Protein Structure Prediction

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2024
Anthony Gitter
gitter@biostat.wisc.edu
Goals for lecture

• Why is protein structure important
• Elements of proteins
• AlphaFold2 and its impact
• What can we do with predicted structures
• What comes next for proteins
WHY IS PROTEIN STRUCTURE IMPORTANT?
Protein structure determines function

Ion channel protein with $K^+$ ions at membrane

Protein structure determines function

Hexokinase binding glucose

Vishal Bhoir https://www.youtube.com/watch?v=U7GJWrt_VPA
Protein structure determines function

DNA-binding domain (DBD)

E2F8 protein bound to DNA

Morgunova Nat Commun 2015 doi:10.1038/ncomms10050
The function of a protein is determined in large part by its 3D shape (*fold, conformation*)

Can we predict the 3D shape of a protein given only its 1D amino-acid sequence?
ELEMENTS OF PROTEINS
Protein Architecture

• Proteins are polymers consisting of amino acids linked by *peptide* bonds
• Each amino acid consists of
  – a central carbon atom (α-carbon)
  – an amino group, NH₂
  – a carboxyl group, COOH
  – a side chain (R)
• Differences in side chains distinguish 20 different amino acids
Amino Acids and Peptide Bonds

- **amino group**
- **side chain**
- **carboxyl group**

α carbon (common reference point for coordinates of a structure)
Amino Acid Side Chains

**Side chains vary in**
- shape
- size
- charge
- polarity
Predicting Side Chains is Hard but Important

AlphaFold  Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn

What Determines Conformation?

• In general, the amino-acid sequence of a protein determines the 3D shape of a protein [Anfinsen et al., 1950s]

• But some qualifications
  – all proteins can be denatured
  – some proteins are inherently disordered (i.e. lack a regular structure)
  – some proteins get folding help from chaperones
  – there are various mechanisms through which the conformation of a protein can be changed in vivo
    – post-translational modifications such as phosphorylation
    – prions
    – etc.
What Determines Conformation?

- Which physical properties of the protein determine its fold?
  - rigidity of the protein backbone
  - interactions among amino acids, including
    - electrostatic interactions
    - van der Waals forces
    - volume constraints
    - hydrogen, disulfide bonds
  - interactions of amino acids with water
    - hydrophobic and hydrophilic residues
Levels of Description

• Protein structure is often described at four different scales
  – primary structure
  – secondary structure
  – tertiary structure
  – quaternary structure
Levels of Description

(a) **Primary structure**
The amino acid sequence itself

(b) **Secondary structure**
“local” description of structure: describes it in terms of certain common repeating elements

3D conformation of a polypeptide

(c) **Tertiary structure**
3D conformation of a complex of polypeptides

Hydrogen bonds between amino acids at different locations in polypeptide chain

α helix

β polypeptide

(d) **Quaternary structure**
Secondary Structure

• Secondary structure refers to certain common repeating structures
• It is a “local” description of structure
• Two common secondary structure
  - $\alpha$ helices
  - $\beta$ strands/sheets (pleated sheet on previous slide)
• A third category, called coil or loop, refers to everything else
Secondary Structure

“Is the neural network an essential tool for the most accurate secondary structure prediction?”
- Burkhard Rost, 1998
Ribbon Diagram Showing Secondary Structures

- α helix
- β strand
- loop
- DNA (not a protein)
STRATEGIES FOR PREDICTING STRUCTURE
Determining Protein Structures

- Protein structures can be determined experimentally (in most cases) by
  - x-ray crystallography
  - nuclear magnetic resonance (NMR)
  - cryo-electron microscopy (cryo-EM)
- But this is very expensive and time-consuming
- There is a large sequence-structure gap
  - ≈ 1B protein sequences available
  - ≈ 100K protein structures in PDB database

- Key question: can we predict structures by computational means instead?
Determining Protein Structures

Folded structure has lowest energy

Dill Science 2012 doi:10.1126/science.1219021
Existing 3D structure prediction ideas

Homology modeling

Threading

Fragment assembly (Rosetta)

Molecular dynamics
Evolutionary conservation

• Multiple sequence alignments provide information about 3D structure

ALPHAFOLD2
Cannot Understate the Leap AlphaFold2 Made

- Critical Assessment of Structure Prediction (CASP) is community challenge to predict new held out structures
- Run since 1994
- DeepMind competed with AlphaFold in CASP13 in 2018
- Then AlphaFold2 in 2020…

https://predictioncenter.org/index.cgi
Cannot understate the leap AlphaFold2 made

- Image circulating Twitter before CASP14 conference started

https://predictioncenter.org/casp14/zscores_final.cgi
Cannot understate the leap AlphaFold2 made

- CASP performance over the years

Side chains resemble model

Cannot underestimate the leap AlphaFold2 made

- Zhavoronkov speculates in *Forbes* about AlphaFold winning Nobel prize
- Mohammed AlQuraishi’s famous blog post
  - “my expectation… not until the late 2020s would we see >90 GDT_TS for most targets.” (AlphaFold2 median 92.4)
  - “The core field has been blown to pieces; there’s just no sugar-coating it.”
  - “This was captured poignantly by a panelist at the very last session of the conference who remarked that CASP14 feels a bit like when one’s child leaves home for the very first time.”

https://moalquraishi.wordpress.com/2020/12/08/alphafold2-casp14-it-feels-like-ones-child-has-left-home/
The AlphaFold2 model

Input: amino acid sequence
Primary output: 3D coordinates for atoms

The AlphaFold2 model

Use input seq to search huge sequence database
Build multiple sequence alignment (MSA)

Use input seq to structure database
Find similar structure templates
The AlphaFold2 model

Iteratively improve embedding of MSA

Iteratively improve embedding of residue pairs
Share information across these embeddings

The AlphaFold2 model

Convert abstract sequence representation to 3D coordinates
AlphaFold2 input
AlphaFold2 input

Extra MSA seqs

Input sequence

Clustered MSA seqs

Template inputs

Compute embeddings
AlphaFold2 Evoformer

Input sequence
- Genetic database search
- Pairing
- Structure database search

MSA

+-

MSA representation (b, r, c)

Evoformer (48 blocks)

Pair representation (r, r, c)

Single repr. (r, c)

Pair representation (r, r, c)

Structure module (8 blocks)

High confidence

Low confidence

3D structure

Recycling (three times)
AlphaFold2 Evoformer

Iteratively improve embedding of MSA

Iteratively improve embedding of residue pairs
Share information across these embeddings

AlphaFold2 Evoformer: MSA

MSA row-wise gated self-attention with pair bias

Pair representation influences the attention calculations
AlphaFold2 Evoformer: MSA

MSA column-wise gated self-attention
AlphaFold2 Evoformer: pairs

Outer product mean

Extract pairwise information from updated MSA
AlphaFold2 Evoformer: pairs

Triangle multiplicative updates and self-attention

Pair representation
\((r,r,c)\)

Corresponding edges in a graph

Triangle multiplicative update using 'outgoing' edges

Triangle multiplicative update using 'incoming' edges

Triangle self-attention around starting node

Triangle self-attention around ending node
AlphaFold2 structure module

Input sequence

Genetic database search

Pairing

MSA

Templates

Structure database search

MSA representation (α, β, γ)

Evoformer (48 blocks)

Pair representation (r, θ, θ)

Pair representation (r, θ, θ)

Single representation (θ, β)

Structure module (8 blocks)

3D structure

High confidence

Low confidence

Recycling (three times)
AlphaFold2 structure module

Residue modeled as a triangle of three backbone atoms
Learn the side chain angles

Learn a rotation and translation for each residue in the sequence

AlphaFold2 structure module

Residue modeled as a triangle of three backbone atoms
Learn the side chain angles

Torsion angles of residue

BMC Bioinformatics 2011 doi:10.1186/1471-2105-12-S14-S10
AlphaFold2 structure module

Pair representation \((r,r,c)\)

8 blocks (shared weights)

Predict \(\gamma\) angles and compute all atom positions

Single repr. \((r,c)\)

 IPA module

Predict relative rotations and translations

Single repr. \((r,c)\)

Backbone frames \((r, 3\times3)\) and \((r,3)\) (initially all at the origin)

Backbone frames \((r, 3\times3)\) and \((r,3)\)
AlphaFold2 structure module

Main loss: frame aligned point error (FAPE)

Considers all atoms, must get side chains and chirality correct
Many auxiliary losses

Masked MSA

Intermediate FAPE

Confidence

Distogram loss
Many other important details

• Ensemble 5 models
• Relax structures with OpenMM
• Recycle the predictions multiple times
• Train on large number of predicted structures (self-distillation)
• Predict per-residue accuracy (predicted Cα local-distance difference test, pLDDT)
POST-ALPHAFOLD2
Academic inspiration

RoseTTAFold inspired by ideas from CASP14

OpenFold reproduces AlphaFold2 including training, improves efficiency
Access to AlphaFold2 predictions

• After source code released, still challenging to run
• Requires and a GPU or TPU
• Requires over 2.5 TB of data
  – 5 GB models
  – 238 GB structures
  – Everything else sequences
• Need to run MSA and template search preprocessing before model inference
Access to AlphaFold2 predictions

ColabFold v1.5.5: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and AlphaFold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.4, v1.5.1, v1.5.2, v1.5.3-natch


Input protein sequence(s), then hit Runtime -> Run all

- Use : to specify inter-protein chainbreaks for modeling complexes (supports homo- and hetero-oligomers). For example PI...SK:PI...SK for a homodimer

- **query_sequence:** "SKGEELFTGVPILVILEDGVEHGHFKFRVSVSEGEKDATYGKLTLKFICTTGKLPVPWPTLVTTLSYGVQFSPDYPMQOHDFFKSAMPPEGYVQERTIFKDDGNYKTRAEV"

- **jobname:** "GFP_O10"

- **num_relax:** 1

- specify how many of the top ranked structures to relax using amber

- **template_mode:** none

- none = no template information is used. pdb100 = detect templates in pdb100 (see notes). custom - upload and search own templates (PDB or mmCIF format, see notes)

ColabFold makes it trivial to generate and visualize a single structure prediction

Mirdita Nat Methods 2022 doi:10.1038/s41592-022-01488-1
Access to AlphaFold2 predictions

AlphaFold Protein Structure Database contains over 200M predicted structures, integrated into UniProt

Subjectively sensed a shift in the community

• Curiosity to see what AlphaFold2 predictions could bring to one’s problem

• Algorithm developers could assume (model of) protein structure available

• Structure models help interpret experimental data
Proteins can have conserved structure without conserved sequence

Myoglobin proteins with sequence identity to human

Cluster and analyze similarity within AlphaFold Protein Structure Database

Barrio-Hernandez Nature 2023 doi:10.1038/s41586-023-06510-w
https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/
AlphaFold-Multimer

Extension to protein complexes

(a) A2B2C2 heteromer
TM-score = 97.4, \(N_{\text{res}} = 1,246\), PDB ID = 6E3K

(b) A3B3 heteromer
TM-score = 85.4, \(N_{\text{res}} = 795\), PDB ID = 7KHD

(c) Protein-peptide complex
TM-score = 96.6, DockQ = 0.954,
\(N_{\text{res}} = 385\), PDB ID = 6JMT

(d) A2B2 heteromer
TM-score = 98.5, \(N_{\text{res}} = 716\), PDB ID = 6IWD

Evans bioRxiv doi:2021.10.04.463034
AlphaMissense

Input
Reference:
DNA
CAG
MDVAMVNQTVATMIS
Protein
Missense variant:
DNA
CGG
MDVAMVNRTVATMIS
Protein

AlphaMissense

1. Structure context
2. Protein language modeling
3. Training variants

Output
Alpha Missense pathogenicity:
1 Pathogenic
0 Uncertain
0 Benign

For all 71M possible missense variants in the human proteome:
11% likely pathogenic
32% likely benign
57% likely benign

Predict pathogenicity of genetic variants

Cheng Science 2023 doi:10.1126/science.adg7492
Single-sequence prediction

Train a language model on all natural protein sequences
Use instead of MSA
Protein structures in diverse complexes

doi:10.1126/science.adl2528
Conclusions

• Protein structure prediction from sequence was an open problem for over 50 years
• One version of it is now largely solved by AlphaFold2
• AlphaFold2 combines expert modeling of MSAs, templates, and protein geometry; professional deep learning engineering; large sequence and structure databases
• Protein machine learning and computational structural biology have flourished in the wake of AlphaFold2
Resources

- https://moalquraishi.wordpress.com/2020/12/08/alphafold2-casp14-it-feels-like-ones-child-has-left-home/
- https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3A_Biochemistry_Free_For_All_(Ahern_Rajagopal_and_Tan)/02%3A_Structure_and_Function/203%3A_Structure_Function_-_Proteins_I
- https://predictioncenter.org/index.cgi