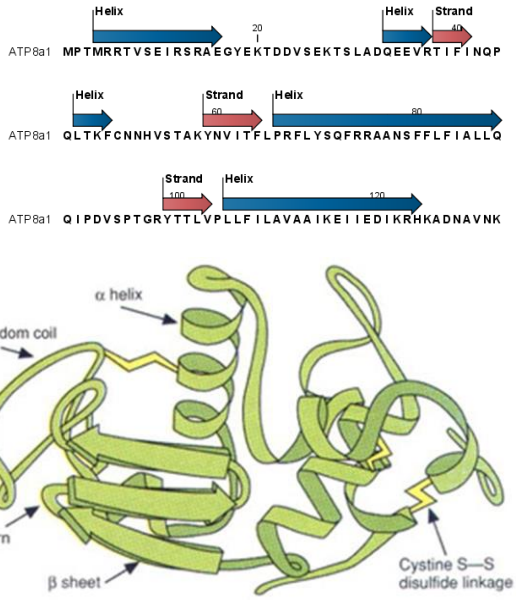
Novel Techniques

E.M. Bakker

Secondary Structure Prediction



8-class (Q8)	3 class (Q3)	Name
H	H	α-helix
E	E	β-strand
L	C	loop or irregular
T	C	β-turn
S	C	bend
G	H	3_{10}-helix
B	E	β-bridge
I	C	π-helix

K. Paliwal, J. Lyons, R. Heffernan A Short Review of Deep Learning Neural Networks in Protein Structure Prediction Problems, Adv Tech Biol Med 3:3, 2015.

Deep Neural Network Architectures

Lot's of available data

Data to be normalized

Various deep neural network architectures [3]:

- deep feed-forward neural networks
- recurrent neural networks
- neural Turing machines
- memory networks.

Feed Forward Neural Networks

Many current state-of-the-art protein predictors are based on feedforward DNNs

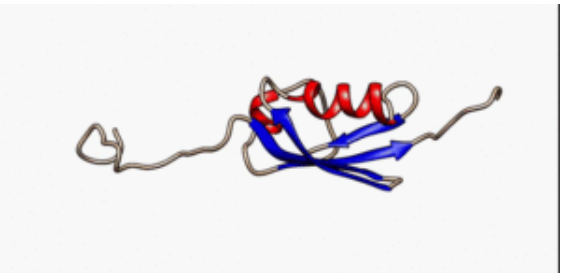
- use a fixed-width window of amino acids, centered on the predicted residue.
- The window is moved over the protein so that predictions can be made for each residue.

Examples:

- PSIPRED (1999) a protein secondary structure predictor based on a neural network with a single hidden layer
  - accuracies of around 80% when predicting 3 states (Q3): helix, coil and sheet.
- SPINE-X, Scorpion, DNSS and SPIDER-2 use deeper neural networks and increased accuracy to ~82% for the 3 states (Q3).
- Deep neural networks for more states (e.g. 8 state (Q8), etc.) helix coil, sheet, Accessible Surface Area (ASA), phi and psi angles, theta, tau angles, and disorder prediction.

Feed Forward Neural Networks

- Deep neural networks for more states e.g. 8 state (Q8) than helix coil, sheet (Q3).
- Accessible Surface Area (ASA)
- Phi and psi angles, theta, tau angles
- Protein Disorder prediction.



Other DNN Architectures

Recurrent Neural Networks (RNNs)

Used for time series, are also used for sequence prediction problems:

- Secondary structure prediction
- Protein disorder prediction
- 2-D RNNs for protein contact map prediction (for every pair of residues in a protein)
- Prediction of disulfide bridges.

- Pass information from one time step to the next.
- Context information contained earlier in the sequence can be utilized later in the sequence.
- Bidirectional Recurrent Neural Networks (BRNNs)
  - Can utilize information along the entire sequence.

RNNs with Memory

RNN's

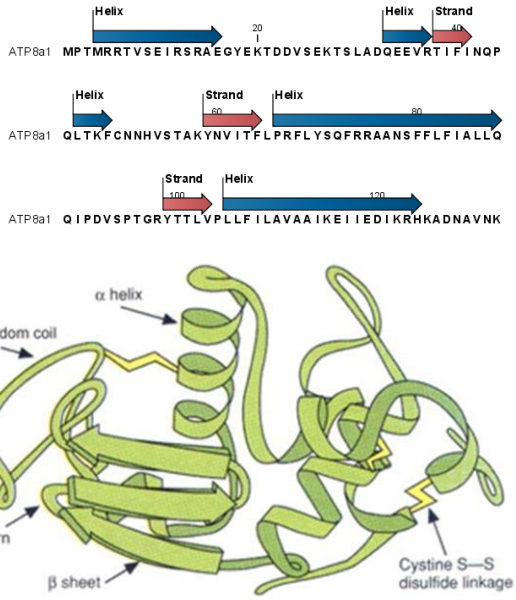
- Can remember information over longer time periods.
- Are widely used for sequence prediction tasks.

Long Short Term Memory (LSTM) RNN's and BRNN's, etc.

- These networks can be trained to solve problems that basic RNNs are incapable of solving, e.g., given examples of sorted and unsorted data, learn to sort new unseen data.
- These architectures have not yet been applied to protein prediction problems (2015) and it remains to be seen whether they will be able to succeed where simpler architectures have not.

S.K. Sønderby, O. Winther, Protein Secondary Structure Prediction with Long Short Term Memory Networks, arXiv:1412.7828v2 [q-bio.QM] 4 Jan 2015. [7]

Secondary Structure Prediction



8-class (Q8)	3 class (Q3)	Name
H	H	α-helix
E	E	β-strand
L	C	loop or irregular
T	C	β-turn
S	C	bend
G	H	3_{10}-helix
B	E	β-bridge
I	C	π-helix

Secondary Structure Prediction

Traditionally

- non-sequential models, typically feed-forward neural networks or SVM's [Hua & Sun, 2001; Jones, 1999].
- Models originally for classifying fixed dimensional vector data: for sequences/streams => sliding window are used
- These methods only learn dependencies within the input window

Recent methods (2014 -)

- learning other dependencies beyond window size
- Conditional random field hybrid models
- RNN's can be applied to sequential data of any length => able to learn long-term dependencies.

Secondary Structure Prediction

But RNN have some problems:

- Exploding or vanishing gradients [Bengio et al., 1994]
- Baldi et al. 1999: their RNN's still were only able to learn dependencies of 15 amino acids relative to the target

Solution to RNN Problem

- Long Short Term Memory (LSTM) RNN's [Graves, 2012] against the vanishing gradients problem
- LSTM networks able to learn dependencies over 100's of time steps.

S.K. Sønderby, O. Winther [7] propose a bidirectional LSTM network for protein secondary structure prediction.

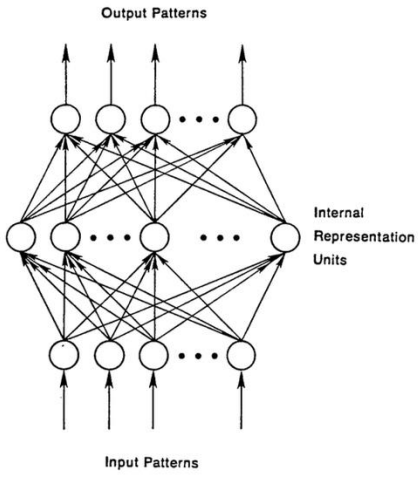
Deep Neural Networks for Protein Structure Prediction

- DNN's state-of-the-art in speech recognition, image recognition, natural language processing tasks, games such a chess and go, etc.

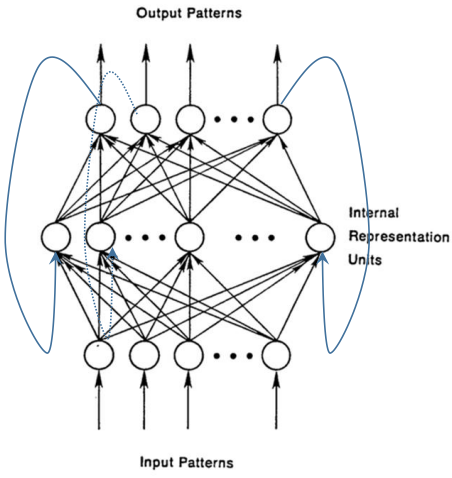
Various deep neural network architectures [3]:

- deep feed-forward neural networks (DNN)
- recurrent neural networks (RNN)
- …
- Long Short Term Memory (LSTM) networks.

Secondary Structure Prediction



Feed Forward Neural Network



Recurrent Neural Network

DNN: AlexNet, VGG16, ResNet, etc.

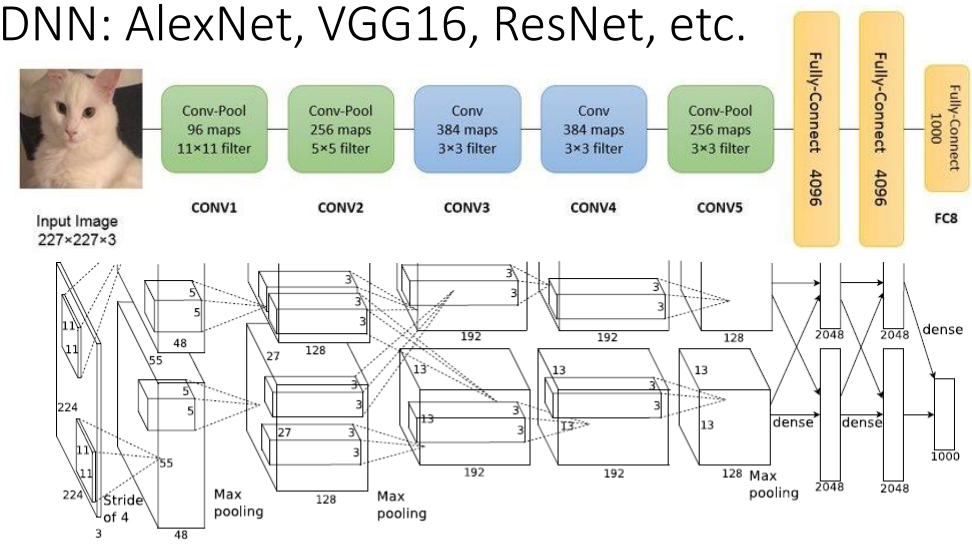


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

Deep Visualization Toolbox

yosinski.com/deepvis

#deepvis


Jason Yosinski


Jeff Clune


Anh Nguyen


Thomas Fuchs

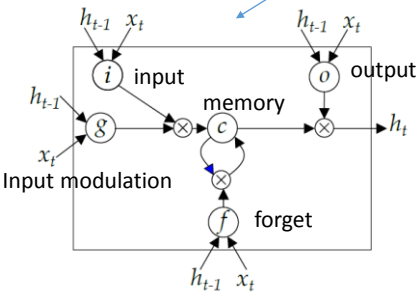

Hod Lipson

Cornell University

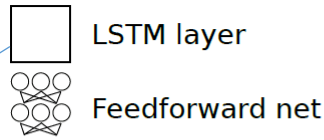
UNIVERSITY OF WYOMING

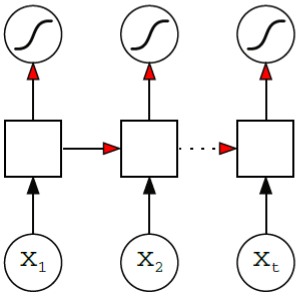
NASA Jet Propulsion Laboratory
California Institute of Technology

Bi-directional LSTM RNN's



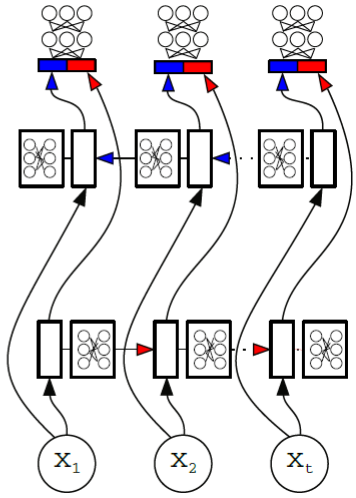
LSTM memory cell:
i: input gate,
f : forget gate,
o: output gate,
g: input modulation gate,
c: memory cell.





Protein Sequence Input

Uni-directional LSTM



Protein Sequence Input

Bi-directional LSTM RNN
S.K. Sønderby, O. Winther [7]

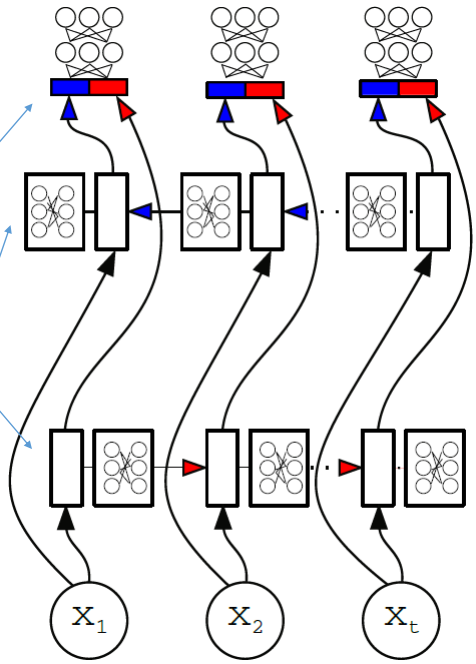
Bi-directional LSTM RNN

When predicting the secondary structure, the whole Amino Acid Sequence $x_1,\ldots,x_N$ is known.

- In forward RNN's only the past sequence $x_1,\ldots,x_t$ (t<N) is used to do the next prediction.
- In our case this is not optimal.

Therefor bidirectional RNN's were introduced:

- Two separate RNN's, the forward RNN starts the recursion from x_1 and goes forwards
- The backwards model starts at x_N and goes backwards.
- The predictions from the forward and backward networks are combined and normalized.
- Normalize the activations from each layer in a softmax layer.



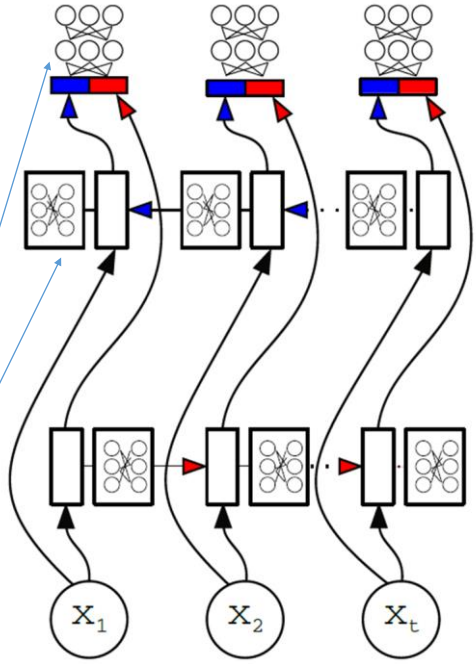
Bi-directional LSTM RNN

Therefor bidirectional RNN's were introduced:

- Two separate RNN's, the forward RNN starts the recursion from x_1 and goes forwards
- The backwards model starts at x_N and goes backwards.
- The predictions from the forward and backward networks are combined and normalized.
- Normalize the activations from each layer in a softmax layer.

Here the standard stacked bidirectional LSTM model is extended:

- by a feed-forward network responsible for concatenating the output from the forward and backward networks into a single softmax prediction
- And a feed-forward network between recurrent hidden states, along with shortcut connections between the recurrent hidden layers.



Data

Troyanskaya (2014)

- Amino acid sequences labeled with secondary structure.
- Sequences and structures were downloaded from PDB and annotated with the DSSP program [Kabsch & Sander, 1983].
- The 8-class DSSP output, the harder problem, is used.
- Each amino acid is encoded as an 42 dimensional vector, 21 dimensions for orthogonal encoding and 21 dimensions for sequence profiles.
- For further descriptions see Troyanskaya 2014.

Filtering and division:

- The full dataset has 6128 non-homologous sequences (identity less than 30%).
- This set is further filtered such that no sequences has more than 25% identity with the CB513 dataset [Cuff & Barton, 1999].
- The dataset is divided into a training (n=5278) and a validation set (n=256).
- The CB513 dataset is used for testing.

DSSP (Define Secondary Structure Protein) "The DSSP program defines secondary structure, geometrical features and solvent exposure of proteins, given atomic coordinates in Protein Data Bank format. The program does NOT PREDICT protein structure."

Results

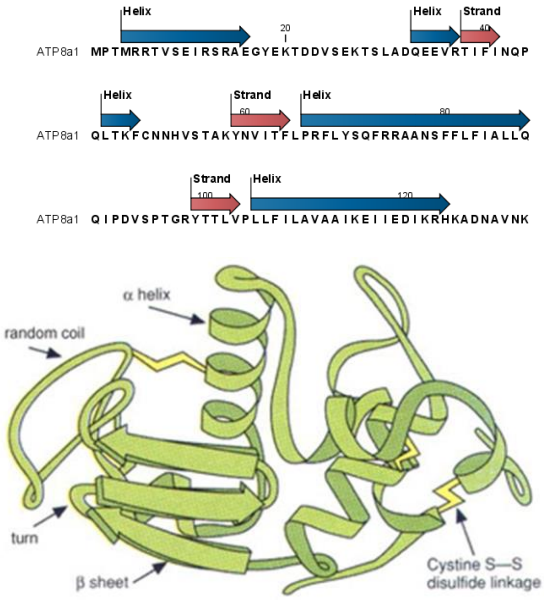
Table 2. Test set per amino acid accuracy for CB513. *Reported by Wang et al. 2011

	Q8 accuracy
[Pollastri et al., 2002] (BRNN)*	0.511
Wang et al. 2011 (CNF - 5-model ensemble)	0.649
Troyanskaya 2014 (GSN)	0.664
LSTM small	0.671
LSTM large	**0.674**

The LSTM network obtains a correct classification rate of 0.674, improves:

- 0.664 by generative stochastic network (GSN) [Bengio & Thibodeau-Laufer, 2013; Troyanskaya, 2014]
- 0.649 by conditional neural field (CNF) [Lafferty et al., 2001; Peng et al., 2009].
- 0.511 by bidirectional RNN (BRNN) having a correct classification rate of 0.511 [Pollastri et al., 2002]

Secondary Structure Prediction



8-class (Q8)	**3 class (Q3)**	**Name**
H	H	α-helix
E	E	β-strand
L	C	loop or irregular
T	C	β-turn
S	C	bend
G	H	3_{10}-helix
B	E	β-bridge
I	C	π-helix

Secondary Structure Prediction (Wikipedia 2019)

Name	Method description	Type	Link	Initial release
Porter 5	Fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes	Webserver/downloadable	server/download	2018
SPIDER2	The most comprehensive and accurate prediction by iterative Deep Neural Network (DNN) for protein structural properties including secondary structure, local backbone angles, and accessible surface area (ASA)	Webserver/downloadable	server/download	2015
s2D	Predicts disorder and secondary structure in one unified framework. Trained on solution-based NMR data.	Webserver/downloadable	server/download	2015
RaptorX-SS8	predict both 3-state and 8-state secondary structure using conditional neural fields from PSI-BLAST profiles	Webserver/downloadable	server download	2011
NetSurfP	Profile-based neural network	Webserver/downloadable	server	2009
GOR	Information theory/Bayesian inference	Many implementations	Basic GOR GOR V	2002 (GOR V)
Jpred	Multiple Neural network assignment from PSI-BLAST and HMMER profiles. Predicts secondary structure and solvent accessibility	Webserver	server and API	1998
Meta-PP	Consensus prediction of other servers	Webserver	server	1999
PREDATOR	Knowledge-based database comparison	Webserver	server	1997
PredictProtein	Profile-based neural network	Webserver	server	1992
PSIPRED	two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST	Webserver	server	1999
SOPMA	Self OPtimised Prediction Method from multiple Alignments (based on nearest neighbour method)	Webserver	server	1995
Homology Modeling Professional for HyperChem	Frequency analysis of amino acid residues observed in proteins	Commercial	algorithm	2002
SymPred	an improved dictionary based approach which captures local sequence similarities in a group of proteins	Webserver	server	2004
YASSPP	Cascaded SVM-based predictor using PSI-BLAST profiles	Webserver	server	2006
PSSpred	Multiple backpropagation neural network predictors from PSI-BLAST profiles	Webserver/downloadable program	server and downloadable program	2012
HCAM	Hidropathy Clustering Assisted Method by detection of physicochemical patterns	downloadable program plus database	main page	2013
Frag1D	Prediction of both secondary structure and Shape Strings (discrete states of dihedral angles) using profile based fragment matching	Webserver/downloadable program	main page	2010

B. Zhang, J. Li, Q. Lü, Prediction of 8-state protein secondary structures by a novel deep learning architecture. BMC Bioinformatics, 19:293, 2018. [8]

A novel deep learning architecture for protein secondary structure prediction by integrating a convolutional neural network, residual network, and bidirectional recurrent neural network.

- A local block comprised of convolutional filters and original input is designed for capturing local sequence features.
- A subsequent bidirectional recurrent neural network consisting of gated recurrent units can capture global context features.
- The residual network improves the information flow between the hidden layers and the cascaded recurrent neural network (integration).

B. Zhang, J. Li, Q. Lü, Prediction of 8-state protein secondary structures by a novel deep learning architecture. BMC Bioinformatics, 19:293, 2018. [8]

The novel deep network achieved:

- 71.4% accuracy on the benchmark CB513 dataset for the 8-state prediction (Q8)
- The ensemble learning by our model achieved 74% accuracy.
- Better than the state-of-the-art methods on three other independent datasets CASP10, CASP11 and CASP12 for both 8- and 3-state prediction.

Conclusion: Our experiment demonstrates that it is a valuable method for predicting protein secondary structure, and capturing local and global features concurrently is very useful in deep learning.

CRRNN Architecture

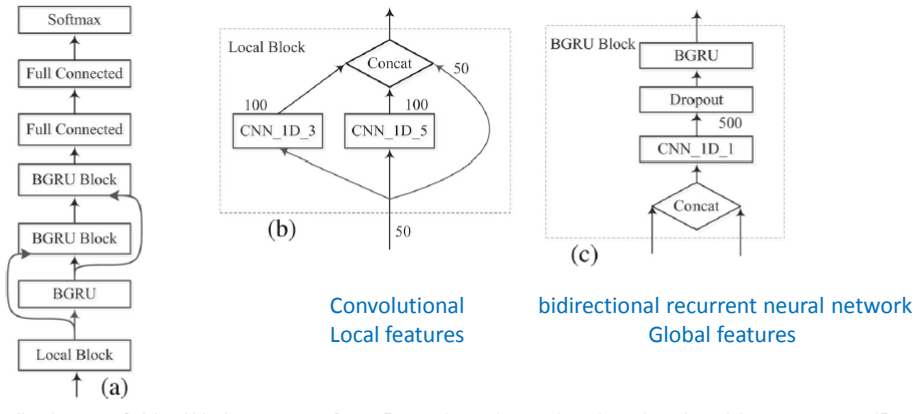


Fig. 1 a CRRNN overall architecture. **b** A local block comprising of two 1D convolutional networks with 100 kernels, and the concatenation (Concat) of their outputs with the original input data. **c** the BGRU block. The concatenation of input from the previous layer and before the previous layer is fed to the 1D convolutional filter. After reducing the dimensionality, the 500-dimensional data is transferred to the next BGRU layer

Data

Table 1 Training and test data used in our work

Label	Types	TR6614		TR5534		CB513		CASP10		CASP11		CASP12	
		Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
H	α-helix	517653	0.352	405560	0.345	26143	0.309	6544	0.297	6330	0.309	3550	0.337
B	β-bridge	15321	0.010	12096	0.010	1180	0.014	227	0.010	221	0.011	113	0.011
E	β-strand	321156	0.218	255887	0.218	17994	0.212	5225	0.237	5089	0.248	2223	0.211
G	3_{10}helix	55994	0.038	46019	0.039	3132	0.037	797	0.036	716	0.035	320	0.030
I	π-helix	281	0	209	0	30	0	5	0	0	0	0	0
T	Turn	160753	0.109	132980	0.113	10008	0.118	2811	0.128	2299	0.112	1164	0.111
S	Bend	118800	0.081	97298	0.083	8310	0.098	1780	0.081	1751	0.085	955	0.091
L	Coil	282584	0.192	225493	0.192	17904	0.211	4652	0.211	4092	0.200	2201	0.209
All		1472542		1175542		84701		22041		20498		10526	

Results

Table 2 Q8 predictive precision of individual secondary structures from CB513

Q8 Label	CRRNN	NCCNN	MUFOLD-SS	DCRNN2[a]	DCRNN	DeepCNF
H	**0.86**	0.841	0.855	0.863	0.832	0.849
B	0.466	**0.676**	0.571	0.571	0.554	0.433
E	**0.797**	0.767	0.764	0.768	0.753	0.748
G	0.466	**0.487**	0.413	0.419	0.429	0.49
I	0	0	0	0	0	0
T	0.556	**0.577**	0.572	0.562	0.559	0.53
S	0.494	**0.548**	0.522	0.509	0.518	0.487
L	**0.603**	0.565	0.586	0.571	0.573	0.571

[a]Data is generated by our experiment
Boldface numbers indicate best performance

Table 4 A comparison of the Q8 accuracy(%) on CB513, CASP10, CASP11 and CASP12 between CRRNN and other state-of-the-art methods

method	CB513	CASP10	CASP11	CASP12
GSN	66.4	-	-	-
BLSTM	67.4	-	-	-
DeepCNF	68.3	71.8	71.7[b]	0.694[b]
DCRNN	69.7	-	-	-
DCRNN2[a]	70.4	73.9	71.2	68.8
NCCNN	70.3	-	-	-
NCCNN[a]	71.4	-	-	-
MUFOLD-SS[b]	70.5	74.2	71.6	69.5
CRRNN	71.4±0.2	73.8±0.5	71.6±0.7	68.7±0.8
eCRRNN[a]	74	**76.3**	**73.9**	**70.7**

[a]indicates ensemble model
[b]Data is generated by our experiment
Boldface numbers indicate best performance

Table 5 F1 score of individual secondary structure labels using CB513

Q8 Label	CRRNN[a]	CRRNN	NCCNN[a]	NCCNN	MUFOLD-SS	DCRNN2	DCRNN[a]	DeepCNF
H	**0.903**	0.892	0.889	0.884	0.886	0.891	0.880	0.876
B	0.138	**0.139**	0.089	0.077	0.000	0.006	0.050	0.049
E	**0.834**	0.814	0.805	0.793	0.789	0.803	0.789	0.788
G	**0.463**	0.413	0.374	0.360	0.387	0.387	0.317	0.340
I	0	0	0	0	0	0	0	0
T	**0.594**	0.555	0.565	0.549	0.561	0.550	0.540	0.529
S	**0.433**	0.397	0.343	0.334	0.373	0.342	0.336	0.335
L	**0.660**	0.629	0.631	0.621	0.622	0.611	0.610	0.611
macro-F1	**0.503**	0.480	0.462	0.452	0.452	0.449	0.440	0.441
micro_F	**0.74**	0.714	0.714	0.704	0.705	0.704	0.697	0.683

[a]indicates ensemble model
Boldface numbers indicate best performance

F1-score: $$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Q3 Results

Table 6 Q3 accuracy(%) comparison on CB513 and CASP datasets

Method	CASP10	CASP11	CASP12	CB513
PSIPRED	81.2	80.7	80.5[a]	79.2
JPRED	81.6	80.4	78.8[a]	81.7
DeepCNF	84.4	84.7	83.2[a]	82.3
DCRNN	-	-	-	84
NCCNN	-	-	-	-
MUFOLD-SS[a]	84.3	82.3	81.1	82.7
CRRNN	86.1±0.6	84.2±0.5	82.6±1.2	85.3±0.4
eCRRNN	**87.8**	**85.9**	**83.7**	**87.3**

[a]Data is generated by our experiment
Boldface numbers indicate best performance

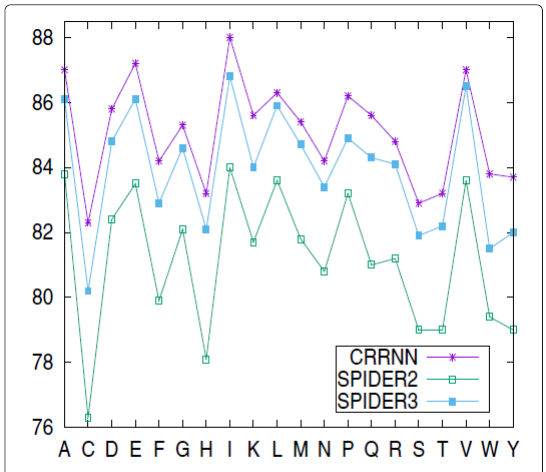
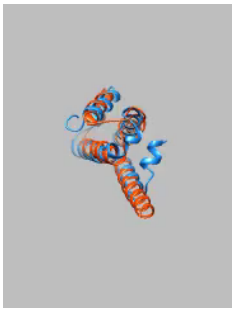


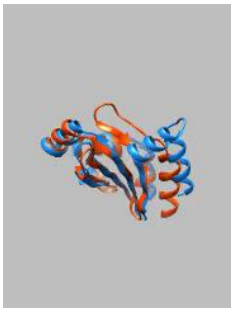
Fig. 2 The accuracy of 3-state secondary structure prediction for individual amino acids as compared with CRRNN, SPIDER3 and SPIDER2 on the TS1199 dataset

Critical Assessment of protein Structure Prediction (CASP)
http://predictioncenter.org/

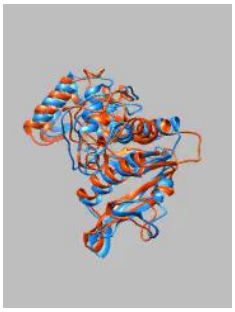
- CASP7 – CASP11



CASP7



CASP9



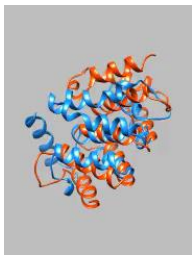
CASP11

Template Based Modeling

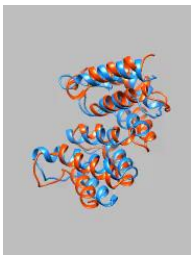
- Most accurate models.
- First 10 years of CASP enormous improvements. Unmatched until CASP12.
- From CASP11 to CASP12 the backbone accuracy of the submitted models improved more than in the preceding 10 years.
- This was due to:
  - more accurate alignment of the target sequence to that of available templates
  - combining multiple templates
  - improved accuracy of regions not covered by templates
  - successful refinement of models
  - better selection of models from decoy sets due to improved methods for estimation of model accuracy.

[Kryshtafovych et al, 2018]

Contact Prediction



No contact restrictions



With contact restrictions

- Progress in recent CASPs (2014, 2016) due to better methods for predicting three-dimensional contacts between pairs of residues in structures.
- Average precision of the best CASP12 compared to that of the best CASP11 predictor increased from 27% to 47%.
- Overall 26 methods in CASP12 showed better results than the best method in CASP11. [Schaarschmidt et al, 2018]
- Theoretical advance in contact prediction lead to improved accuracy of 3D models, especially for ab initio modeling.
- CASP13 (2018) big improvement in accuracy of contact prediction, with the average precision of the best contact prediction group increasing by 23% (compared to CASP12) and reaching 70%.

Ab Initio Modeling

- Modeling proteins with no or marginal similarity to existing structures (ab initio, new fold, non-template or free modeling) is the most challenging task in tertiary structure prediction.
- Probably the first ab initio model of reasonable accuracy was built in CASP4.
- Since then progress but mainly for small proteins (120 residues or less).

- In CASP11 for the first time a larger new fold protein (256 residues, sequence identity to known structures <5%) was built with unprecedented before accuracy for targets of this size.
- CASP11 and CASP12 experiments (2014, 2016) also showed a new trend in building better non-template models by successful utilizing predicted contacts.

- CASP13 witnessed yet another substantial improvement in accuracy of template-free models likely due to employing deep learning artificial intelligence techniques.
- The best models submitted on difficult for modeling targets showed substantial increase in average GDTTS (see next slide) scores going from 52.9 in CASP12 up to 65.7 in CASP13.

CASP GDTTS Scores

CASP's metric to quantify the quality of individual models is the Global Distance Test Total Score3 (GDTTS)

- a metric for model topology assessment.
- GDTTS reports an average of the maximum number of model residues that can be superimposed onto the target under cutoffs of 1, 2, 4, and 8 Å.
- The GDTTS score ranges between 0 and 100.
- <20 generally indicates poor models
- >50 generally indicates overall good topology.
- = 100 corresponds to a model that matches the full structure within 1 Å deviation in the Calpha coordinates of all residues.

J. Hou, et al. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13, 2019.

- CASP10 (2012) deep learning for contact and distance distribution prediction.
- CASP11 (2014) prediction of residue-residue distance relationships (e.g. contacts) is the key direction to advance protein tertiary structure prediction.
- CASP11 and CASP12: successes of residue-residue co-evolutionary analysis

CASP13 (2018) MULTICOM (3rd place) a protein structure prediction system with three major deep learning components:

- contact distance prediction based on deep convolutional neural networks
- contact distance-driven template-free (ab initio) modeling
- protein model ranking empowered by deep learning and contact prediction
- further components: template library, sequence database, and alignment tools.
- MULTICOM was ranked 3rd out of all 98 predictors in both template-free and template-based protein structure modeling in CASP13.

co-evolutionary analysis

two amino acids in contact (or spatially close according to a distance threshold such as 8Å) must co-evolve in order to maintain the contact relationship during evolution

- if one amino acid is mutated to a positively charged residue, the other one must change to a negatively charged one to be in contact.
- A number of co-evolutionary methods of calculating direct rather than indirect/accidental correlated mutation scores has been developed

  => improve contact prediction.

The co-evolutionary scores can be used as input for machine learning methods to further improve contact prediction.

MULTICOM ideas

Challenges for accurately predicting protein contact distance:

- few homologous sequences to generate co-evolutionary signals.
- folding proteins from noisy contact distances
- ranking models of hard targets.

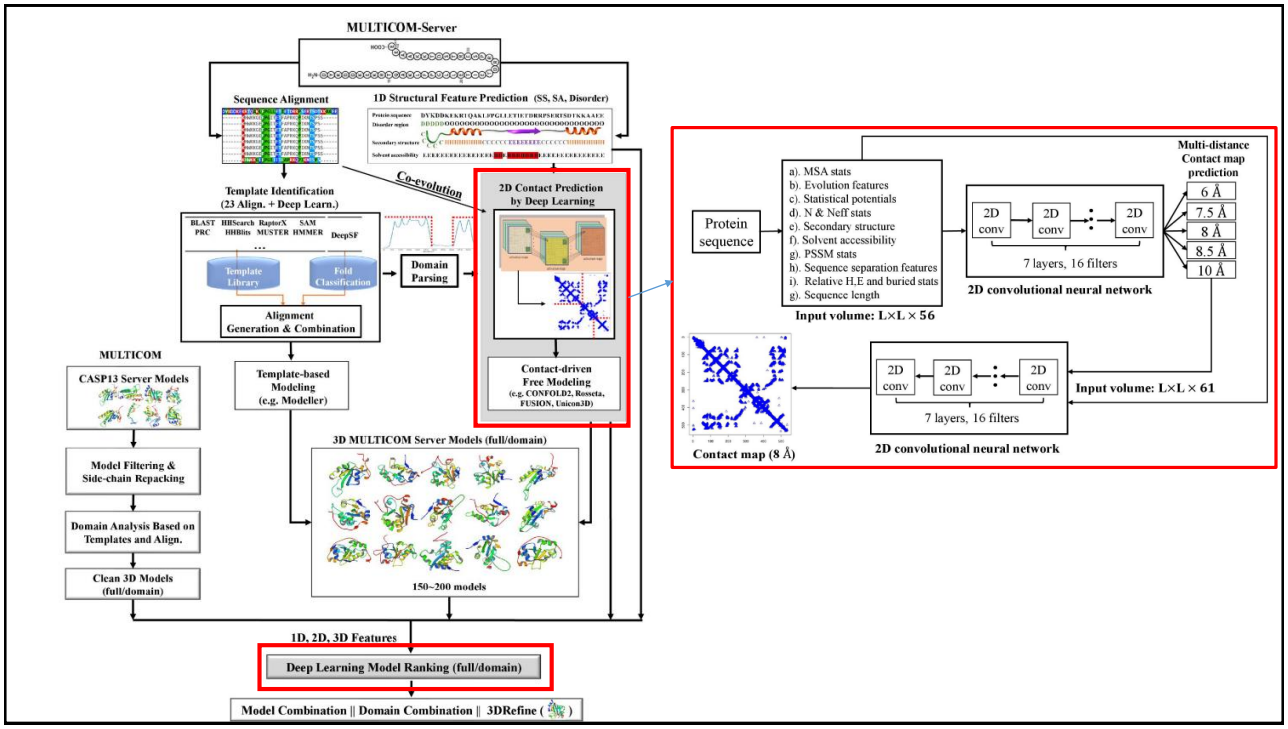
MULTICOM Deep convolutional neural network:

- Utilize global information in pairwise residue-residue features

  co-evolution scores to substantially improve inter-residue contact distance prediction,
- Integrated 1D structural features, 2D contact information, and 3D structural quality scores to improve protein model quality assessment

  Note: the contact prediction enhances ranking of protein models for the first time.

Key: Protein contact distance prediction and model selection using deep learning.



References

[1] L.A. Abriata, G.E. Tam, B. Monastyrskyy, A. Kryshtafovych, M. Dal Peraro, Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. Proteins, Vol. 86, pp. 97-112, 2018.

[2] A. Kryshtafovych, B. Monastyrskyy, K. Fidelis, J. Moult, T. Schwede, A. Tramontano Evaluation of the template-based modeling in CASP12. Proteins, Vol. 86, pp. 321-334, 2018.

[3] K. Paliwal, J. Lyons, R. Heffernan, A Short Review of Deep Learning Neural Networks in Protein Structure Prediction Problems, Adv. Tech. Biol. Med. Volume 3, Issue 3, 2015.

[4] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, A.M.J.J. Bonvin, Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins, Vol. 86, pp. 51-66, 2018.

Not Peer Reviewed

[5] G. Derevyanko, S. Grudininy, Y. Bengiozx, G. Lamoureux, Deep convolutional networks for quality assessment of protein folds. arXiv:1801.06252v1, 18 Jan 2018.

[6] J. Hou, T. Wu, R. Cao, J. Cheng. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13, http://dx.doi.org/10.1101/552422doi: bioRxiv preprint first posted online Feb. 17, 2019.

References

[7] S.K. Sønderby, O. Winther, Protein Secondary Structure Prediction with Long Short Term Memory Networks, arXiv:1412.7828v2 [q-bio.QM] 4 Jan 2015.

[8] B. Zhang, J. Li, Q. Lü, Prediction of 8-state protein secondary structures by a novel deep learning architecture. BMC Bioinformatics, 19:293, 2018.