Structure Prediction

- This mini-lecture will introduce
  - motivations of structure prediction
  - how structure space < sequence space
  - making predictions that maximize template information
  - the new deep learning methods, AlphaFold and RoseTTAFold
- At the end of this mini-lecture, we will have a discussion about:
  - What factors increase/decrease
    - your confidence in a predicted structure?
    - the influences of the prediction algorithm?
  - Can current structure prediction methods be used to predict the effect of a mutations or buffer conditions?

Why perform structure prediction?

- Predict
  - functional differences between homologs (e.g. isoforms or different species)
  - effects of mutations
  - binding sites and drug interactions
- Design the above

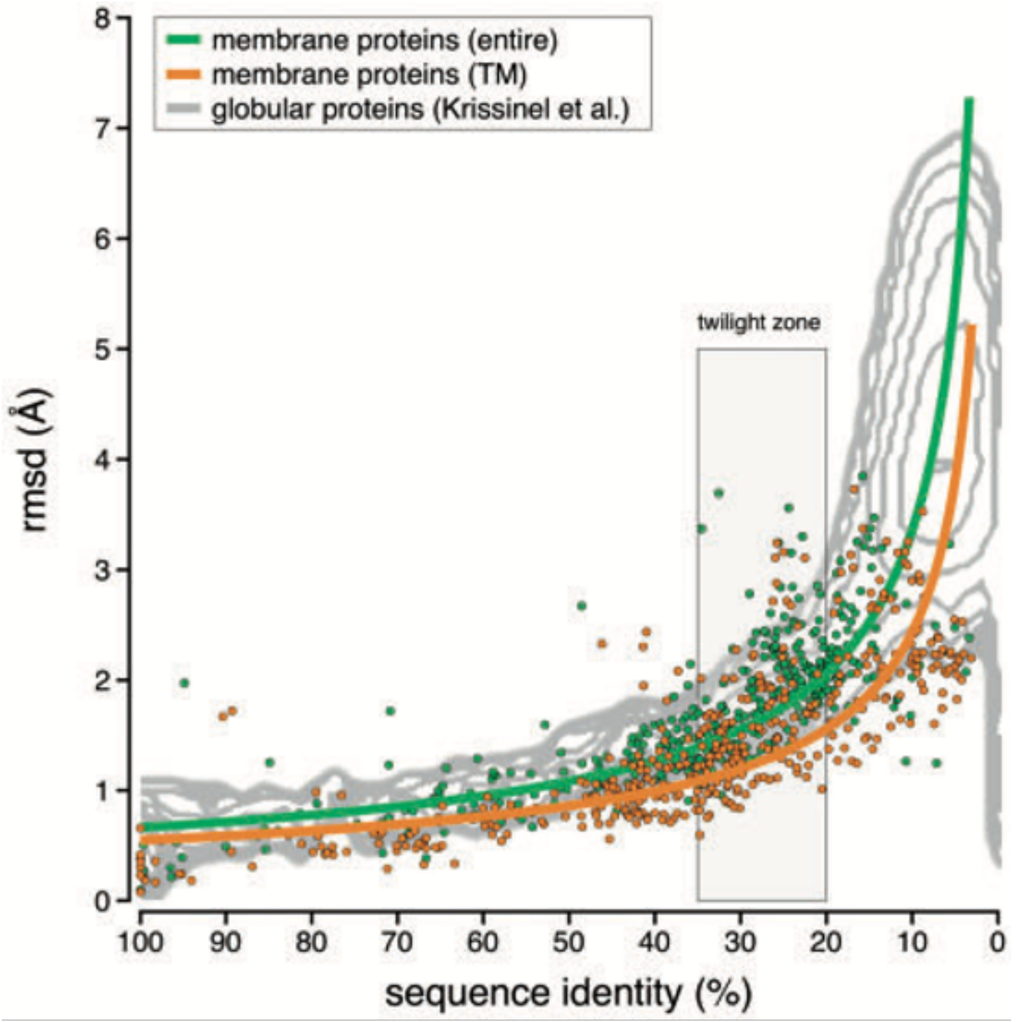
- It's easier than structure determination
- Even with rapid expansion of the PDB, many-fold more sequences have been obtained

- Fortunately…

Structure < sequence space

- The number of unique folds is far less than the number of sequences
- Similar sequences have similar structure - in general, sequence identify only needs to be 30-40%!
- Therefore we can model a sequence of unknown structure based on a homolog with known structure

- But how?



Chothia C & Lesk AM (1986) _EMBO J_ 5(4), 823–826

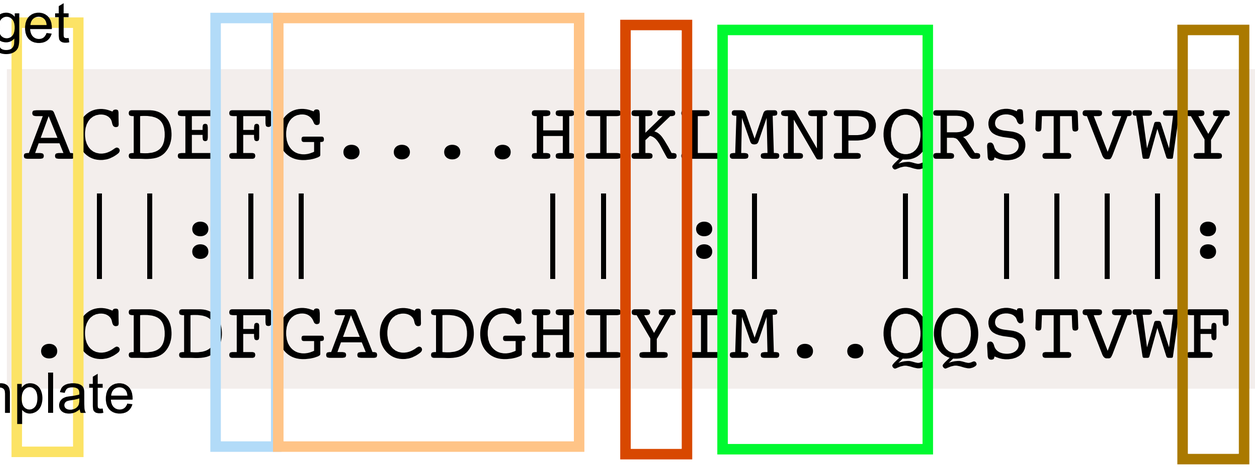
Evolutionary significance of an alignment

Given this alignment...

target

```
ACDEFG....HIKLMNPQRSTVWY
  ||:||    ||  :|   |  ||||:
.CDDFGACDGHIYIM..QQSTVWF
```

template



Biologist infers...

- The gene was extended by one residue at the N-terminus.
- The Phe is conserved.
- Four residue deletion occurred between G to H.
- A non-similar mutation Y->K occurred..
- A two-residue insertion occurred between M and Q.
- A similar mutation F->Y occurred..

Aligned positions share a common ancestral position.

Bystroff, 2020

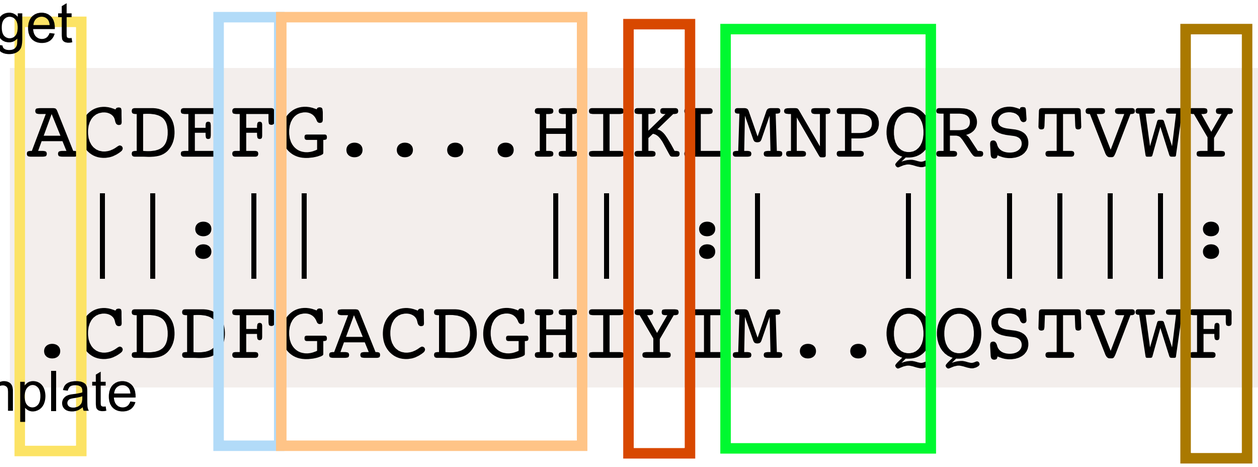
An alignment as modeling instructions

Given this alignment...

target

```
ACDEFG....HIKLMNPQRSTVWY
  ||:||    ||  :|   |  ||||:
.CDDFGACDGHIYIM..QQSTVWF
```

template



- Modeler program should...
- Add Ala to the N-terminal Cys using energy minimization.
- Keep the conserved Phe sidechain and backbone.
- Cut out the four residue insertion and connect G to H.
- Switch non-similar sidechains Y->K. Possibly move backbone. Possibly pick another alignment.
- Cut at M-Q, insert two residues, Asn-Pro.
- Switch similar sidechains F->Y. Keep backbone fixed.

Aligned positions share a common spatial position.

Bystroff, 2020

Choosing Structure Prediction Software

- There are many software tools for protein structure prediction (see https://en.wikipedia.org/wiki/List_of_protein_structure_prediction_software)
- How should you decide which to use?
  - Ease of use
    - Web server - easier for sporadic use
    - Downloadable and scriptable - easier for large-scale applications
  - Accuracy

CASP

- “Critical Assessment of protein Structure Prediction” (CASP) experiments are *blinded* tests of the ability to predict structure from sequence. (see http://www.predictioncenter.org/index.cgi)
  - “I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent community-wide CASP7, CASP8, CASP9, CASP10, CASP11, CASP12, and CASP13 experiments.”
  - “AlphaFold is an AI system developed by DeepMind that predicts a protein’s 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.” “In CASP14, AlphaFold was the top-ranked protein structure prediction method by a large margin.”

AlphaFold and RoseTTA fold

Article | Open Access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper, Richard Evans, ... Demis Hassabis + Show authors

Nature **596**, 583–589 (2021) | Cite this article

566k Accesses | **983** Citations | **2993** Altmetric | Metrics

Abstract

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1,2,3,4], the structures of around 100,000 unique proteins have been determined[5], but this represents a small fraction of the billions of known protein sequences[6,7]. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single

RESEARCH ARTICLE | PROTEIN FOLDING

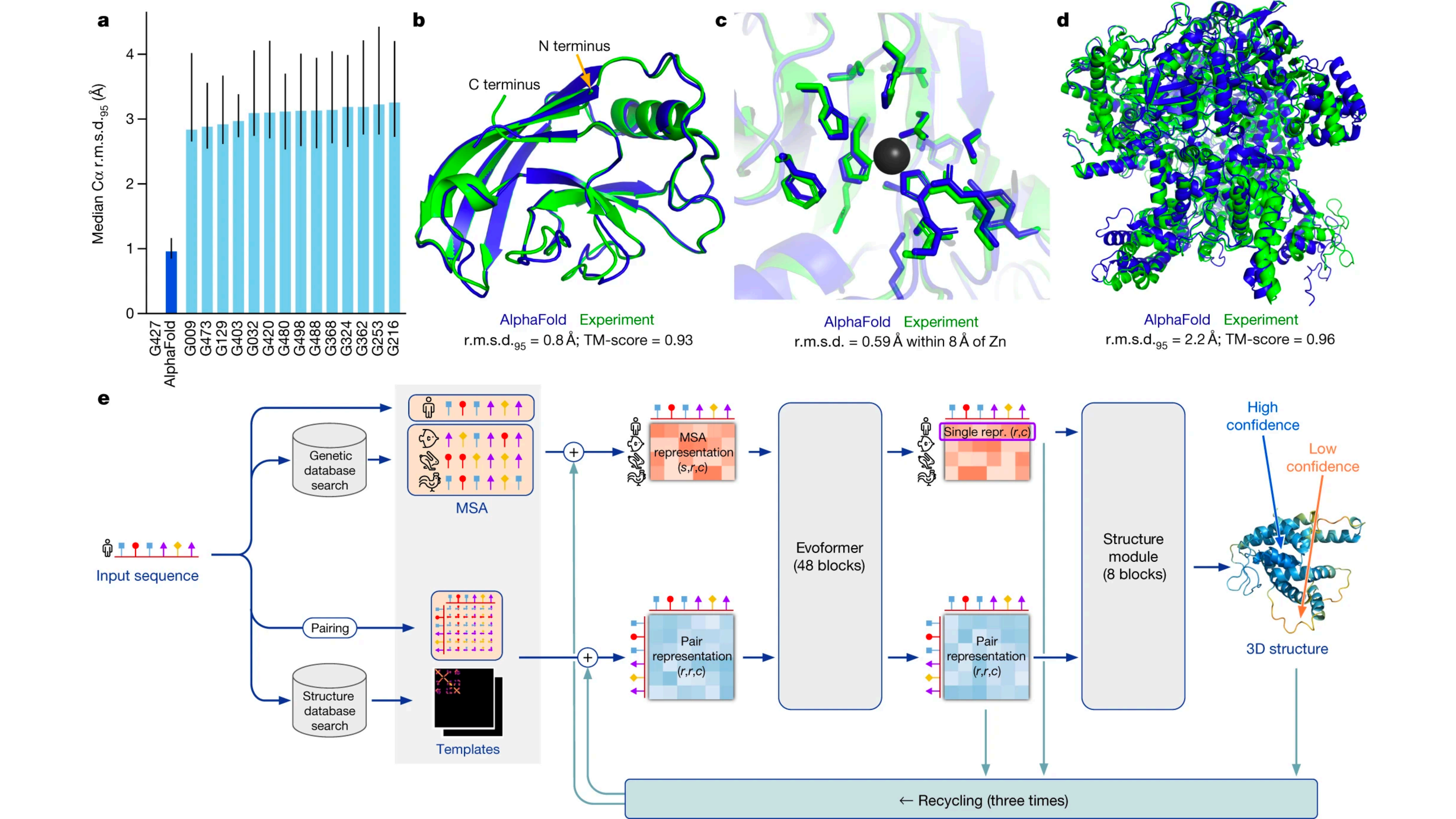
Accurate prediction of protein structures and interactions using a three-track neural network

MINKYUNG BAEK, FRANK DIMAIO, IVAN ANISHCHENKO, JUSTAS DAUPARAS, SERGEY OVCHINNIKOV, GYU RIE LEE, JUE WANG, QIAN CONG, LISA N. KINCH, [...] DAVID BAKER +23 authors Authors Info & Affiliations

SCIENCE · 19 Aug 2021 · Vol 373, Issue 6557 · pp. 871-876 · DOI: 10.1126/science.abj8754

Deep learning takes on protein folding

In 1972, Anfinsen won a Nobel prize for demonstrating a connection between a protein's amino acid sequence and its three-dimensional structure. Since 1994, scientists have competed in the biannual Critical Assessment of Structure Prediction (CASP) protein-folding challenge. Deep learning methods took center stage at CASP14, with DeepMind's Alphafold2 achieving remarkable accuracy. Baek *et al*. explored network architectures based on the DeepMind framework. They used a three-track network to process sequence, distance, and coordinate information simultaneously and achieved accuracies approaching those of DeepMind. The method, RoseTTA fold, can solve challenging x-ray crystallography and cryo–electron microscopy modeling problems and generate accurate models of protein-protein complexes. —VV


a
Median Cα r.m.s.d.95 (Å)
0
1
2
3
4
G427
AlphaFold
G009
G473
G129
G403
G032
G420
G480
G498
G488
G368
G324
G362
G253
G216
b
N terminus
C terminus
AlphaFold Experiment
r.m.s.d.95 = 0.8 Å; TM-score = 0.93
c
AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn
d
AlphaFold Experiment
r.m.s.d.95 = 2.2 Å; TM-score = 0.96
e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Discuss

- What factors increase/decrease
  - your confidence in a predicted structure?
  - the influences of the prediction algorithm?
- Can homology modeling/threading be used to
  - predict the effect of a mutation
    - of a contact with a ligand in a binding site?
    - on a large-scale conformational change?
  - predict the effect of buffer conditions?

References

- Lecture 19 of BIOL 4550 by Chris Bystroff of Rensselaer Polytechnic Institute
- Lab 04 of IIBM3202 Molecular Modeling and Simulation from the Institute for Biological and Engineering at Pontificia Universidad Catolica de Chile
- Chothia C & Lesk AM (1986) _EMBO J_ 5(4), 823–826
- AlphaFold paper: https://www.nature.com/articles/s41586-021-03819-2
- RoseTTA fold paper: https://doi.org/10.1126/science.abj8754