, 2017), but such insight does not exist for the 2D features in protein contact map prediction. Thus, we have trained our baseline model without our 2D feature groups (the evolutionary coupling features and CCMpred output) sequentially to see the effect of the 2D features on our predictions. As is shown in Supplementary Table S6, 2D features from CCMpred and DCA led to significant improvement over the model based on 1D feature only. The improvement is significant in all short, medium and long-range contact pairs. For example, the improvement for top L/10 prediction is 3.7%, 6.5% and 10.3% for short, medium and long-range contacts, respectively. Similar level of improvement in F1 measure is also observed in Supplementary Table S7. The overall improvement in AUC is 3.3% from 0.908 to 0.941.

3.3 Comparison to the 17 other methods

The performance of the 17 other predictors on a subset of Test-Hard can be found in Supplementary Table S8 through Supplementary Table S11. The corresponding Receiver-Operating Characteristic (ROC) and PR curves are shown in Figures 2 and 3. Mean precisions given by different methods for short, medium and long-range contacts are also shown in Figure 4. Predictors such as CMapPro and PSICOV have maximum length or a minimum number of sequence homologues requirements, which trimmed the size of our dataset from 277 to 228. SPOT-contact significantly outperforms all compared models over all performance evaluations on this dataset. For example, as shown in Supplementary Table S10, the largest AUC is 0.958 given by SPOT-contact, followed by SPOT-Contact-2016 (0.950), RaptorX-Contact (0.909) and DNCON2 (0.886). The AUC values for all other methods are less than 0.84. A two-tailed P-value of $< 10^{-6}$ is obtained when comparing SPOT-contact's AUC value against all external predictors. This improvement is not dependent on the length cutoff, as Supplementary Table S10 shows that the increase of AUC for SPOT-Contact is present over all residue-separation cutoffs. For example, the AUC values for short-, medium- and long-range contacts are 0.935, 0.949 and 0.888 by RaptorX-Contact, respectively, compared to 0.960, 0.965 and 0.951 by SPOT-Contact, respectively.

We also compare the PR curves of each predictor, and calculate the AUC of these curves in Supplementary Table S11. SPOT-Contact obtains the highest AUC, even when segmented by residue

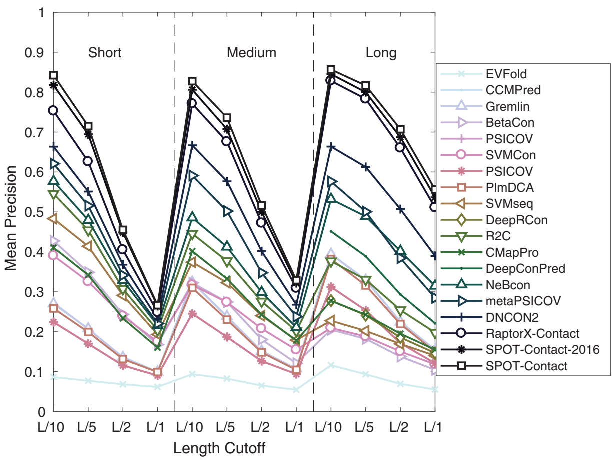

Fig. 4. Mean precisions for 19 predictors on the Test-Hard subset

separation. SPOT-Contact increases on RaptorX-Contact's overall PR AUC by 0.1 from 0.554 to 0.658, 0.076 from 0.584 to 0.660 and 0.07 from 0.457 to 0.527 for the short-, medium- and long-range contacts, respectively.

Most significantly, SPOT-contact increases the already-outstanding long-range contact precision scores of RaptorX-contact (e.g. from 51.1 to 55.6 in top *L* long-range prediction), further increasing the gap between modern machine learning techniques and ECA methods. This is independent of sequence profile database selection, as SPOT-Contact-2016 also improved on RaptorX-Contact in each analyzed metric.

In particular, SPOT-contact is the only method to achieve more than 80% precision for short, medium and long-distance contacts for top *L*/10 predictions. This happened without specific training for precision. It can be noticed that the ECA-based predictors perform poorly on this dataset, due many proteins in this dataset not having a large number of sequence neighbors in the existing sequence library. SPOT-contact receives slightly higher evaluation metrics on this subset of 228 proteins in Test-Hard than the results in Supplementary Table S2 because the maximum length and minimum number of sequence homologue restrictions placed by other predictors makes this subset easier to predict than the full 277-protein Test-Hard dataset. This is confirmed by a similar decrease in the evaluation metrics from RaptorX-Contact; for instance the mean precision of the top *L* long-range residue pairs decreases from 51.09% in the subset to 49.31% in the full set.

To confirm the dependence of method-performance on the number of homologous sequences, we present the mean precision values of the top *L*/5-ranked predictions as a function of the maximum cumulative Neff values (the number of effective homologous sequences, ranging from 1 to 20) from HHblits, in Supplementary Figure S1. All methods had their lowest performance for lower Neff sequences. SPOT-Contact improves over RaptorX-Contact at all Neff values, with the largest improvement for low-medium range Neff sequences. We further analyze the results in accordance with the number of contacts a residue has from the reduced Test-Hard. We bin each residue in our database depending on the number of contacts it has, and calculate the mean precision of the top *L*/5 long-range precisions. As shown in Supplementary Figure S2, residues with fewer contacts (surface contacts) are much harder to discriminate from their non-contacts, with each additional contact bringing an almost linear increase in performance to all predictors. SPOT-Contact and RaptorX-Contact show a distinct advantage over all other methods, with our method maintaining an increase in performance over RaptorX-Contact for all contact numbers.

We further examined the dependence of prediction precision on protein secondary structure. Supplementary Table S12 compares the performance of the four top-performing methods (metaPSICOV, RaptorX-Contact, SPOT-Contact-2016 and SPOT-Contact) for residues with different secondary structure elements on the full set of Test-Hard. The contacts between sheet residues have the highest precision for all three methods. Using the updated sequence profiles, SPOT-Contact increases on SPOT-Contact-2016 over all secondary structure pairs for all length cutoffs. SPOT-Contact also increases upon the performance of RaptorX-Contact, which attained similar performance for some residue pairs with SPOT-Contact-2016.

3.4 CASP results

For completeness, we compared SPOT-Contact the other predictors in Supplementary Table S13 for the available TFM CASP targets (22 proteins only). At the time of testing, several servers were not available and others are unable to predict the full 22-protein set. Only the remaining servers were provided in Supplementary Table S13. SPOT-Contact achieves the highest AUC of both the ROC and PR curves (0.906 compared to the next highest 0.862, and 0.443 compared to 0.369, respectively), and also achieves the highest mean precision values across all length cutoffs and distance separations. For example, the *L*/5 cutoff scores for RaptorX-Contact (the CASP12 winner) and SPOT-Contact are 59.76% and 64.06%, 56.14% and 64.41% and 58.99% and 63.95% for the short-, medium- and long-range contacts, respectively.

4 Conclusion

In this paper, we have developed a new predictor called SPOT-Contact for protein contact maps. This method is built on the previous success of residual CNN in contact map prediction by RaptorX-Contact and the capability of capturing long-range interactions by LSTM-RNN networks by inputting whole sequences. In addition, the ensemble of six predictors with different combinations of networks removes prediction noise and makes prediction more generalizable. Using 228 recently-solved structures as an independent test set, SPOT-Contact is consistently more accurate in contact prediction across contacts at different sequence separations (Fig. 4), across proteins with different number of effective homologous sequences (Supplementary Fig. S1) and across residues with different number of contacts (Supplementary Fig. S2). The improvement in AUC is 5% over the next best technique RaptorX-Contact, and is statistically significant ($P < 10^{-6}$). The result highlights the usefulness of coupling ResNets with two-dimensional LSTM networks.

Funding

This work was supported by Australia Research Council DP180102060 to Y.Z. and K. P., and in part by National Health and Medical Research Council (1121629) of Australia to Y.Z., the National Natural Science Foundation of China (U1611261, 61772566) and the program for Guangdong Introducing Innovative and Entrepreneurial Teams (2016ZT06D211) to Y.Y. We also gratefully acknowledge the use of the High Performance Computing Cluster 'Gowonda' to complete this research, and the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

Conflict of Interest: none declared.

References

Abadi,M. *et al.* (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems. *CoRR*, Abs/1603.04467.

Adhikari,B. *et al.* (2017) DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*, **1**, 7.

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.

Baldi,P. and Pollastri,G. (2003) The principled design of large-scale recursive neural network architectures–dag-rnns and the protein structure prediction problem. *J. Mach. Learn. Res.*, **575**, 602.

Betancourt,M.R. and Thirumalai,D. (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.*, **8**, 361–369.

Cheng,J. and Baldi,P. (2005) Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, **21**, i75–i84.

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

Clevert,D.A. *et al.* (2015) Fast and accurate deep network learning by exponential linear units (elus.). *arXiv Preprint arXiv: 1511.07289*.

Di Lena,P. *et al.* (2012) Deep architectures for protein contact map prediction. *Bioinformatics*, **28**, 2449–2457.

Ekeberg,M. *et al.* (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Phys. Rev. E*, **87**, 012707.

Ezkurdia,I. *et al.* (2009) Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Prot. Struct. Func. Bioinform.*, **77**, 196–209.

Fawcett,T. (2006) An introduction to ROC analysis. *Patt. Recogn. Lett.*, **27**, 861–874.

Göbel,U. *et al.* (1994) Correlated mutations and residue contacts in proteins. *Prot. Struct. Funct. Bioinform.*, **18**, 309–317.

Graves,A. *et al.* (2007) Multi-dimensional recurrent neural networks. *CoRR*, Abs/0705.2011.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hansen,L.K. and Salamon,P. (1990) Neural network ensembles. *IEEE Trans. Patt. Anal. Mach. Intel.*, **12**, 993–1001.

Hanson,J. *et al.* (2017) Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, **33**, 685–694.

He,B. *et al.* (2017) NeBcon: protein contact map prediction using neural network training coupled with naïve bayes classifiers. *Bioinformatics*, **33**, 2296–2306.

He,K. *et al.* (2016a) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 770–778.

He,K. *et al.* (2016b) Identity mappings in deep residual networks. In: *European Conference on Computer Vision*. Springer, Amsterdam, pp. 630–645.

Heffernan,R. *et al.* (2017) Capturing non-local interactions by long short term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure. *Bioinformatics*, **33**, 2842–2849.

Hinton,G.E. *et al.* (2006) A fast learning algorithm for deep belief nets. *Neural Comput.*, **18**, 1527–1554.

Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.

Islam,S.A. *et al.* (1995) Identification and analysis of domains in proteins. *Prot. Eng.*, **8**, 513–526.

Jones,D.T. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Jones,D.T. *et al.* (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, **31**, 999–1006.

Kaján,L. *et al.* (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, **15**, 85.

Kamisetty,H. *et al.* (2013) Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proc. Natl. Acad. Sci.*, **110**, 15674–15679.

Kim,J. *et al.* (2017) Residual LSTM: design of a deep recurrent architecture for distant speech recognition. *CoRR*, Abs/1701.03360.

Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. *CoRR*, Abs/1412.6980.

LeCun,Y. *et al.* (1989) Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, **1**, 541–551.

Lei Ba,J. *et al.* (2016) Layer normalization. *ArXiv e-Prints, Abs/1607.06450*.

Li,Q. *et al.* (2016) Kscons: a bayesian approach for protein residue contact prediction using the knob-socket model of protein tertiary structure. *Bioinformatics*, **32**, 3774–3781.

Meiler,J. *et al.* (2001) Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Mol. Model. Annu.*, **7**, 360–369.

Min,S. *et al.* (2017) Deep learning in bioinformatics. *Brief. Bioinformatics*, **18**, 851–869.

Mirdita,M. *et al.* (2016) Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.*, **45**, D170–D176.

Miyazawa,S. and Jernigan,R.L. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.*, **108**, E1293–E1301.

Oh,K.S. and Jung,K. (2004) GPU implementation of neural networks. *Patt. Recogn.*, **37**, 1311–1314.

Ovchinnikov,S. *et al.* (2017) Protein structure determination using metagenome sequence data. *Science*, **355**, 294–298.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Rumelhart,D.E. *et al.* (1985) Learning internal representations by error propagation. *Tech. Rep. DTIC Document*.

Schaarschmidt,J. *et al.* (2018) Assessment of contact predictions in casp12: co-evolution and deep learning coming of age. *Prot. Struct. Funct. Bioinform.*, **86**, 51–66.

Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, **45**, 2673–2681.

Seemayer,S. *et al.* (2014) CCMpredfast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, **30**, 3128–3130.

Shah,A. *et al.* (2016) Deep residual networks with exponential linear unit. In: *Proceedings of the Third International Symposium on Computer Vision and the Internet VisionNet'16*. ACM, New York, NY, USA, pp. 59–65.

Srivastava,N. *et al.* (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.

Tegge,A.N. *et al.* (2009) NNcon: improved protein contact map prediction using 2d-recursive neural networks. *Nucleic Acids Res.*, **37**, W515–W518.

Vapnik,V.N. (1998) *Statistical Learning Theory*. Vol. **1**, Wiley, New York.

Visin,F. *et al.* (2015) ReNet: a recurrent neural network based alternative to convolutional networks. *CoRR, Abs/1505.00393*.

Wang,S. *et al.* (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.*, **13**, e1005324–e1005334.

Wang,S. *et al.* (2018) Analysis of deep learning methods for blind protein contact prediction in casp12. *Prot. Struct. Funct. Bioinform.*, **86**, 67–77.

Wang,Z. and Xu,J. (2013) Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics*, **29**, i266–i273.

Wu,S. and Zhang,Y. (2008) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics*, **24**, 924–931.

Wuyun,Q. *et al.* (2016) A large-scale comparative assessment of methods for residue–residue contact prediction. *Brief. Bioinform.*, **19**, 219–230.

Xiong,D. *et al.* (2017) A deep learning framework for improving long-range residueresidue contact prediction using a hierarchical strategy. *Bioinformatics*, **33**, 2675–2683.

Yang,J. *et al.* (2016) R2C: improving *ab initio* residue contact map prediction using dynamic fusion strategy and Gaussian noise filter. *Bioinformatics*, **32**, 2435–2443.