

Distance-based Protein Folding Powered by Deep Learning

Jinbo Xu

Toyota Technological Institute at Chicago
(affiliate of Univ. of Chicago)

Available at [bioRxiv](#) and [arXiv](#)

Deep Convolutional Residual Neural Network




Winner of 2018 PLoS CB Research Prize in Breakthrough/Innovation

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model

Sheng Wang , Siqi Sun , Zhen Li, Renyu Zhang, Jinbo Xu 

253
Save

70
Citation

23,843
View

12
Share

Findings:

- 1) First time showing contact prediction can be greatly improved by DL
- 2) In CAMEO, predicted contacts can fold quite a few FM targets on which the other CAMEO servers failed

Deep Convolutional Residual Neural Network




Winner of 2018 PLoS CB Research Prize in Breakthrough/Innovation

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model

Sheng Wang , Siqi Sun , Zhen Li, Renyu Zhang, Jinbo Xu 

253
Save

70
Citation

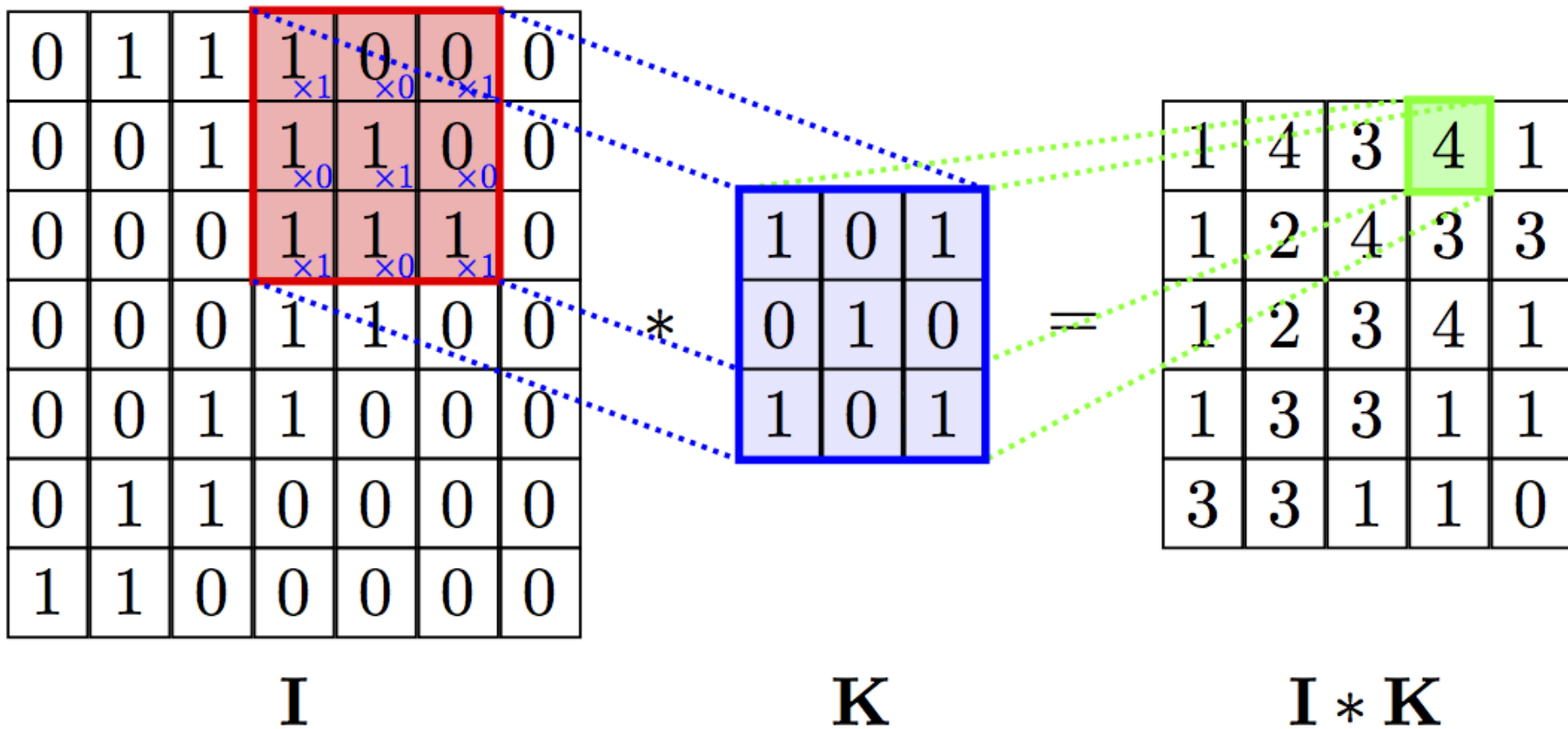
23,843
View

12
Share

Trained by **single-chain, soluble** proteins, works well for

- Membrane proteins (Cell Systems 2017)
- Complex contact prediction (RECOMB 2018, NAR 2018)
- Domain-domain contact prediction (ask Martin Weigt)

Convolution: Pattern Detection & Information Exchange



What's Residual ?

- To predict y from x , first predict $y-x$ from x
- Then add x and predicted $y-x$ to get y
- $y-x$ is the residual

- This makes it easy to stack multiple convolutional layers together to form a very deep network

Contact Prediction Methods

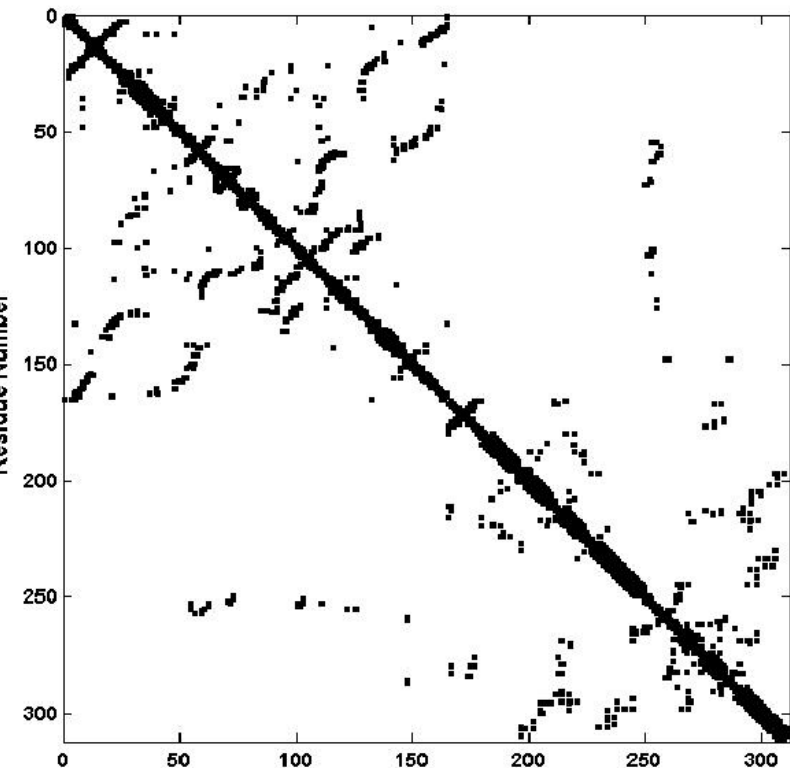
	Local Method	Global Method
Co-variation analysis	Mutual Information Correlation Coefficient	DCA (EVfold, PSICOV, CCMpred, plmDCA, GREMLIN...)

Local method: predict the label of one residue pair
independent of the labels of the others

Global method: predict the label of one residue pair
considering the labels of the others

Contact Prediction Methods

	Local Method	Global Method
Co-variation analysis	Mutual Information Correlation Coefficient	DCA (EVfold, CCMpred, plmDCA, PSICOV...)
Machine learning	MetaPSICOV, DNCON(Deep Belief Networks), PConsC3, PhyCMAP, SVMSEQ...	



1. Split into small submatrices



2. Assign a label to each submatrix:
1 if its center is in contact, 0 otherwise

3. Predict a label for each submatrix
independent of the others

Contact Prediction Methods

	Local Method	Global Method
Co-variation analysis	Mutual Information Correlation Coefficient	DCA (EVfold, CCMpred, plmDCA, PSICOV...)
Machine learning	MetaPSICOV, DNCON(Deep Belief Networks), PConsC3, PhyCMAP...	Deep ResNet (RaptorX-Contact)

RaptorX-Contact is the first global machine learning method for contact prediction

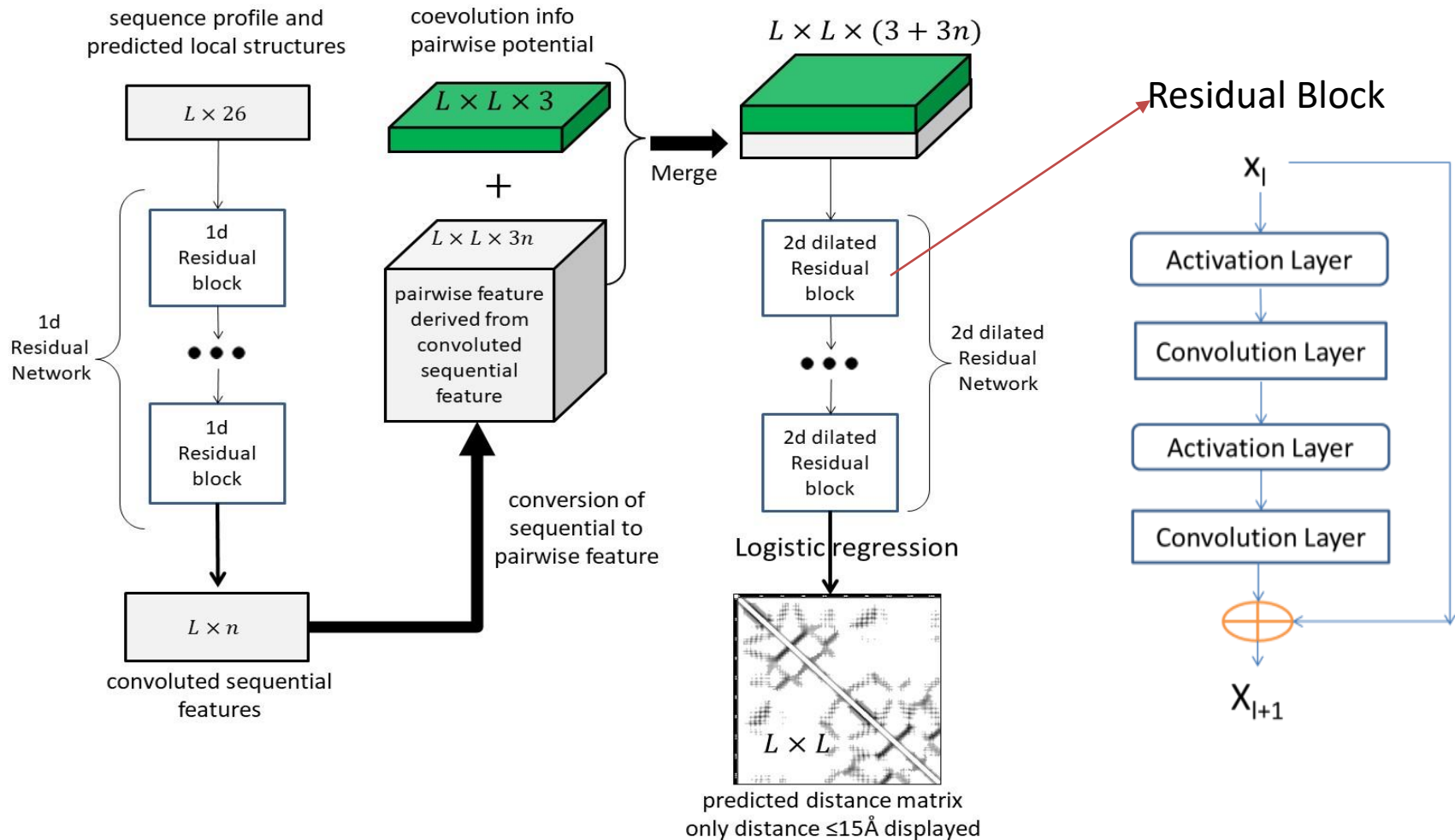
Key Idea

1. Predict the label of one residue pair while considering the labels of the other residue pairs
2. Use contact patterns and global conformation context to predict the label of a single residue pair

Formulate very differently than previous learning methods

3. Treat the whole contact matrix as an image (**do NOT split it !**) and each residue pair as a pixel
4. Pixel-level labeling (i.e. semantic segmentation)
5. Apply convolution to the whole matrix
6. Use residual modules to build **deep** networks to capture **long-range** and **global** information.

Deep ResNet for Contact Prediction



Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning. PLoS Computational Biology, January 2017. **2018 PLoS CB Research Prize in the category of breakthrough/innovation.**

Next step 1:

Finally, instead of predicting contacts, our deep learning model actually can predict inter-residue distance distribution (i.e., distance matrix), which provides finer-grained information than contact maps and thus, shall benefit 3D structure modeling more than predicted contacts.

1. Distance matrix contains more information than contact matrix
2. Trained by distance matrix, DL can learn more about physical properties of a protein and thus, reduce conformation space and improve folding.

Next step 2:

We may also improve the 3D model quality by combining our predicted contacts with energy function and fragment assembly. For example, we may feed our predicted contacts to Rosetta to build 3D models. Compared to CNS, Rosetta makes use of energy function and more local structural restraints through fragment assembly and thus, shall result in much better 3D models. Finally, instead of

Distance Prediction and Folding

- Not new but not well studied

Predict distance distribution by neural networks and derive distance-based statistical potential

- Zhao & Xu (2012). *A position-specific distance-dependent statistical potential for protein structure and functional study*, STRUCTURE.
 - Wang (2016). *Knowledge-based machine learning methods for macromolecular 3D structure prediction*, PhD thesis.
 - Zhu, Bu & Xu (2018). *Protein threading using residue co-variation and deep learning*, ISMB (Bioinformatics)
 - A few other groups also studied it
- Discretize distance into 25 bins, each width 0.5Å
 - Use **same ResNet** to predict distance probability distribution for 5 types of atom pairs: $C_b C_b$, $C_a C_a$, $C_g C_g$, NO, $C_a C_g$

Contact Accuracy on 37 CASP12 Hard Targets

Median family size: 58; Five have >1000 sequence homologs

Below table is taken from our CAPS12 invited paper

Method	Long-range				Medium-range			
	<i>L</i>	<i>L/2</i>	<i>L/5</i>	<i>L/10</i>	<i>L</i>	<i>L/2</i>	<i>L/5</i>	<i>L/10</i>
RaptorX submit	28.63	36.42	46.76	51.50	20.89	31.90	43.92	53.72
RaptorX postdict	40.18	50.20	58.87	63.93	25.28	39.52	56.83	66.54
MetaPSICOV standalone	24.72	31.28	41.66	45.34	18.23	26.12	36.27	42.65
MetaPSICOV submit	27.15	34.00	42.48	46.57	19.79	28.59	39.00	45.60
CCMpred	12.91	17.54	21.13	25.42	9.98	13.95	18.86	25.31
Baker GREMLIN	11.24	17.18	23.22	27.00	7.11	9.53	18.06	22.49

In CASP12

Aug/Sept 2016

RaptorX 43.1 56.9 66.9 73.8 (old MSAs, tested right before CASP13)

Folding on 32 CASP13 FM Targets

1. Feed predicted distance and angles into CNS to build 3D models
2. NO fragments, NO energy function, NO folding simulation

Servers	#models	RMSD	TMscore	GDT	#(TM>0.5)
Zhang-Server	32	8.97	0.5239	0.4531	
QUARK	32	8.85	0.5137	0.4423	
RaptorX-DeepModeller	32	9.79	0.5009	0.4313	17
RaptorX-Contact	32	10.09	0.4983	0.4269	17
Baker-Robetta	32	13.02	0.4298	0.3724	
RaptorX-TBM (threading)	32	11.87	0.4206	0.3579	

Best of top 5 is evaluated

DCA vs. Deep ResNet

Direct Coupling Analysis	Deep ResNet
Linear model	Nonlinear
Mainly pairwise relationship	Learn patterns and global conformation context
Tens of millions of parameters	Millions of parameters
Trained by a single family	Trained by thousands of families
Not highly correlated with distance	Fine-grained distance prediction

What if DCA not used?

(Results on CASP12 hard targets)

Contact Pred Methods	L	L/2	L/5	L/10
Deep ResNet (+MI +DCA)	43.1	56.9	66.9	73.8
Deep ResNet (+MI -DCA)	36.3	48.4	61.9	66.7

Folding Methods	Top 1 TM	Top 5 TM	#(TM>0.5)
Our Distance (+MI +DCA)	0.466	0.476	21
Our Distance (+MI -DCA)	0.400	0.411	12
Our Contact (+MI +DCA)	0.354	0.397	10
Baker-server	0.326	0.370	9
Zhang-server	0.347	0.404	10
Baker-human	0.392	0.422	11
Zhang-human	0.375	0.420	11

Why DL Not Much Better in CASP12?

---Not all DLs are equal

Local Methods and/or Not End-to-End Training

- Deep Belief Networks or Feed Forward Networks
 - DNCON (Cheng, 2012); EPSILON-CP (Brock, 2016)
- Iterative or multi-stage learning
 - PConsC2, PConsC3 (Elofsson, 2014); CMAPpro (Baldi, 2012)
- Cannot capture contact patterns or long-range information and thus, do not work well

Global Methods and End-to-End training

- Deep Convolutional (Residual) Neural Networks
 - RaptorX-Contact (Xu, 2016): not fully implemented in CASP12 (although ranked first), but performed well in CAMEO since Oct. 2016
 - Most groups used deep ResNet or CNN in CASP13 since it works!

Summary

- What went right ?
 - Deep ResNet predicts contact/distance very well even for proteins with ~60 sequence homologs
 - Predicted distance can fold many FM targets without folding simulation (21 of 37 CASP12; 17 of 32 CASP13)
 - Distance prediction can greatly improve folding (0.08-0.1 TMscore on CASP12) and slightly contact accuracy (3-4%)
 - With DL, folding on a personal computer is feasible
- What went wrong ?
 - Not very good for small proteins with <50AAs due to too deep
 - Fix it by a shallower model
- Evaluate Distance Prediction in future CASPs ?

Acknowledgements

- RaptorX contact/distance prediction and ab initio folding server at <http://raptorx.uchicago.edu/AbInitioFolding>
- Funding
 - NIH R01GM089753
 - NSF BIO-1564955
- Computational resources
 - Nvidia