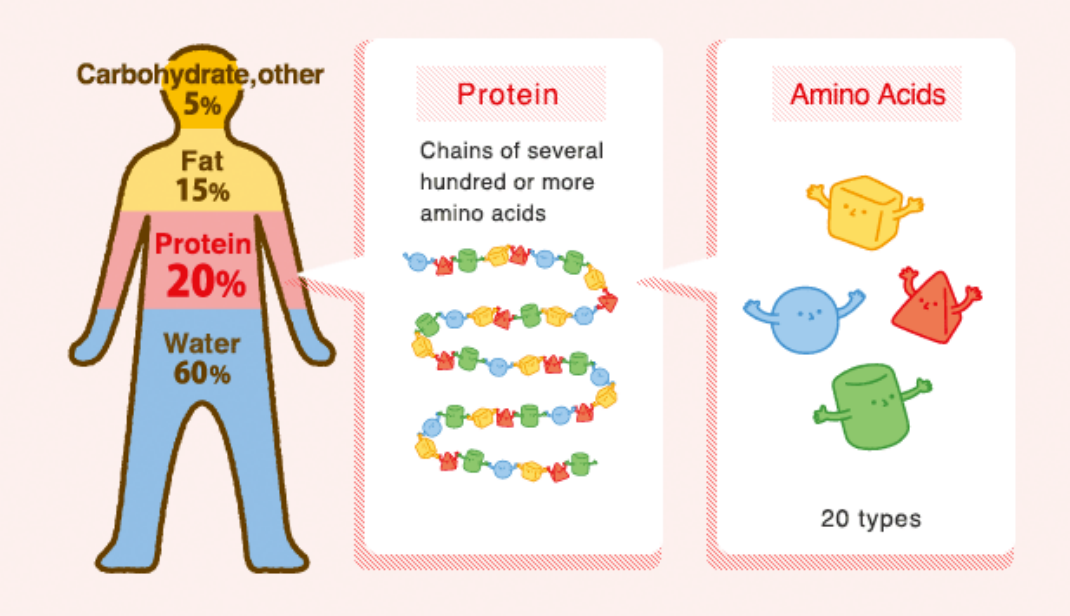
Protein structure Prediction

AI Lab. 2021.05.10

Sung-eun Jang

Background

Protein



drug discovery

antibody in immune system

catalyst of chemical reactions

signal transaction

intercellular molecular transfer

...

- Macromolecular organic material makes up the body of living things in biochemistry

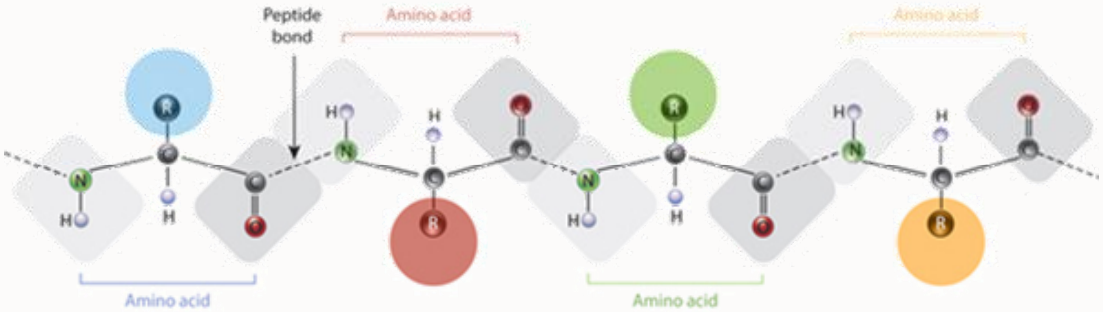
  connections of many amino acids long-linked by chemical

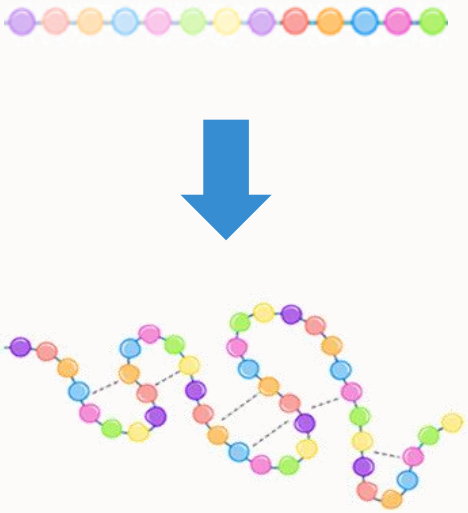
  have their own functions decided by their structure

  affects all processes in living organism

Background

Protein and amino acid





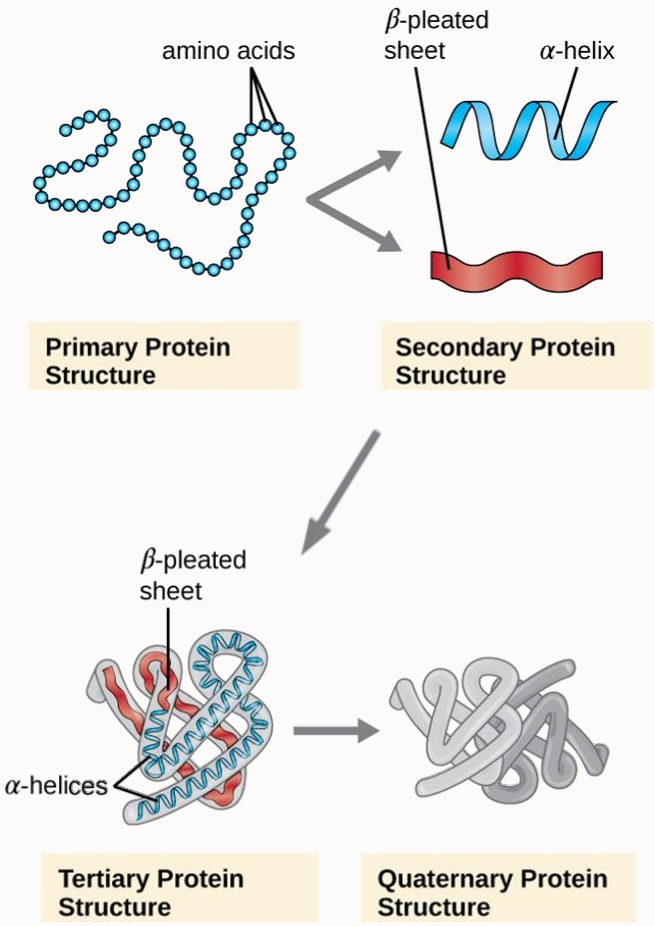
- Amino acids are classified according to side chain (R)

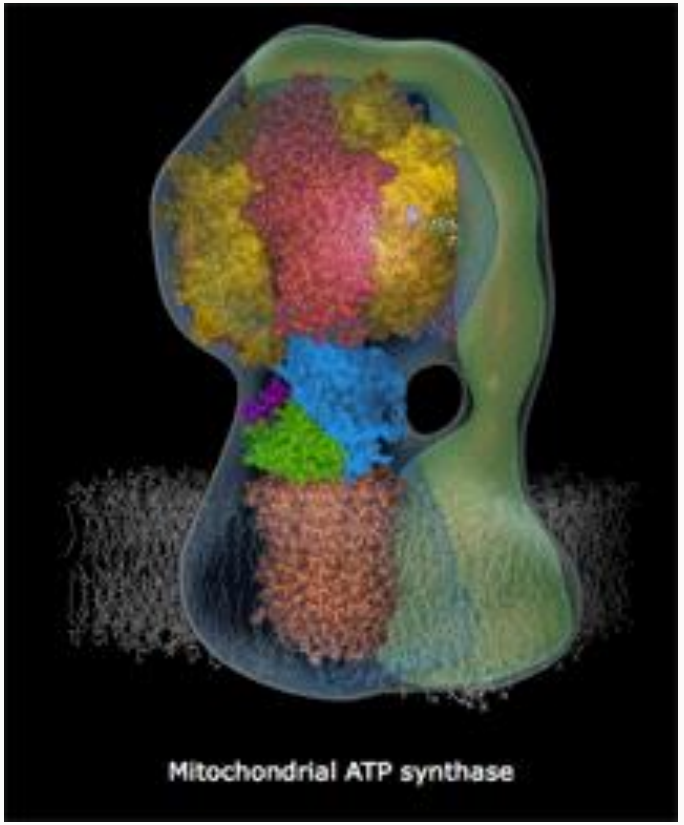
  a large part of the chain is called a backbone (N and two C per amino acid)

- **Protein folding** is determined by an interaction network between amino acids
- Final **structure of the protein** chain is determined by the amino acid sequence

Background

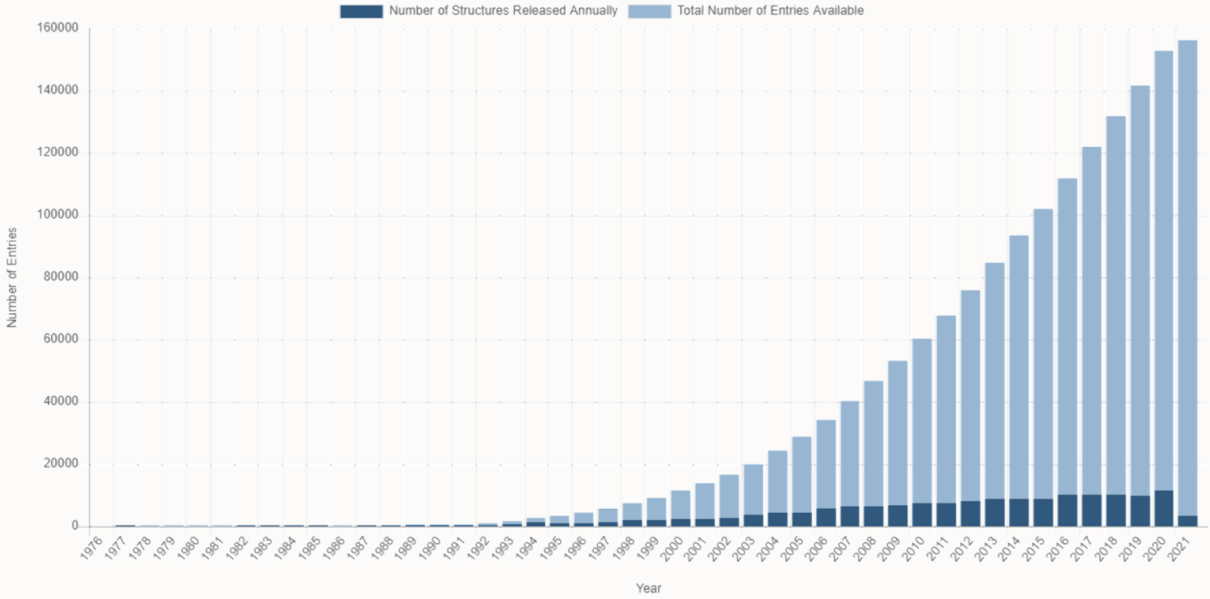
Protein structure





Background

Protein structure prediction



- Protein structure is usually obtained through x-ray determination

  equipment is expensive & protein is difficult to crystallize

  requires thousands of dollars and it's hard to get the right results

- Compared to the protein sequence, the protein structure is less identified

- **Protein structure prediction is being studied to compensate for these differences.**

Machine learning based Protein structure prediction

Computational methods

- De novo modeling methods (Molecular dynamics)

  track the micro-movement of each molecule

  using force between atoms in the physical system

- Template-based methods (homology modeling)

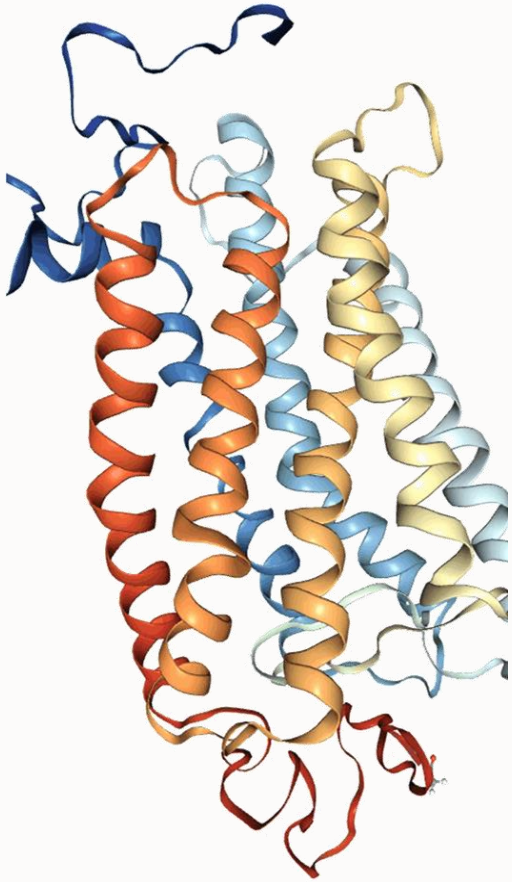
  compares the fragments with protein structure library

  borrows the structure of the similar fragment

- Template-free methods (machine learning based methods)

  build models and accurately predict protein structures

  solely based on amino acid sequences



Machine learning based Protein structure prediction

Computational methods

- De novo modeling methods (Molecular dynamics)

  only applicable to very small proteins

  it takes weeks to calculate a protein

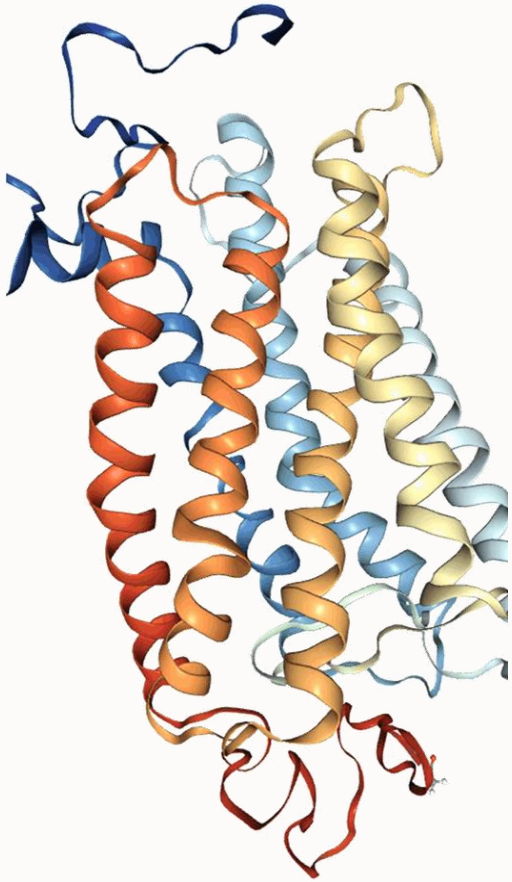
  no guarantee of getting the right structure

- Template-based methods (homology modeling)

  poor accuracy with mutation

  time consuming

- Template-free methods (machine learning based methods)



Machine learning based Protein structure prediction

Database

Database Sources	Websites
PDB	http://www.rcsb.org/pdb/
UniProt	http://www.uniprot.org/
DSSP	http://swift.cmbi.ru.nl/gv/dssp/
SCOP	http://scop.mrc-lmb.cam.ac.uk/
SCOP2	http://scop2.mrc-lmb.cam.ac.uk/
CATH	http://www.cathdb.info/

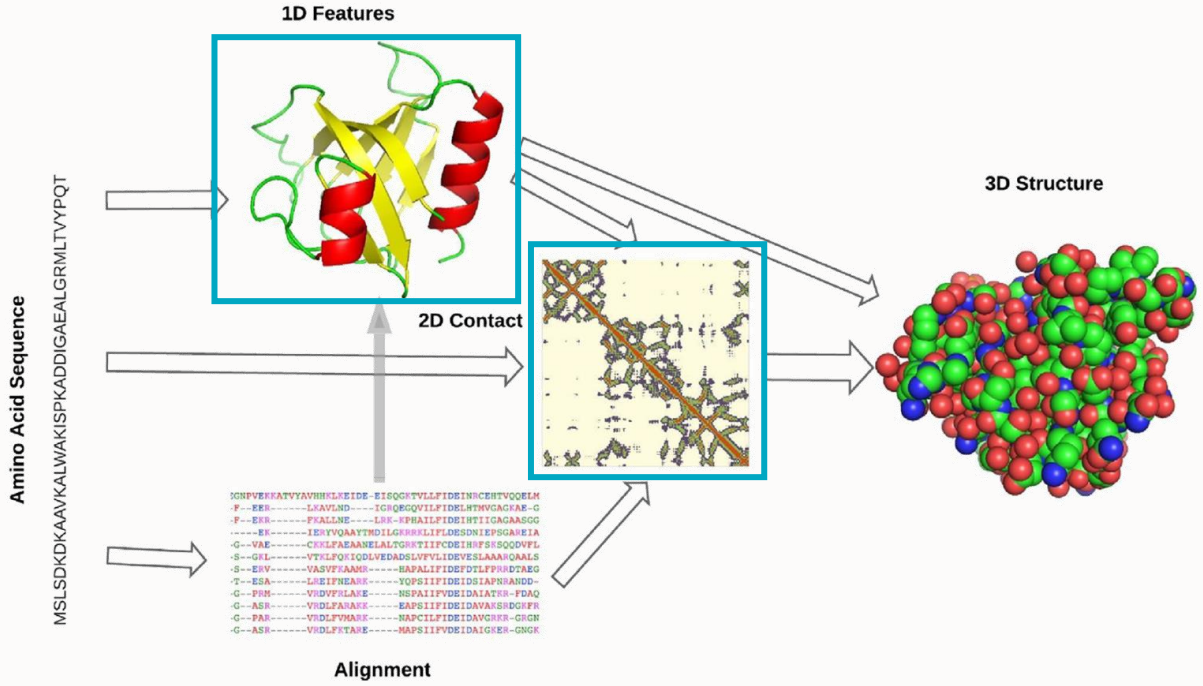
- Organize and annotate the protein structures
- Often includes three-dimensional coordinates as well as experimental information

  unit cell dimensions

  angles for determined structures.

Machine learning based Protein structure prediction

Protein Structure Annotations (PSA)



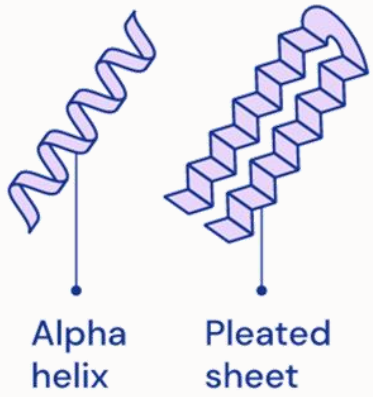
- Intermediate prediction steps for protein structure prediction which are simpler than the full, detailed 3D structure
- Where abstractions are inferred, yet structurally informative

  1D PSA: secondary structure (α, β), angle between amino acids (φ, ψ)

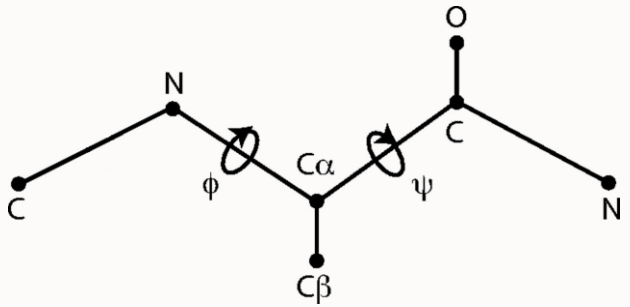
  2D PSA: contact map

Deep learning based 1D Protein Structure Annotation

1D Protein Structure Annotation



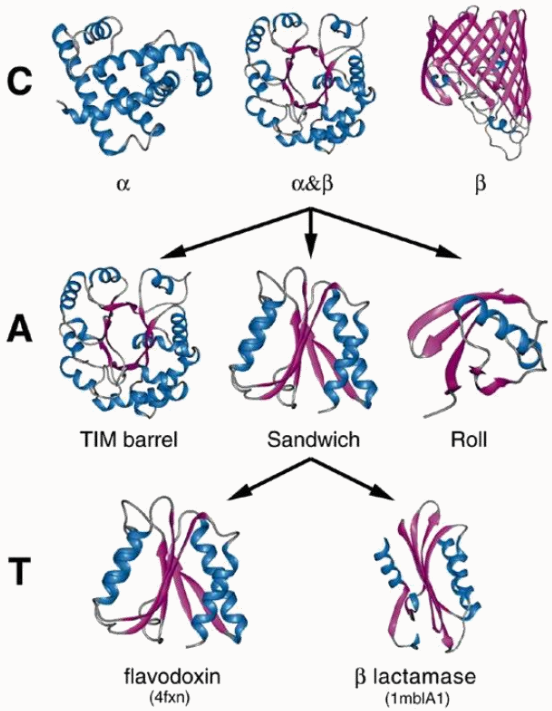
secondary structure (α, β)



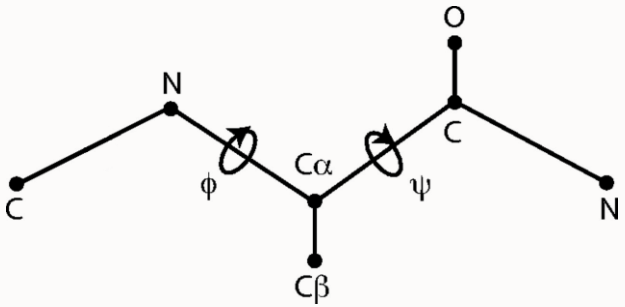
angle between amino acids (φ, ψ)

Deep learning based 1D Protein Structure Annotation

1D Protein Structure Annotation



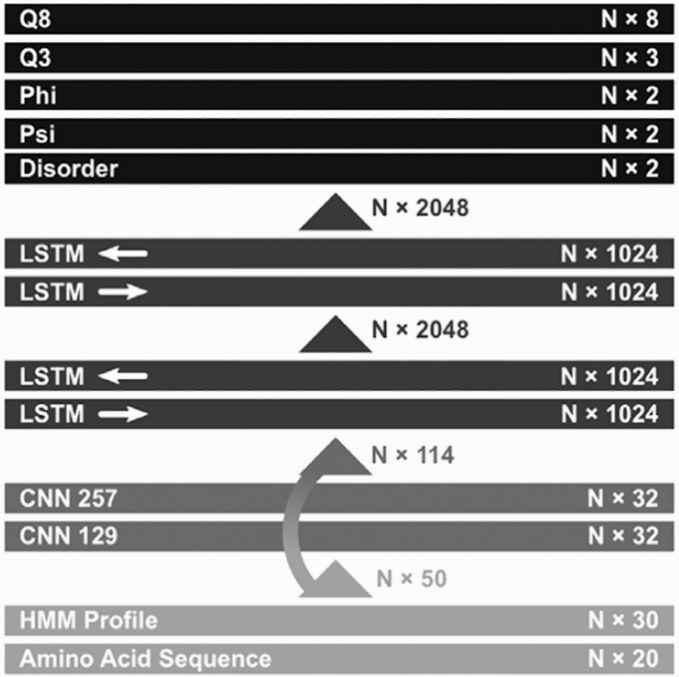
secondary structure (α, β)



angle between amino acids (φ, ψ)

Deep learning based 1D Protein Structure Annotation

NetSurfP-2.0



- NetSurfP-2.0

  sequence-based and uses an architecture composed of convolutional and

  long short-term memory neural networks trained on solved protein structures

  precision of 85% on secondary structure 3-class predictions

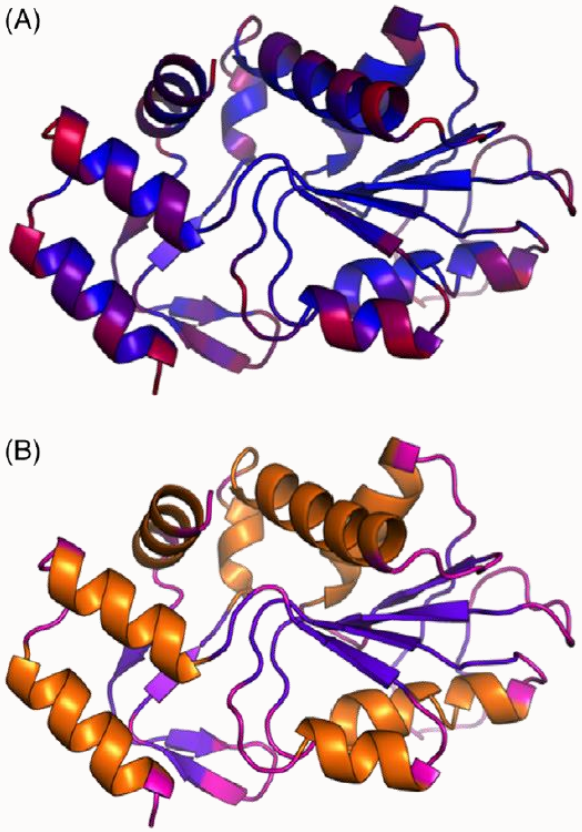
  predicting more than 1000 proteins in less than 2 hours

Deep learning based 1D Protein Structure Annotation

NetSurfP-2.0

TABLE 1 Results of the method's validation on independent test datasets

	SS3 [Q3]	SS8 [Q8]	Phi [MAE]	Psi [MAE]
CASP12				
NetSurfP-2.0 (mmseqs)	*0.820*	*0.703*	*20.3*	*31.8*
NetSurfP-2.0 (hhblits)	**0.824**	**0.711**	**20.0**	**31.2**
NetSurfP-1.0	0.709			
Spider3	0.791		21.6	33.2
RaptorX	0.786	0.661		
Jpred4	0.760			
TS115				
NetSurfP-2.0 (mmseqs)	**0.857**	**0.750**	**17.2**	**25.8**
NetSurfP-2.0 (hhblits)	*0.853*	0.744	*17.5*	*26.5*
NetSurfP-1.0	0.779			
Spider3	0.839		18.5	27.3
RaptorX	0.822	0.716		
Jpred4	0.767			
CB513				
NetSurfP-2.0 (mmseqs)	**0.854**	**0.723**	**20.1**	**28.0**
NetSurfP-2.0 (hhblits)	*0.853*	*0.720*	*20.2*	28.6
NetSurfP-1.0	0.788			
Spider3	0.845		*20.4*	*28.2*
RaptorX	0.827	0.706		
Jpred4	0.779			



Deep learning based 1D Protein Structure Annotation

NetSurfP-2.0

Submit data

Paste in FASTA sequences or choose a file from your computer below. For detailed instructions, see "Help" tab above. Only amino acid input is accepted, maximum 100 sequences or a total of 100,000 residues .

For an input of less than 10 sequences, the HHblits method is used. For 10 and more sequences MMseqs is used to generate the sequence profiles.

For an overview of the methods, performance data and citation information is found under the Abstract/Cite tab above.

Sequence submission: paste the sequence(s) *and/or* upload a local file

```
mtnrtlsreeirkldrdlrilvatngtltrvlnvvaneeivvdiinqqlldvapkipelenlkigr
ilqrdillkgqksgilfvaaeslividllptaittyltkthhpigeimaasrietykedaqvwigd
lpcwladygywdlpkravgrryriiaggqpviitteyflrsvfqdtpreeldrcqysndidtrsgd
rfvlhgrvfknl
```

For example sequences Click here
Format directly from your local disk:
파일 선택 선택된 파일 없음
Submit Clear fields

NetSurfP - 2.0

Protein secondary structure and relative solvent accessibility

Server predicts the surface accessibility, secondary structure, disorder, and phi/psi dihedral angles of amino acids in an amino acid sequence.

There has been some portability issues for the output. This will be taken care for later. It is possible to see and export the output. Notice: it is a slow service.

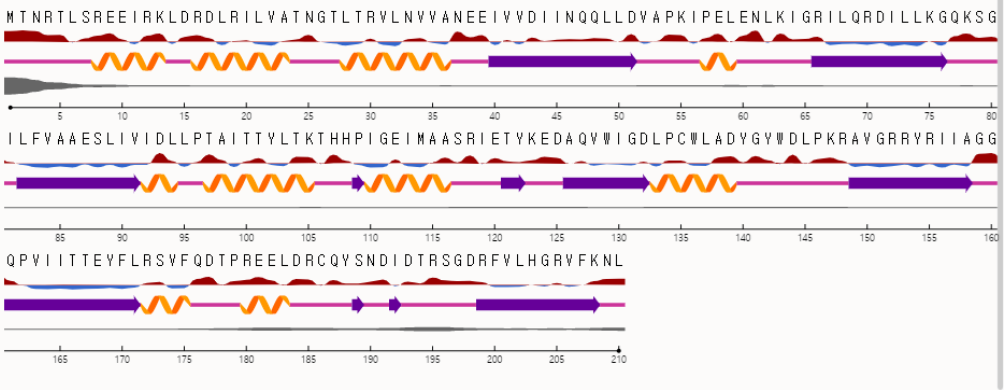
Submission | Abstract | Instructions | Dataset | Downloads

Showing 1 Prediction

Below is a graphical representation of 210 residue predictions across 1 sequence. Running time was 46 seconds (46 seconds per sequence). Hover your mouse over a sequence position to see all outputs.

Relative Surface Accessibility: Red is exposed and blue is buried, thresholded at 25%.
Secondary Structure: Helix, Strand, Coil.
Disorder: Thickness of line equals probability of disordered residue.

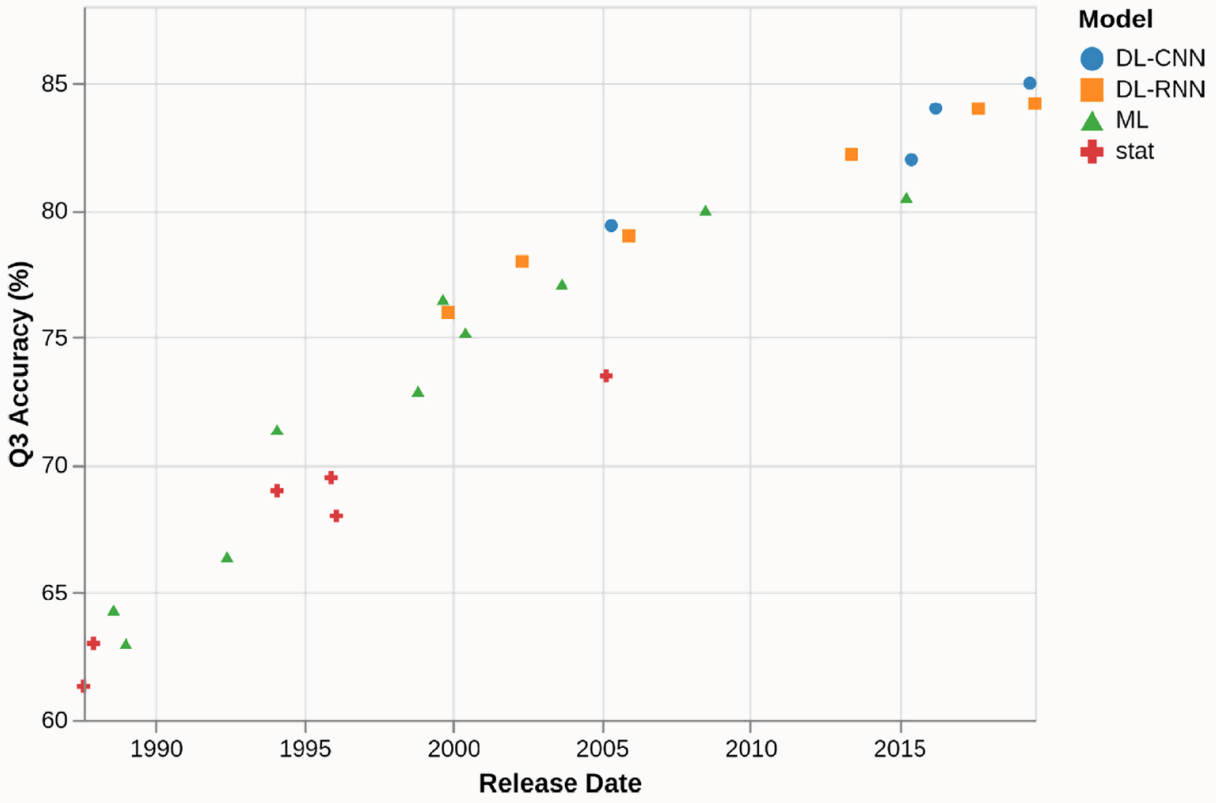
Sequence Export Sequence



https://services.healthtech.dtu.dk/service.php?NetSurfP-2.0

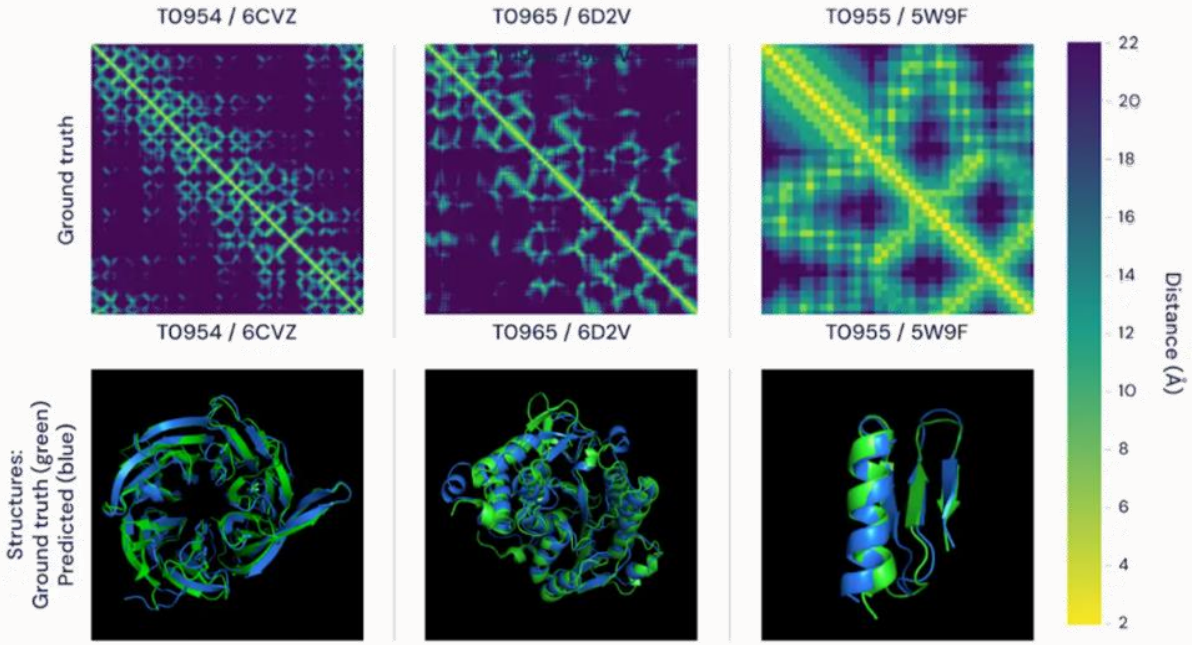
Deep learning based 1D Protein Structure Annotation

RNN based protein structure prediction



Deep learning based 2D Protein Structure Annotation

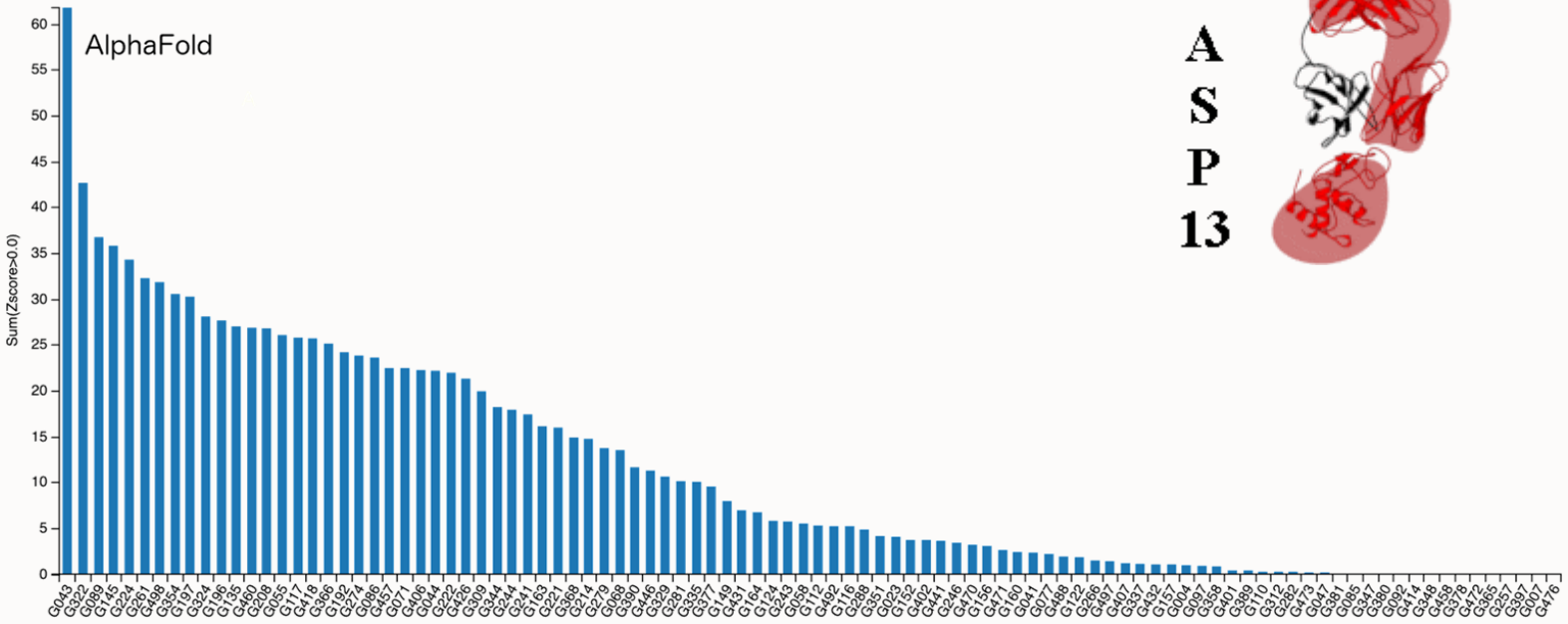
2D Protein Structure Annotation



- Contact maps have been adopted to reconstruct the full 3D protein structure
- Typically lead to more accurate 3D structures than binary maps
- Tend to be more robust when random noise is introduced in the map

Deep learning based 2D Protein Structure Annotation

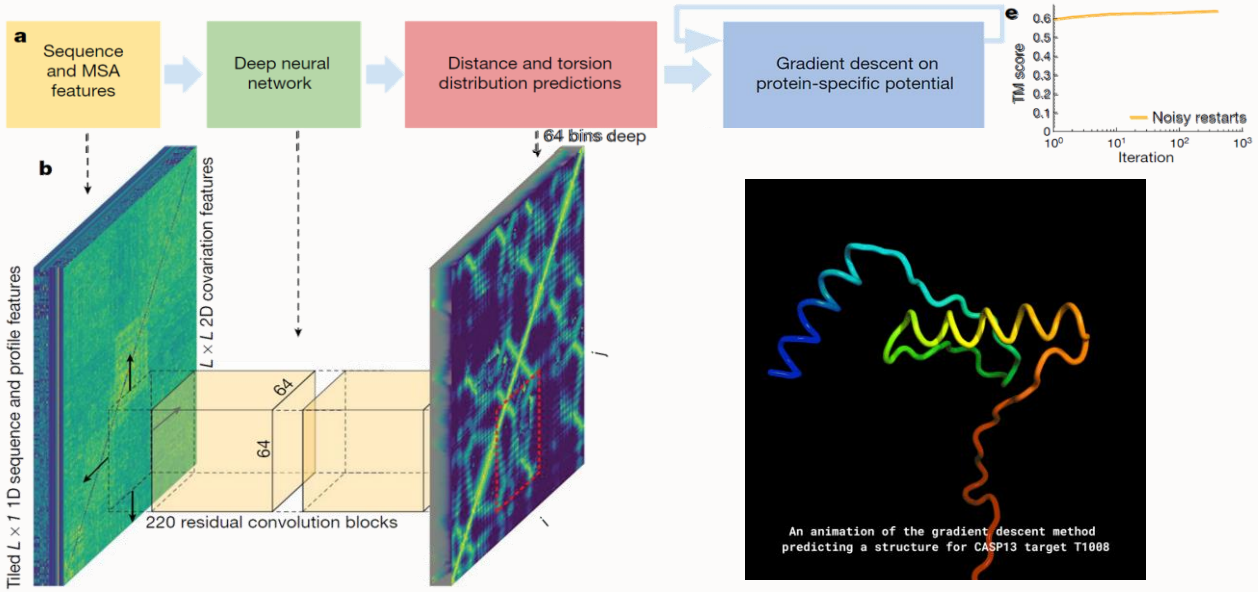
Critical Assessment of Techniques for Protein Structure Prediction (CASP)



- biennial blind protein structure prediction assessment
- run by the structure prediction community to benchmark progress in accuracy

Deep learning based 2D Protein Structure Annotation

AlphaFold



- Make accurate predictions of the distances between pairs of amino acid
- Achieves high accuracy, even for sequences with fewer homologous sequences

Deep learning based 2D Protein Structure Annotation

AlphaFold

- AlphaFold relies on deep neural networks

  the **distances** between pairs of amino acids

  the **angles** between amino acids

- Distances

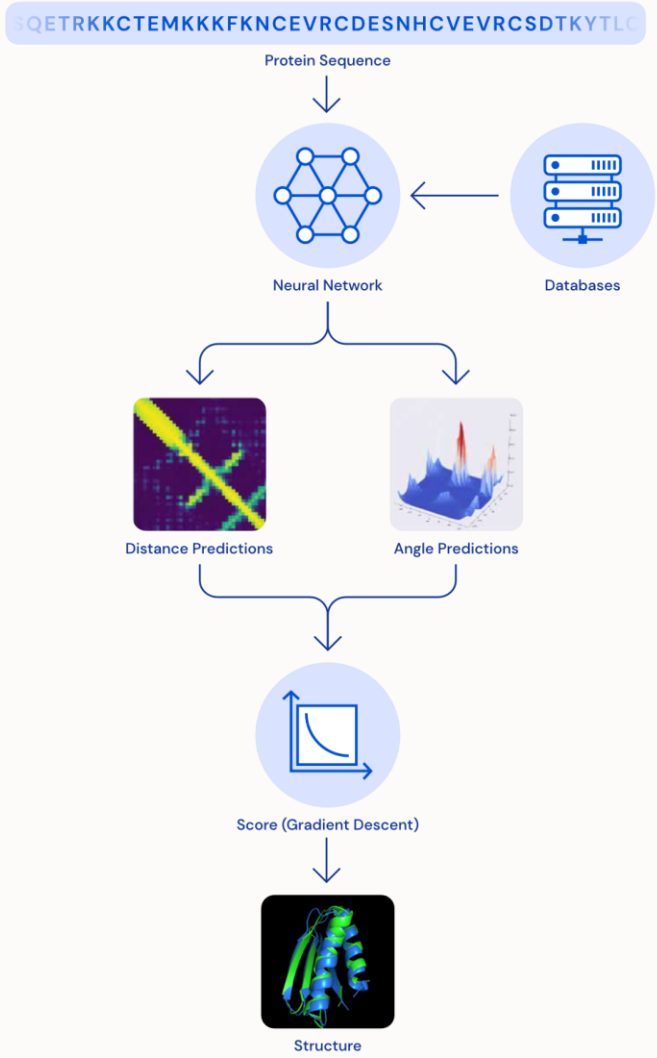
  search the protein landscape to find structures that matched our predictions

  repeatedly replaced pieces of a protein structure with new protein fragments

- Angles

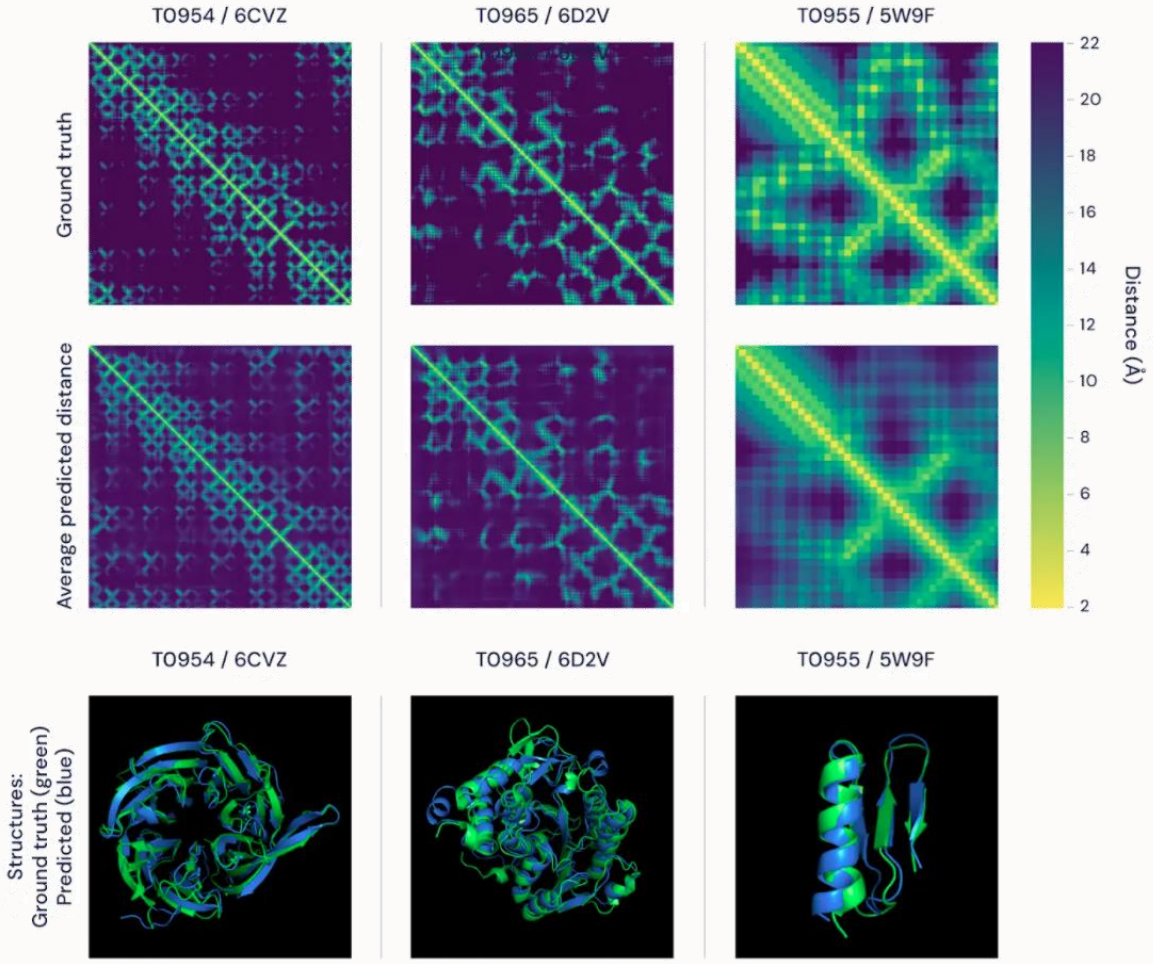
  applied to entire protein chains

  optimised scores through gradient descent



Deep learning based 2D Protein Structure Annotation

AlphaFold



THANK YOU