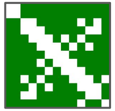
Deep Learning Shines New Hopes on Solving the Half-a-Century-Old Problem of Protein Folding

Badri Adhikari
adhikarib @ umsl.edu

Assistant Professor of CS
Department of Mathematics & Computer Science
University of Missouri-St. Louis

Significance of Protein Contact Prediction

Precise protein contact prediction

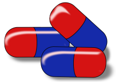


Leads to..

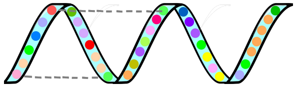
Accurate protein structure / function prediction

Leads to..

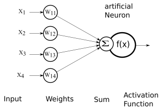
Curing diseases through drug design
(cancer, mental health diseases)



Better understanding of how life works
(through understanding of how proteins work)

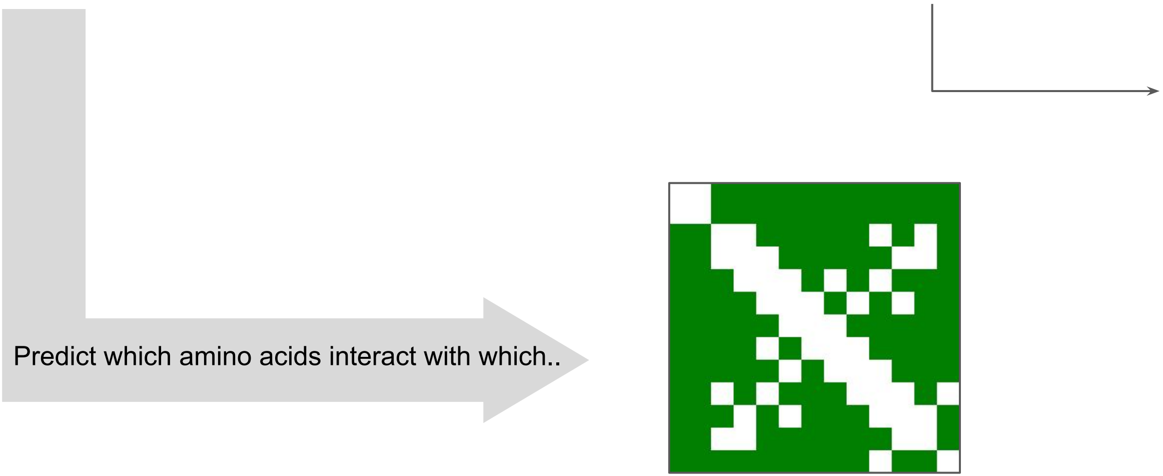


Improvements in Machine / Deep Learning
(because contact prediction is a difficult problem)



What is Protein Contact Prediction?

G F G C N G P W D E D D M

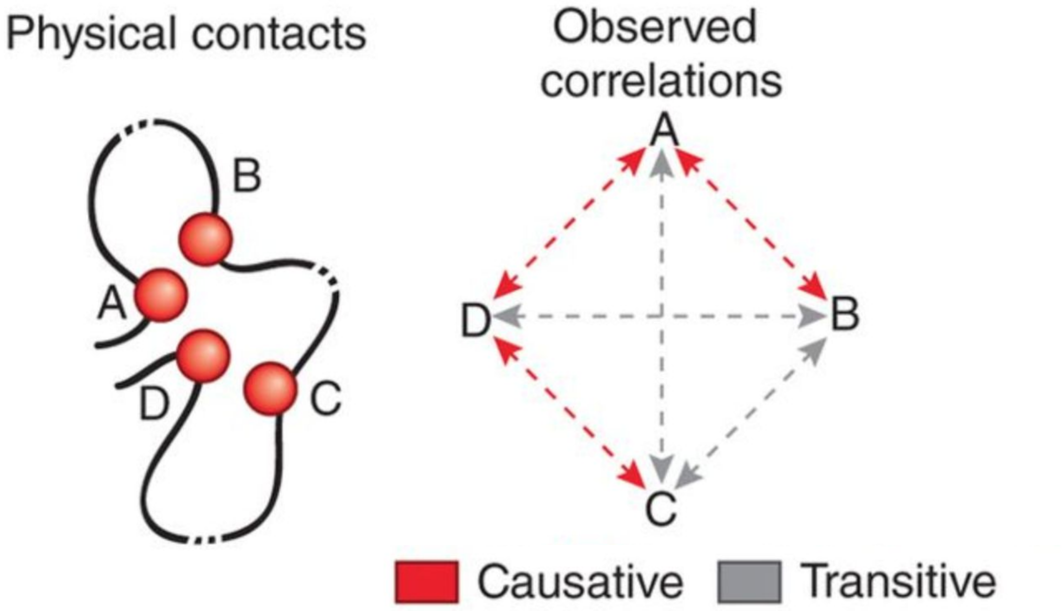


Protein Sequence

Predict which amino acids interact with which..

Contact Map

Transitive Noise - The Roadblocks



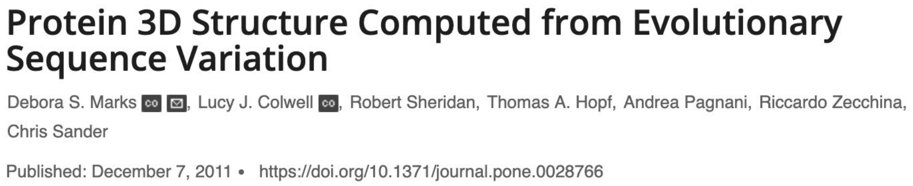
Perspective | Published: 08 November 2012

Protein structure prediction from sequence variation

Debora S Marks, Thomas A Hopf & Chris Sander

Nature Biotechnology **30**, 1072–1080 (2012) | Download Citation

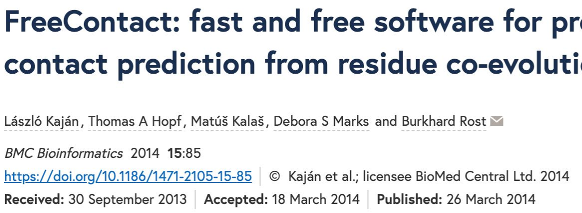
Can We Write Algorithms to Remove Transitive Noise?



Protein 3D Structure Computed from Evolutionary Sequence Variation

Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, Chris Sander

Published: December 7, 2011 • https://doi.org/10.1371/journal.pone.0028766



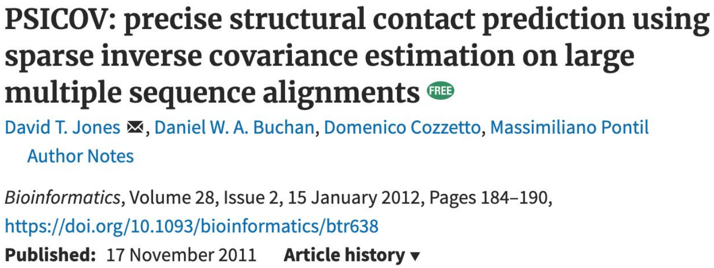
FreeContact: fast and free software for protein contact prediction from residue co-evolution

László Kaján, Thomas A Hopf, Matúš Kalaš, Debora S Marks and Burkhard Rost

BMC Bioinformatics 2014 **15**:85

https://doi.org/10.1186/1471-2105-15-85 | © Kaján et al.; licensee BioMed Central Ltd. 2014

Received: 30 September 2013 | **Accepted:** 18 March 2014 | **Published:** 26 March 2014



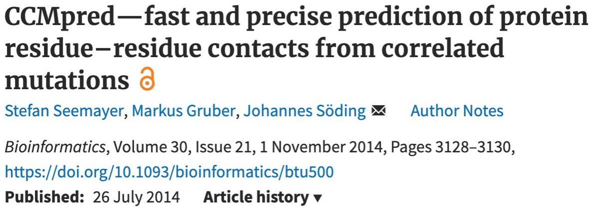
PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments FREE

David T. Jones, Daniel W. A. Buchan, Domenico Cozzetto, Massimiliano Pontil

Author Notes

Bioinformatics, Volume 28, Issue 2, 15 January 2012, Pages 184–190, https://doi.org/10.1093/bioinformatics/btr638

Published: 17 November 2011 **Article history ▾**



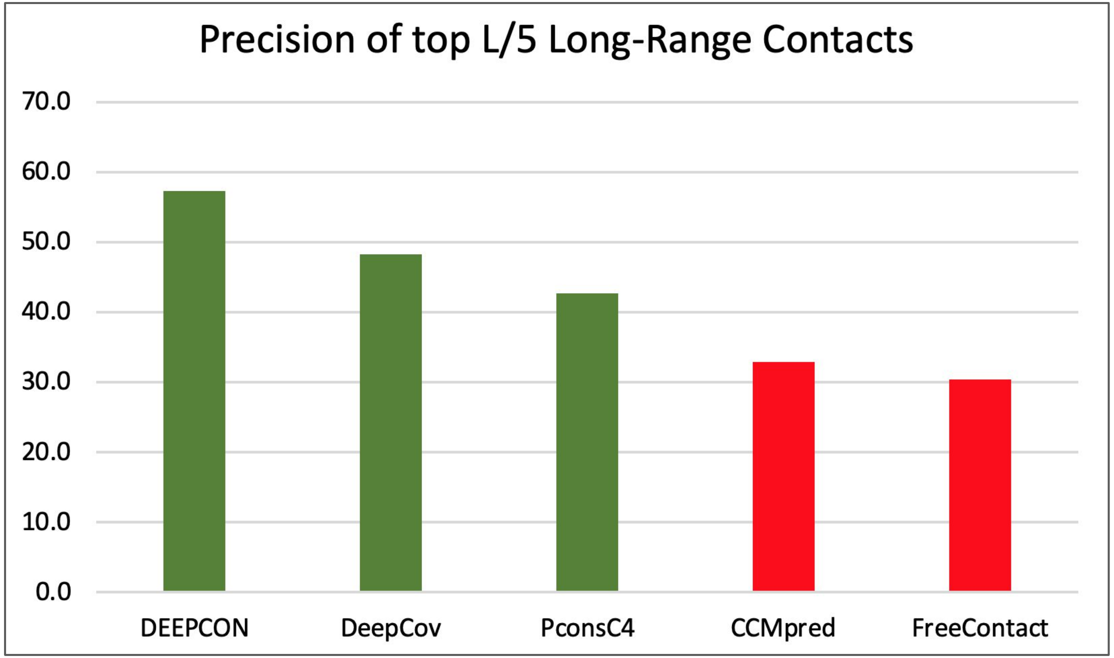
CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations

Stefan Seemayer, Markus Gruber, Johannes Söding Author Notes

Bioinformatics, Volume 30, Issue 21, 1 November 2014, Pages 3128–3130, https://doi.org/10.1093/bioinformatics/btu500

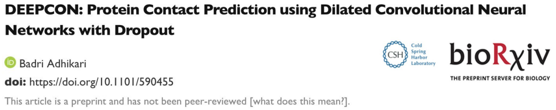
Published: 26 July 2014 **Article history ▾**

Can Deep Learning Remove Transitive Noise?



Deep Learning Methods

Algorithms



DEEPCON: Protein Contact Prediction using Dilated Convolutional Neural Networks with Dropout

Badri Adhikari

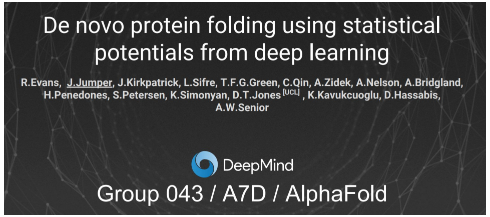
doi: https://doi.org/10.1101/590455

This article is a preprint and has not been peer-reviewed [what does this mean?].

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY
CSH
Cold Spring Harbor Laboratory

What ConvNet Architectures are Best Fit for Contact Prediction?

Top Methods in the most recent CASP Competition

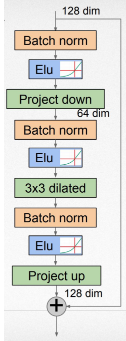


De novo protein folding using statistical potentials from deep learning

R.Evans, J.Jumper, J.Kirkpatrick, L.Sifre, T.F.G.Green, C.Qin, A.Zidek, A.Nelson, A.Bridgland, H.Penedones, S.Petersen, K.Simonyan, D.T.Jones [UCL], K.Kavukcuoglu, D.Hassabis, A.W.Senior

DeepMind

Group 043 / A7D / AlphaFold

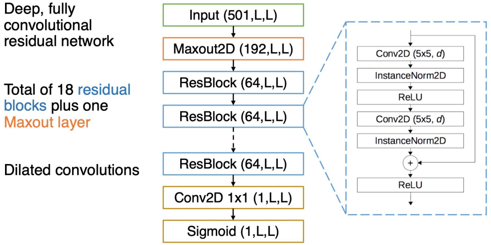


UCL

DeepMetaPSICOV (DMP) in CASP13

Shaun M Kandathil
University College London
&
The Francis Crick Institute

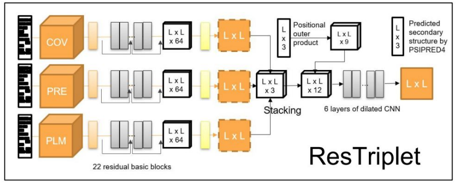
DeepMetaPSICOV model architecture



ResTriplet/TripletRes: Learning contact-maps from a triplet of coevolutionary matrices

Eric W. Bell, Yang Li, Chengxin Zhang, Dong-Jun Yu, Yang Zhang

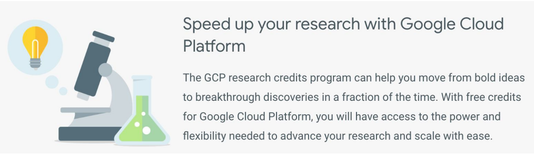
Department of Computational Medicine and Bioinformatics, University of Michigan - Ann Arbor



All these results show that residual networks are best architectures (for this problem)

What Variations of Residual Architectures are Best Fit?

- To obtain an answer we have to try ‘almost’ all possible architectures
  - A lot of computing resources (GPUs)
- The input data for training is [2 GB to 200 GB+]
  - In one epoch (less than 20 minutes) we need to read 200 GB of data
  - In Lewis cluster, training takes at least 10 days (with regular hard-drives)
  - We need SSDs (SATA & M.2)
- Applied to Google for resources
  - $5000 worth of Google Cloud Credits
  - Finished them in less than a week and requested more

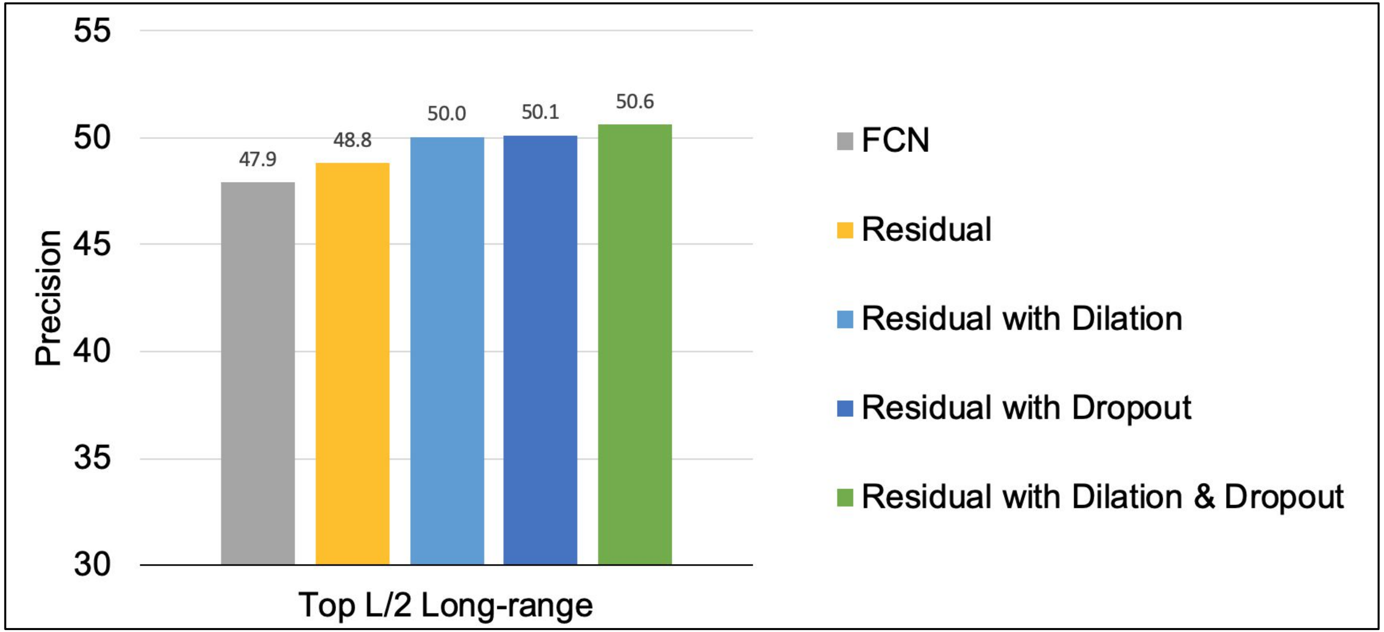


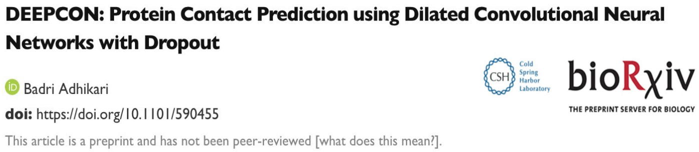
Speed up your research with Google Cloud Platform

The GCP research credits program can help you move from bold ideas to breakthrough discoveries in a fraction of the time. With free credits for Google Cloud Platform, you will have access to the power and flexibility needed to advance your research and scale with ease.

- Applied to NVIDIA for resources
  - Awarded a Quadro P6000 GPU (performs similar to V100s; extremely useful)

Residual Networks with Dilation & Dropout Perform Best




DEEPCON: Protein Contact Prediction using Dilated Convolutional Neural Networks with Dropout

Badri Adhikari

doi: https://doi.org/10.1101/590455

This article is a preprint and has not been peer-reviewed [what does this mean?].

CSH Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

But.. Is There Room for Improvement? YES

> "It was good to see Google DeepMind win this time..
> I was sick of seeing Rosetta win since almost two decades.."

- a senior scientist at the CASP13 conference

- We are still far from end-to-end deep learning
  - Where Deep Learning will do the magic!

Conclusions

1) Groups who were good at exploring ‘new flavors’ did well
   - Learn various deep learning methods, even when you don’t see a direct fit to your problem

2) Balanced efforts of ML experts and domain experts brought success

3) When end-to-end is not possible, correct feature engineering becomes important
   - For example, for standard images, we don’t need feature engineering

Acknowledgements

Research Support & Contribution

Cezary Janikow

Sharlee Climer

Cynthia Jobe

Anthony Ackah-Nyanzu

Patrick Kong

Sri Harsha Akurathi

IT Support

Philip Reiss

Kenneth Voss

Michael Remier

MU - Research Computing Support Services Team (RCSS)

Computing Resources

Google

NVIDIA.

University of Missouri System
COLUMBIA | KANSAS CITY | ROLLA | ST.LOUIS

THANK YOU