UMSL

Machine Learning Methods for Protein Contacts Prediction

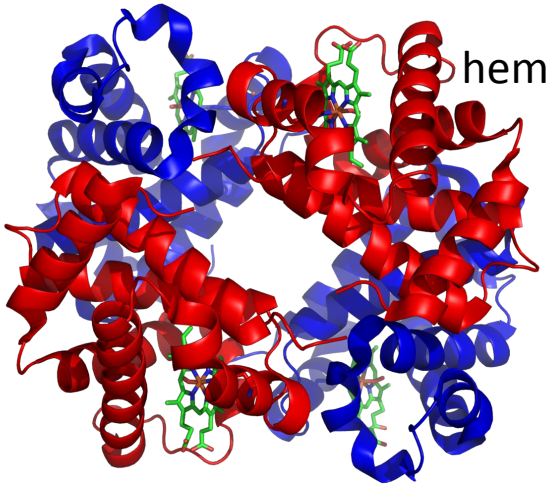
Badri Adhikari, PhD
Mathematics and Computer Science Department
University of Missouri - St. Louis

11 - 16 - 2017

Proteins we eat and proteins inside our body

Proteins we eat

Decompose and build new



hemoglobin

Proteins inside our body

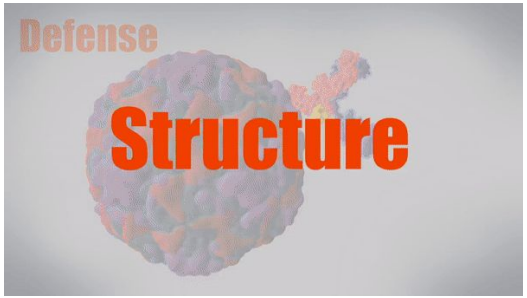
Proteins are made up of amino acids MKTFLYFCLLFIVQTAFAADSIYVREQ

Functions of proteins (some examples)

Recognize and bind to foreign molecules – prevent viral DNA/RNA to enter the cell

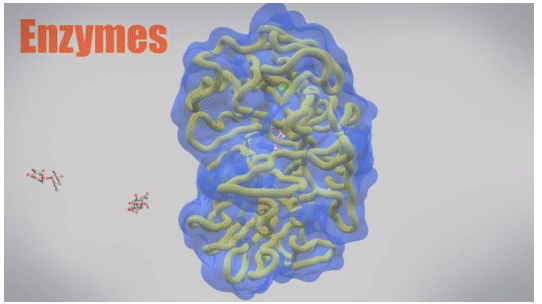
Alpha-amylase protein in saliva helps breaks down carbohydrates

Insulin regulates the blood sugar level



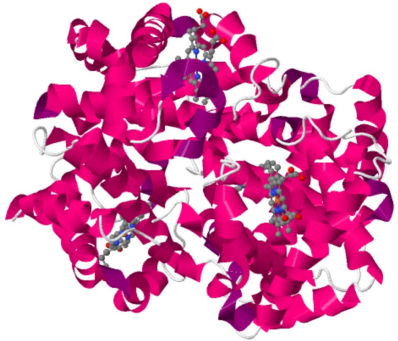
Collagen provides structural support – skin, cartilage, bones, etc.

Calcium pumps allow synchronized muscles contraction

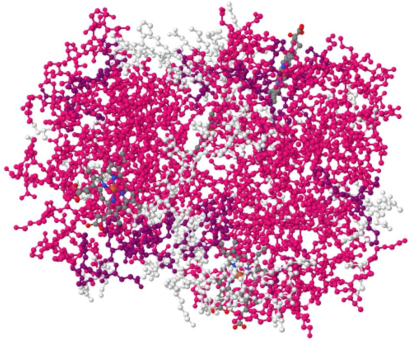


Ferritin forms a hollow shell to store iron from our food

3D structure of Hemoglobin

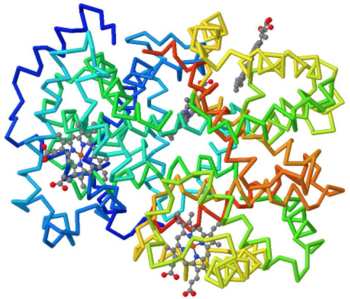


Cartoon representation



All atoms in a protein

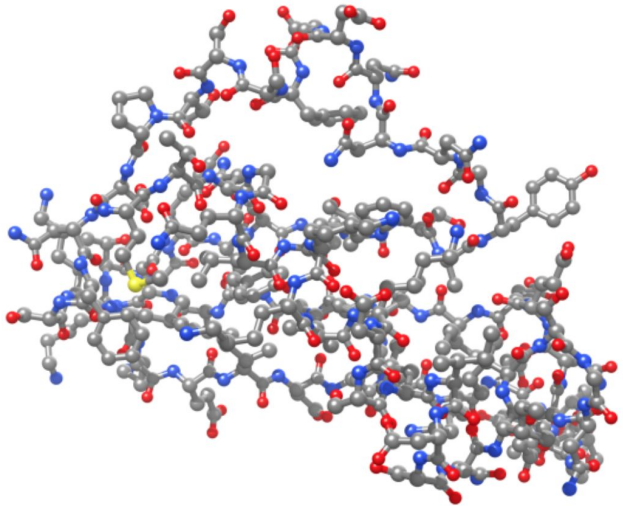
Binding sites



Backbone representation

https://www.rcsb.org/pdb/explore/jmol.do?structureId=1LFZ&opt=3&bionumber=1

Computer representation



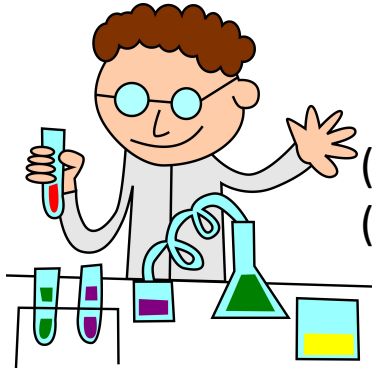
How biologists and chemists see a protein

	Atom		X	Y	Z
1	N	LYS	0.000	0.388	2.413
2	CA	LYS	0.000	0.000	3.819
3	C	LYS	1.210	-0.884	4.102
4	O	LYS	1.454	-1.845	3.376
5	CB	LYS	-1.307	-0.698	4.207
6	H	LYS	0.000	1.357	2.176
7	N	THR	1.967	-0.564	5.155
8	CA	THR	3.147	-1.342	5.515
9	C	THR	3.473	-1.119	6.989
10	O	THR	2.930	-0.206	7.607

How computer scientists see a protein

How are 3D structures determined?

MKTFLYFCLLFIVQTAFAADSIYVREQ → ?



(1) NMR
(2) X-Ray Crystallography

Structure DB
(~ 100K proteins)
Publicly available

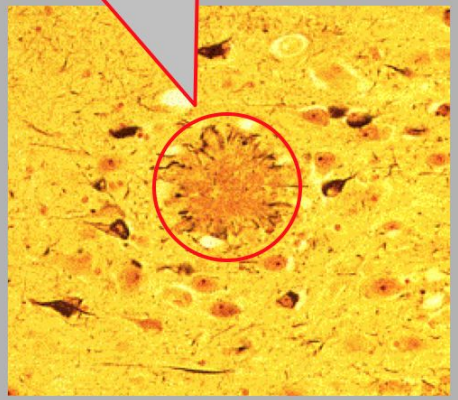
- Accurate
- Takes months to years
- Expensive

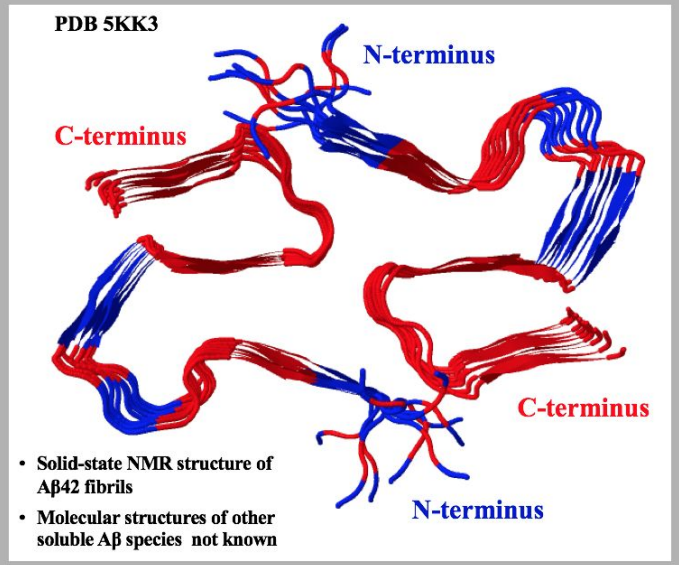
Protein oligomerization in Alzheimer's disease

Amyloid-β protein/peptide (Aβ) hydrophilic hydrophobic

H_2N-DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA-COOH

Aβ fibers/fibrils are main component of senile plaques in AD brains





MT Colvin et al., *J Amer Chem Soc*, 2016, 138, 9663

"**These fibrils are of considerable medical importance** because they are **associated with more than 40 different diseases** including Alzheimer's disease."

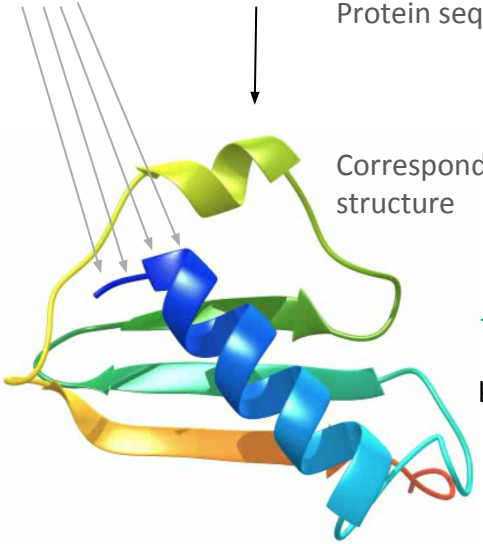
Dr. Michael Nichols
Department of Chemistry and Biochemistry
(UMSL)

" .. my laboratory involves mechanistic studies of Aβ aggregation .. "

What are protein contacts?

MKTFLYFCLLFIVQTA ... FAADSIYVREQ

Protein sequence



Corresponding structure

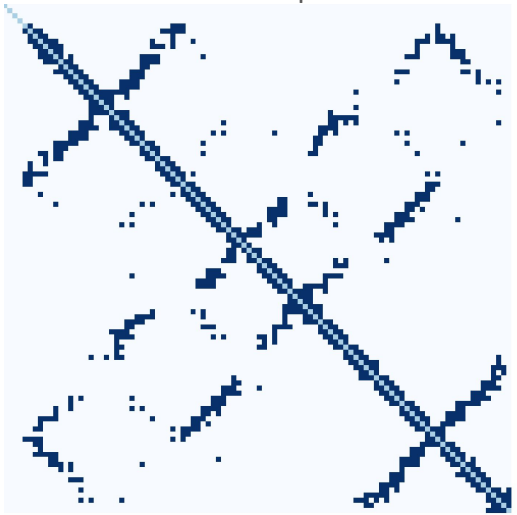
3D Structure

Calculate pairwise distance

Distance matrix

if $d < 8Å$ $c_{ij} = 1$

Dark spots = Contacts

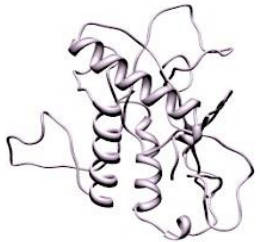


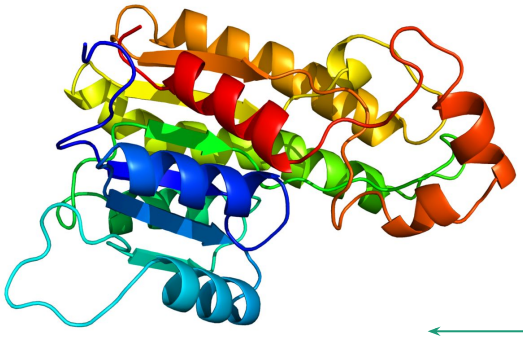
Contact Map (**or Contacts**)

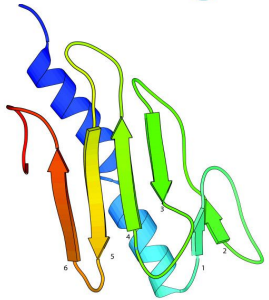
3D

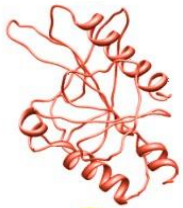
2D

Protein and contact map sizes

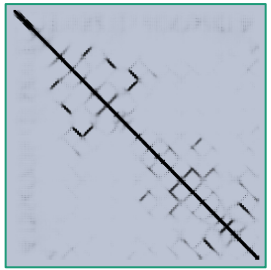


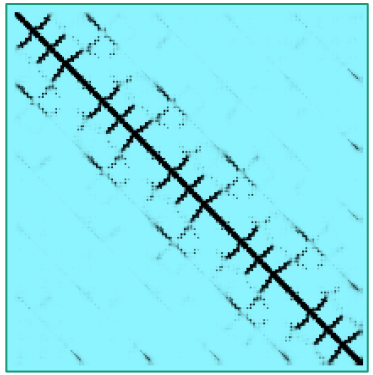


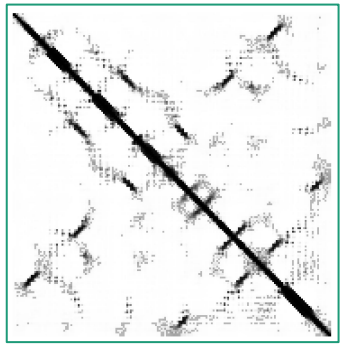


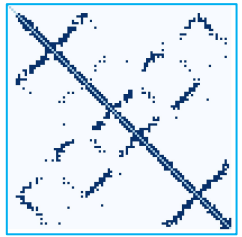


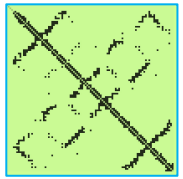
1-to-1 relationship
(mostly)





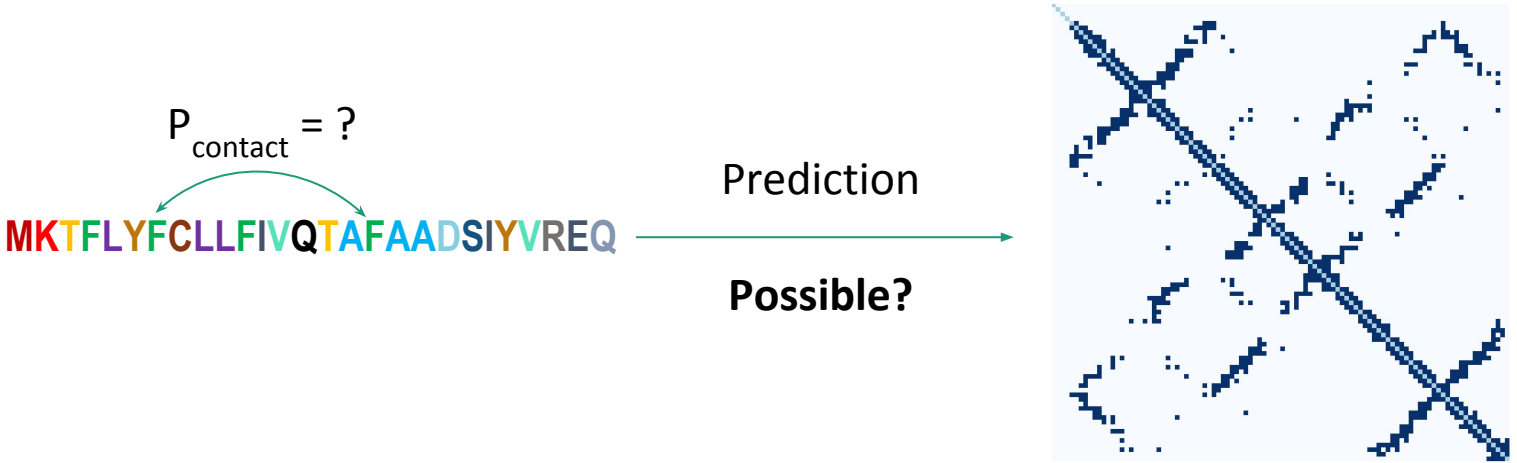






Can we rescale structures or maps?

What is protein contact prediction?



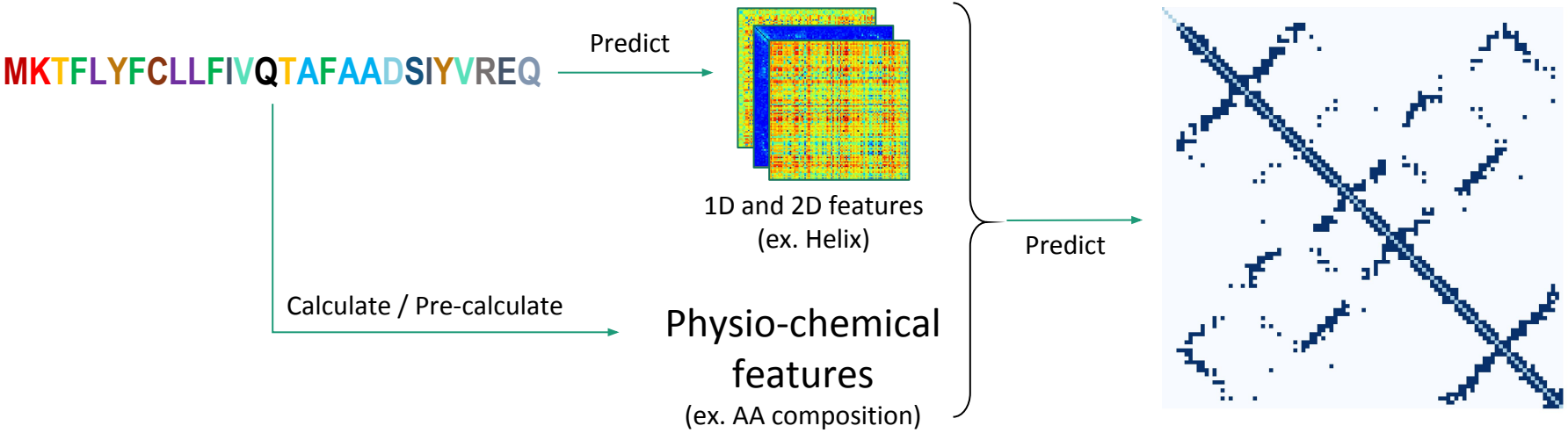
Protein sequence

Contact Map (contacts)

1D

2D

Features and Data for Contacts prediction

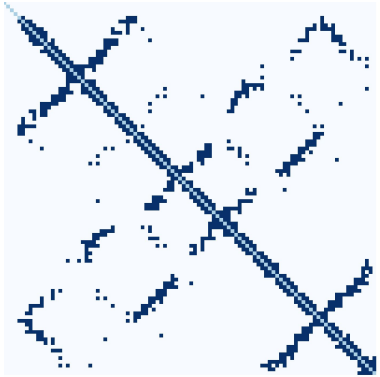


Significance of Contact prediction

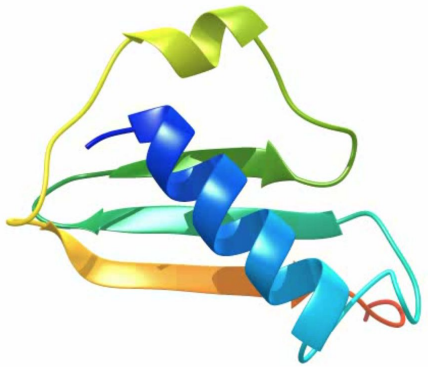
1D 2D 3D

MKTFLYFCLLFIVQTAFAADSIYVREQ

Predict



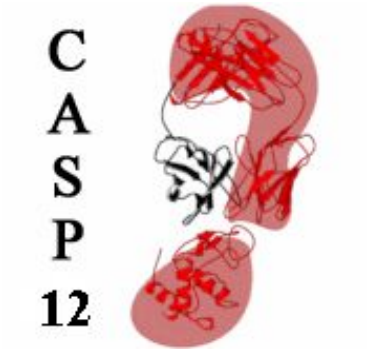
3D modeling



Sequence Contacts Structure

- Save scientists' time
- Save money (millions)
- Drug design

How accurately can we predict structures today?



World-wide competition held every two years (3 months long)


vs


most recent competition

root mean square deviation

a top participant

Competition: CASP12 (2016)

Predictor: Baker-Rosetta (UW)

Protein		RMSD		
Type	**Count**	**Best**	**Median**	**Worst**
Template-based	57	0.69	4.7	24.2
Template-free	58	2.04	12.9	22.8

dataset

99% similarity (experimental biologists' are happy)

random prediction

What is Machine Learning (ML) ?

Glossary of AI terms

ARTIFICIAL INTELLIGENCE

Enabling computers to mimic human intelligence

MACHINE LEARNING

Enabling machines to improve at tasks with experience (Statistical techniques)

DEEP LEARNING

Many layers of neural networks (vast amount of data)

Data (Experience)

ex. face database

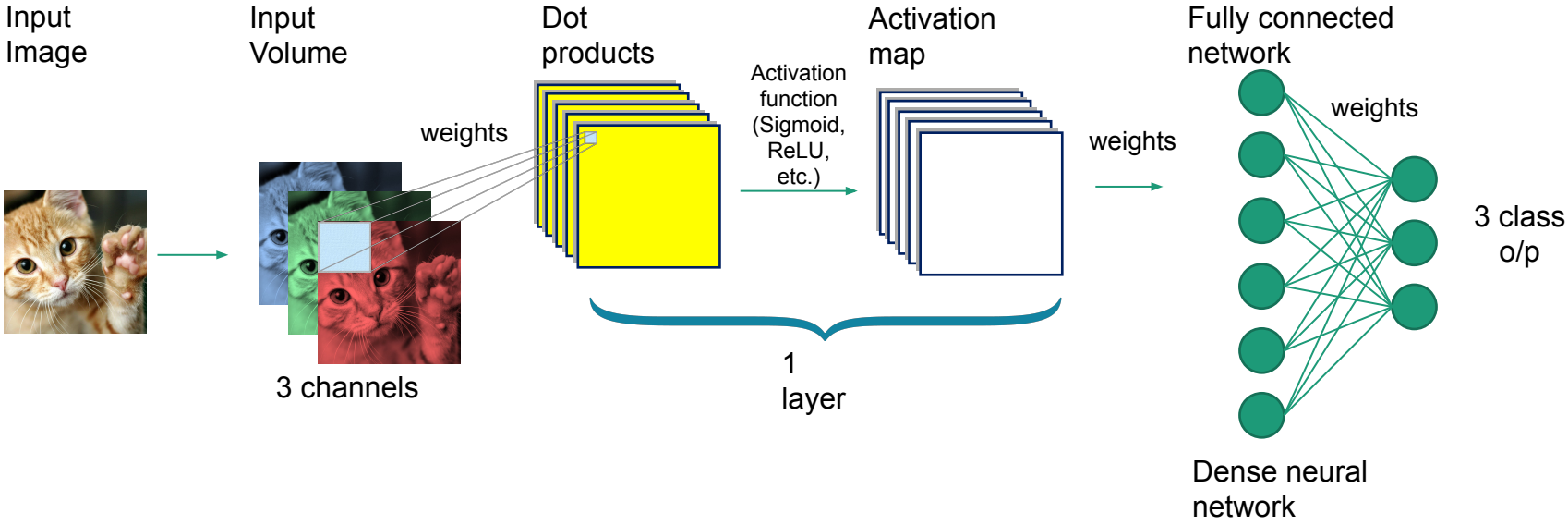
training

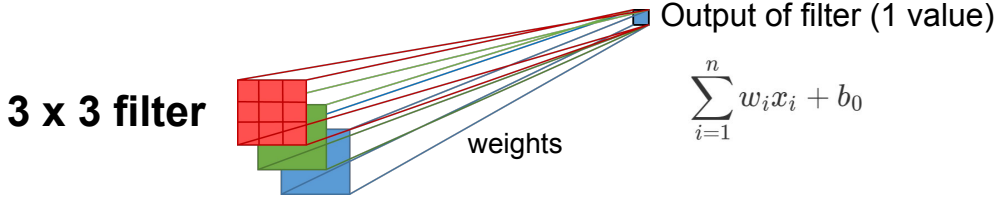
Examples: Neural Networks (NN), Support Vector Machines (SVM), Random Forests, etc.

Examples: Restricted Boltzmann Machines, Deep Neural Networks, Deep RNNs, Deep CNNs, etc.

Technologies: Batch normalization, Dropouts, Residual neural networks, etc.

Example: Convolutional Neural Network for Image Classification





Is ML applicable to problems in biology and medicine?

Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks

P Lakhani, B Sundaram - Radiology, 2017 - pubs.rsna.org

"accurately classify TB at chest radiography with an AUC of 0.99"

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

Nature **542**, 115–118 (02 February 2017)

doi:10.1038/nature21056

"artificial intelligence is capable of classifying skin cancer with a level of competence comparable to dermatologists" (performance tested against 21 board-certified dermatologists on biopsy-proven clinical images)

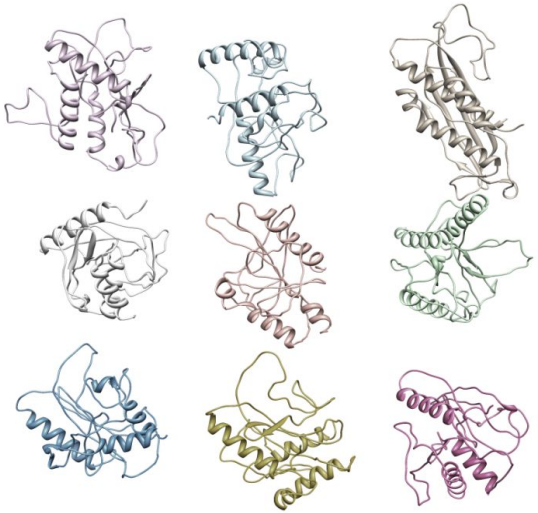
CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng

(Submitted on 14 Nov 2017)

Diagnose pneumonia from chest X-rays better than radiologists

Contact prediction as a ML problem



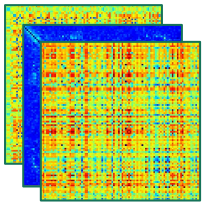
PDB DB

(~100K structures reduced to < 5K)

Sequences

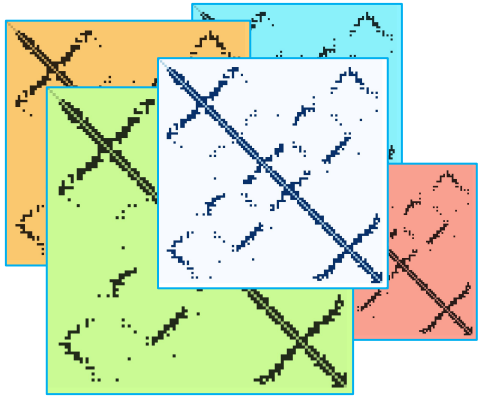
YFCLLFIVQTAFAAD
FIVQTAFAADSIYVREQ
TFLYFCLLFIVQTAFAAD
YFCLLFIVQTAFAAD

Prediction



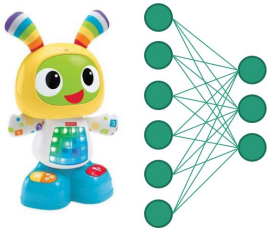
1D and 2D features

(1) Make prediction



Contact maps

(2) Calculate error



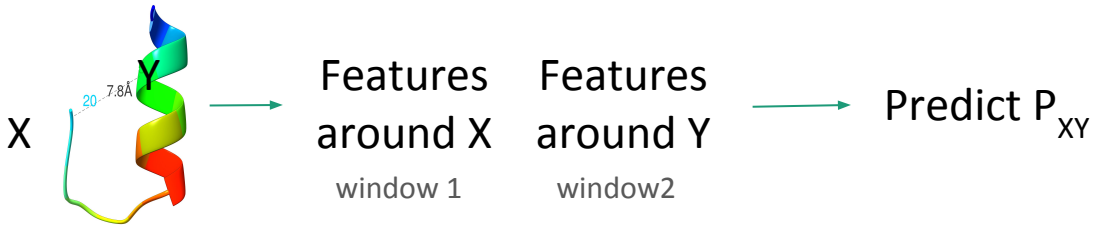
Binary cross-entropy

$$-\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)\right]$$

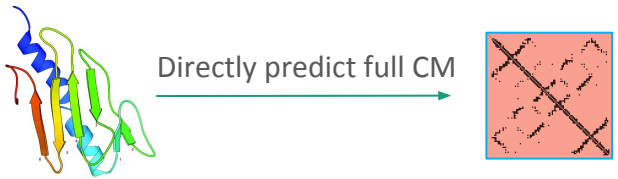
Current methods for Contact prediction

(1) Correlated mutation based methods
- algorithm based and non ML methods
- used as features

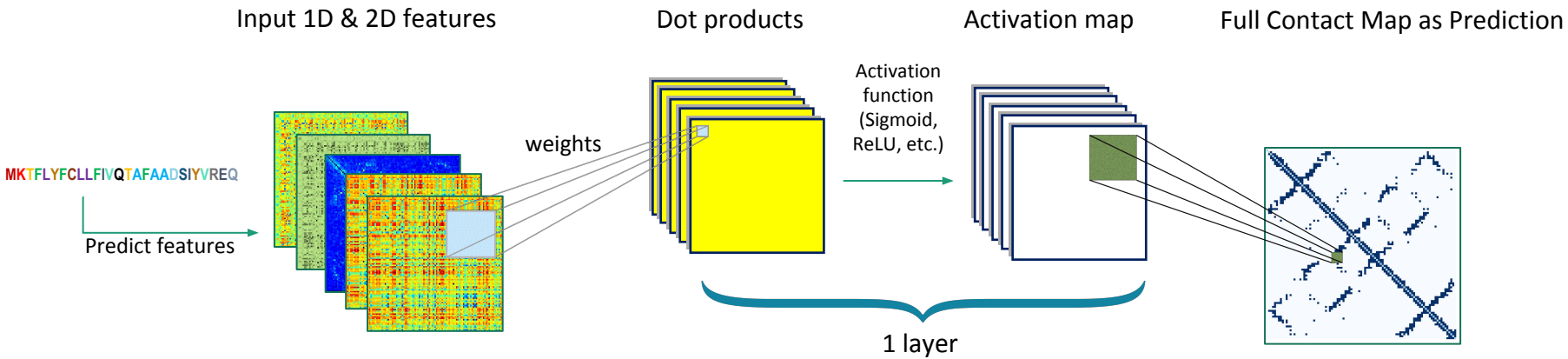
(2) Window-based methods
(DNCON and MetaPSICOV)



(3) Deep CNN-based methods (last 1 year)
(DNCON2 method)



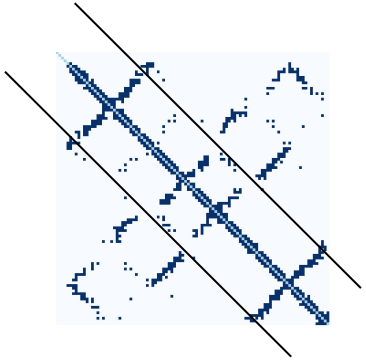
Convolutional Neural Network for Contact Prediction

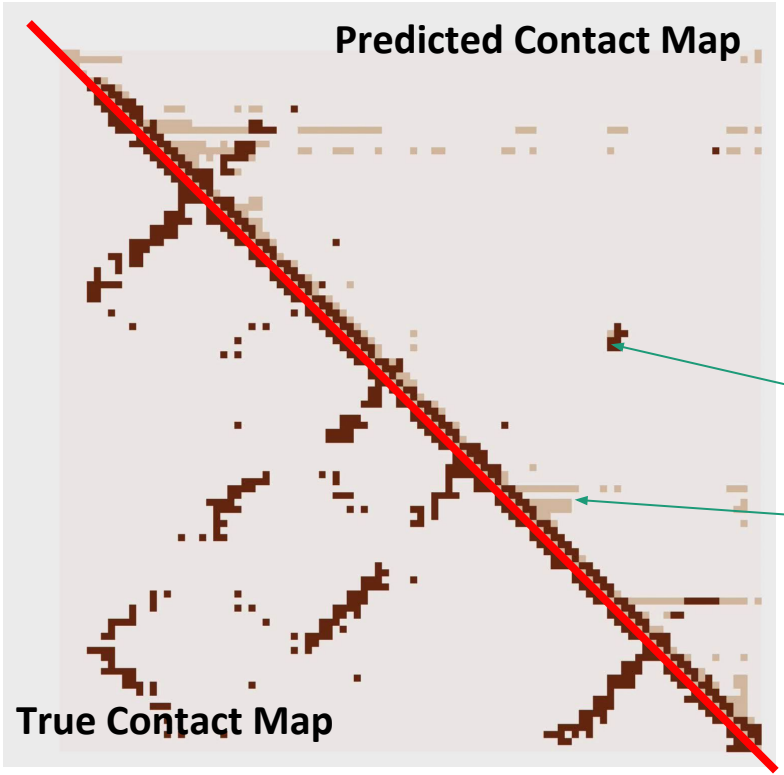


Additional challenges (in comparison to image classification problems):

- Varied and Large input volume
- Highly unbalanced data
- Limited data
- Features are predicted
- Visualization
- Actual contact map as output

An example of CNN learning and predicting





Only top N_c contacts are evaluated

Prediction over increasing epochs

Datasets

Training and Validation

- Used the dataset in original DNCON method
- 1426 proteins from the PDB (0-2 Å resolution)
- L = 30 to 300
- 1230 training proteins
- 196 validation proteins
- The two sets have less than 25% sequence identity

Testing

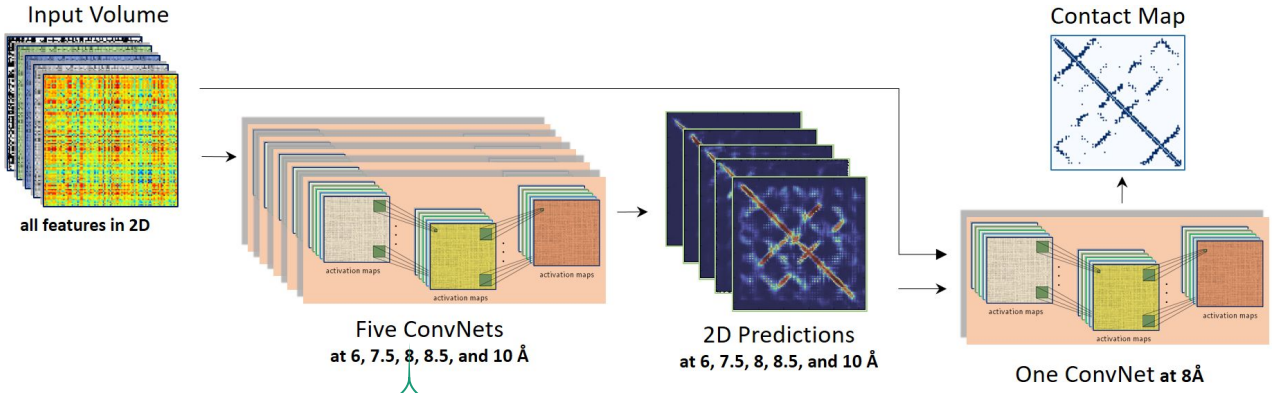
- CASP10, 11, and 12 free-modeling (FM) datasets

All features used

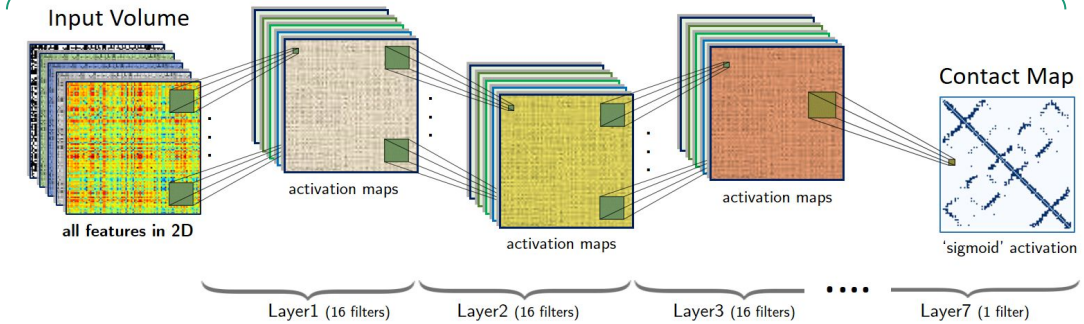
Feature	Number of Dimensions	Number of Features	Number of 2D Features
Log of sequence length	Scalar	1	1
Log of number of sequences in the alignment	Scalar	1	1
Log of N_{eff}	Scalar	1	1
Ratio of number of 'buried' residues to L (SCRATCH)	Scalar	1	1
Ratio of number of 'E' (and 'H') residues to L (SCRATCH)	Scalar	2	2
Binary predictions for H, C, and E residues (SCRATCH)	1D	3	6
Solvent accessibility - 'buried' flag (SCRATCH)	1D	1	2
PSSM and Atchley features	1D	8	16
Sequence separation information	2D	5	5
Pre-computed statistical potentials	2D	6	6
Probabilities of PSIPRED predictions for H, C, E residues	1D	3	6
Probabilities of PSISOLV predictions for solvent accessibility	1D	1	2
Shannon entropy sum of the alignment columns	2D	1	1
CCMpred prediction	2D	1	1
FreeContact prediction	2D	1	1
PSICOV prediction	2D	1	1
Alignment related features (from alignment)	2D	3	3
Total		**40**	**56**

Block diagram and CNN architecture

Overall architecture



Architecture of one CNN



Python Code create a Deep CNN Architecture

```
def build_model_for_this_input_shape(model_arch, X):
        model = Sequential()
        for layer in range(1, 1000):
                if not model_arch.has_key('layer' + str(layer)):
                        break
                parameters = model_arch['layer' + str(layer)]
                cols = parameters.split()
                num_kernels = int(cols[0])
                filter_size = int(cols[1])
                b_norm_flag = cols[2]
                activ_funct = cols[3]
                if layer == 1:
                        model.add(Convolution2D(num_kernels, filter_size, filter_size, border_mode='same', input_shape=X[0, :,
                else:
                        model.add(Convolution2D(num_kernels, filter_size, filter_size, border_mode='same'))
                if b_norm_flag == '1':
                        model.add(BatchNormalization())
                model.add(Activation(activ_funct))
        model.add(Flatten())
        return model
```

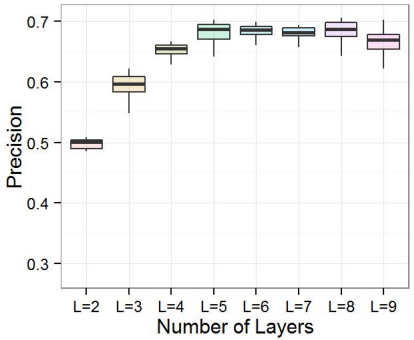
```
# Model architecture description
# Layer Number-of-kernels Filter-size Batch-norm-flag Activation-function
layer0 16 5 1 relu
layer1 16 5 1 relu
layer2 16 5 1 relu
layer3 16 5 1 relu
layer4 16 5 1 relu
layer5 16 5 1 relu
layer6  1 5 0 sigmoid
```

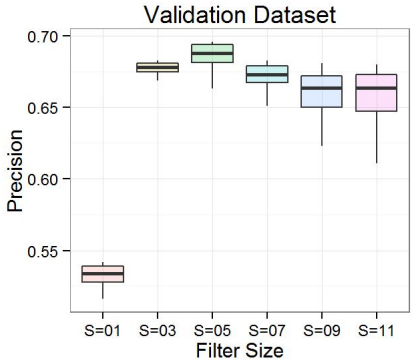
https://github.com/multicom-toolbox/DNCON2/

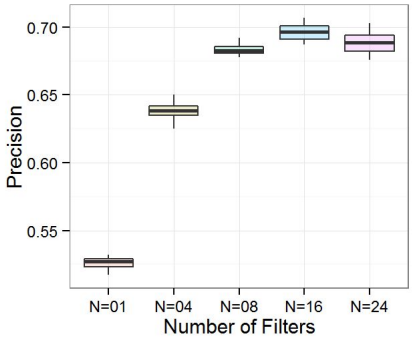
Hyper-parameters Optimization

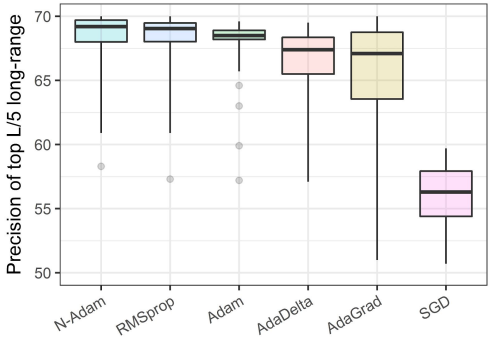
Create and test various brain structures and neural connections

- Depth of the network = 7
- Filter sizes in each layer = 5
- Number of filters in each layer = 16
- Batch normalization = at each layer
- Batch size = 30
- Optimization function = Nesterov Adam (NAdam)
- Activation function = Rectified Linear Units (ReLU)









Comparison with the state-of-the-art

FM Dataset	Domain Count	Precision of top L/5 long-range contacts (%)		
		Top CASP Group	MetaPSICOV	DNCON2
CASP10	15	18.1 (DNCON 1.0)	30.6	35.0
CASP11	30	29.7 (CONSIP2)	34.4	50.0
CASP12	37	46.3 (Raptor-X)	42.9	53.4

Most difficult dataset

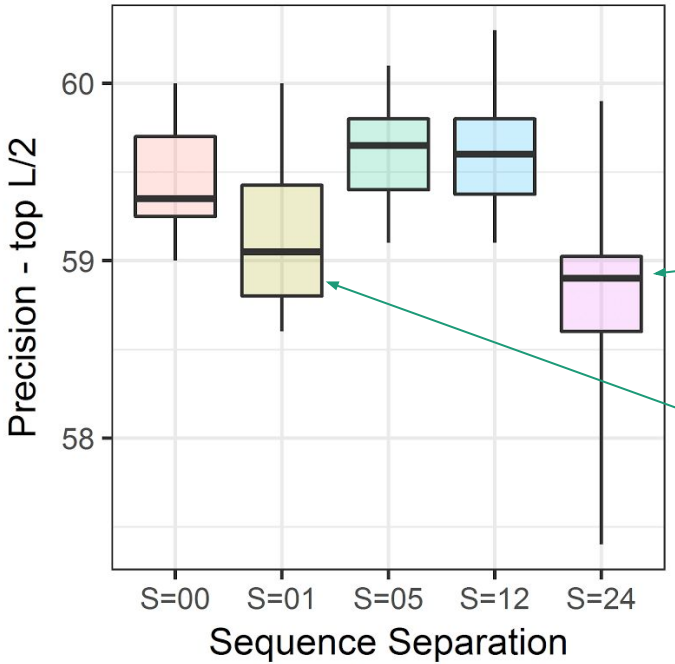
Number of proteins

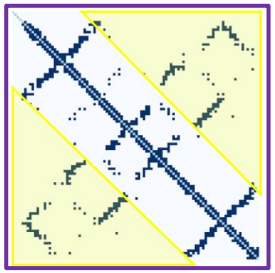
Results of the 2016 CASP competition

A benchmark method that uses similar features but NN (not CNN)

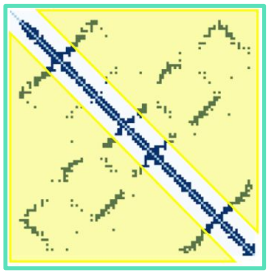
My method

Short- and medium-range HELP long-range prediction

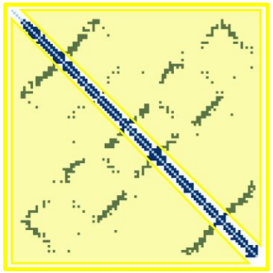




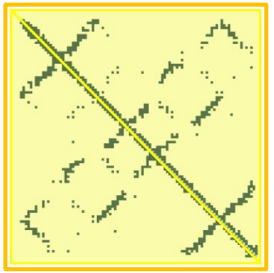
Training using
Long-range only



Training using
Short-, Medium-, and Long-range

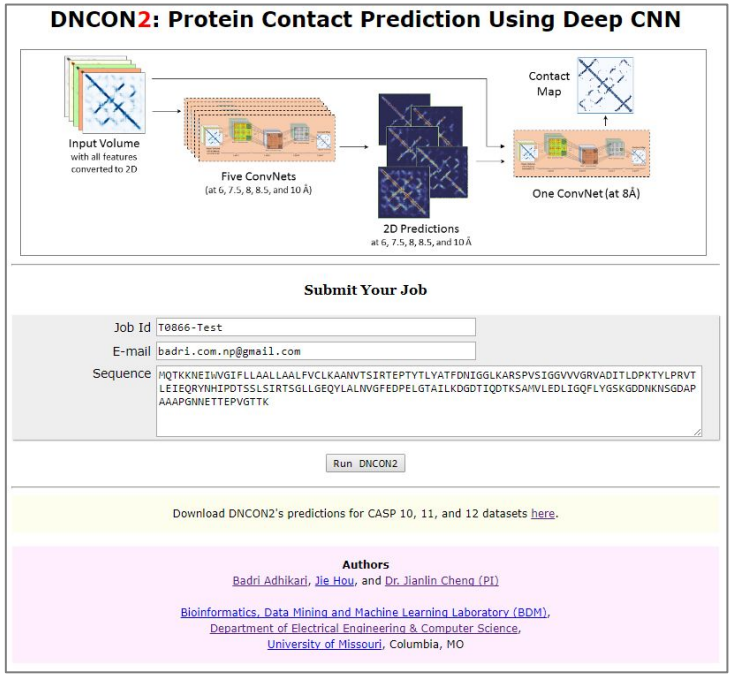


Training using all
except the diagonal

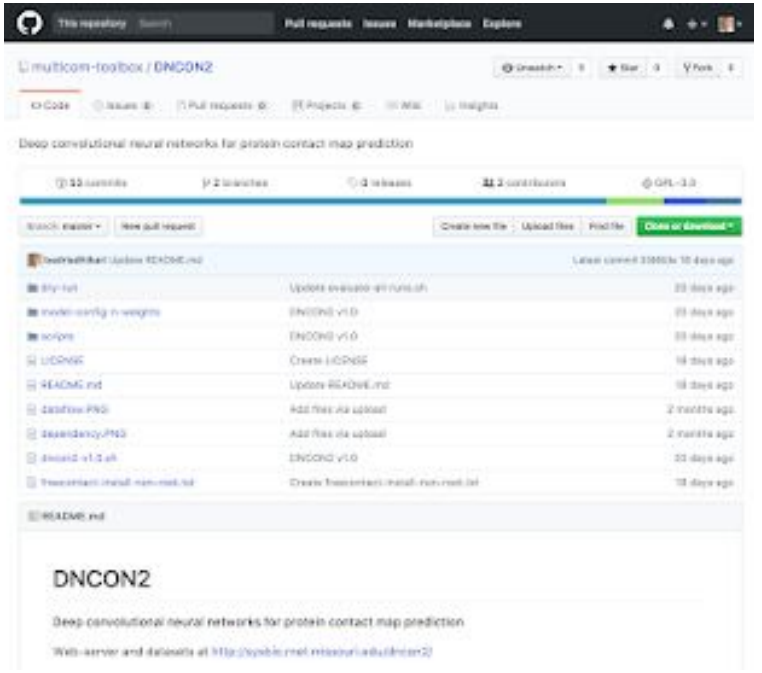


Training using everything

Availability of DNCON2



http://sysbio.rnet.missouri.edu/dncon2/



https://github.com/multicom-toolbox/DNCON2/

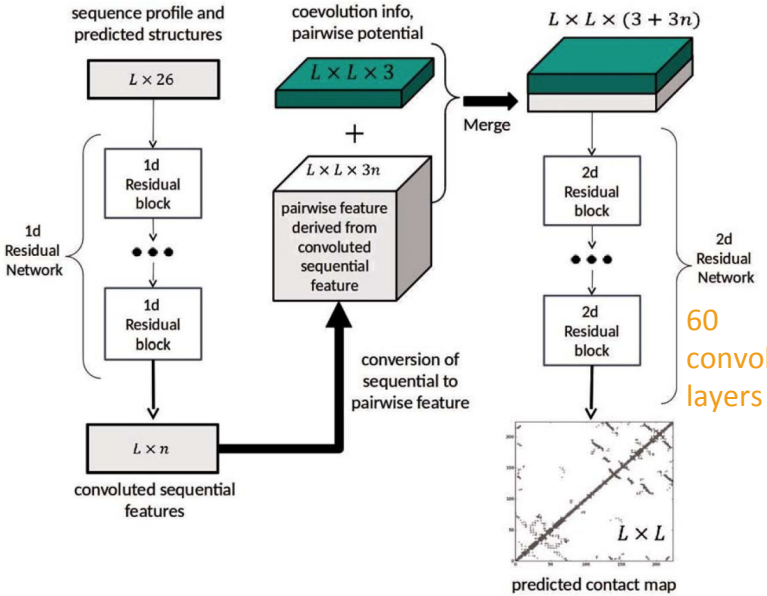
Re-submitted
(after a major revision)

Structural Bioinformatics

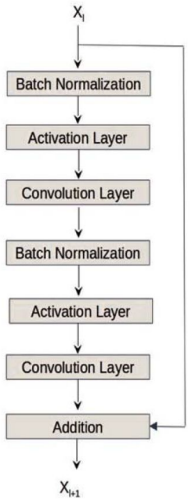
DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks

Badri Adhikari, Jie Hou, and Jianlin Cheng

Raptor-X



60 convolution layers



Method	Long	
	L/5	L/2
EVfold	0.68	0.53
PSICOV	0.70	0.52
CCMpred	0.76	0.62
plmDCA	0.76	0.61
Gremlin	0.76	0.63
MetaPSICOV	0.87	0.74
Our method	0.96	0.89

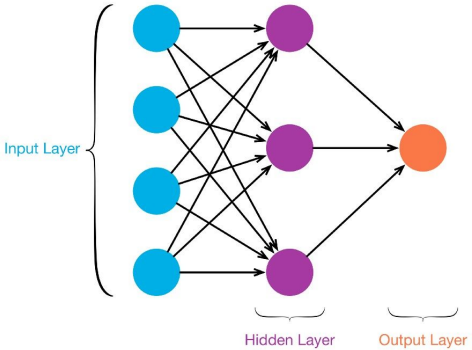
Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model

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

Toyota Technological Institute at Chicago, Chicago, Illinois, United States of America

MetaPSICOV

- classic feed-forward neural networks
- with 55 hidden units and a single output unit



Precision of L/5 long –range contacts

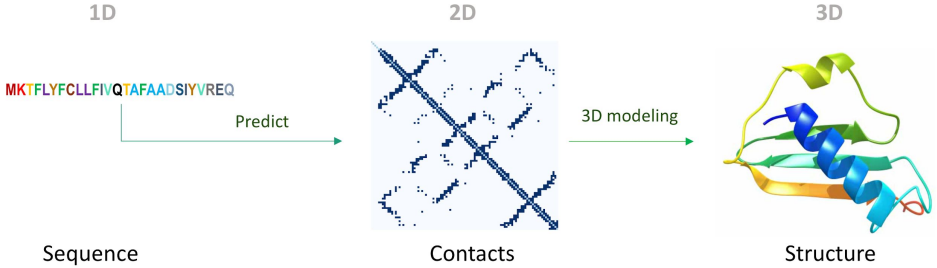
PSICOV	0.65
DCA	0.64
CCMpred	0.71
Consensus only	0.70
PconsC	0.75
MetaPSICOV (stage 1)	0.78
MetaPSICOV (stage 2)	0.83

MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins

David T. Jones, Tanya Singh, Tomasz Kosciolek, Stuart Tetchner

Bioinformatics, Volume 31, Issue 7, 1 April 2015, Pages 999–1006, https://doi.org/10.1093/bioinformatics/btu791

Some Current and Future projects



Publications and Book Chapters:
- DNCON2 / MULTICOM-CLUSTER
- CONEVA
- CONFOLD
- Reconstruction studies

(1) Extend DNCON2 and CONFOLD for membrane proteins' structure prediction

(2) Predict contacts at various distance thresholds and study improvement in protein modeling (Deep CNN problem)

(3) Study how recent deep learning technologies like Boosting, bagging, residual networks, deeper networks will improve the precision (Deep CNN problem)

(4) Study what CNN layers learn! (Deep CNN problem)

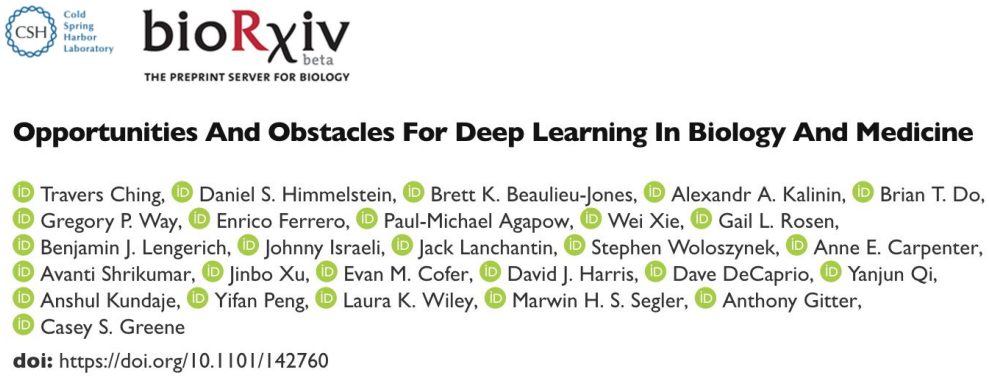
Interested undergraduate/graduate students are welcome join!

Summary

- Protein structure prediction using computational methods has high stakes
- Contact prediction is at the heart of the protein structure prediction problem (a five decade old problem)
- Almost all machine learning methods have been applied to solve the contact prediction problem
- Results show that convolutional neural networks are the future for solving the problem

Conclusion

"We are extremely **optimistic about the future of deep learning in biology and medicine**. It is by no means inevitable that **deep learning will revolutionize these domains**, but given how rapidly the field is evolving, we are confident that **its full potential in biomedicine has not been explored**."



Thank you Dr. Clingher
and
Thank you ALL