(*In silico*) model building

BCH394P/BCH364C
Systems Biology & Bioinformatics

Caitie McCafferty
clmccafferty@utexas.edu

1

Today's agenda

- Sequence to structure
- Types of *in silico* modeling
  - Comparative modeling (homology modeling)
  - Evolutionary coupling modeling
  - *Ab initio* modeling with neural networks

2

Anfinsen's dogma

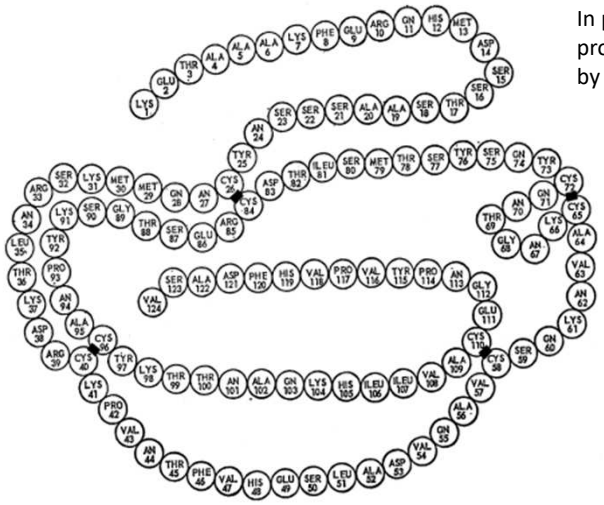


Fig. 1. The amino acid sequence of bovine pancreatic ribonuclease (*50*).

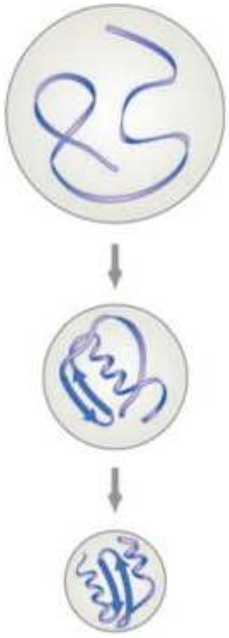
In physiological conditions, a globular protein's native structure is determined by the protein's amino acid sequence.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*.

3

Protein folding problem

1. What is the folding code?
2. What is the folding mechanism?
3. Can we predict a native protein structure from its primary, amino acid sequence?



Dill, K. A., et al. (2008). The protein folding problem. *Annu. Rev. Biophys.*

4

Why might we need to build computational models?

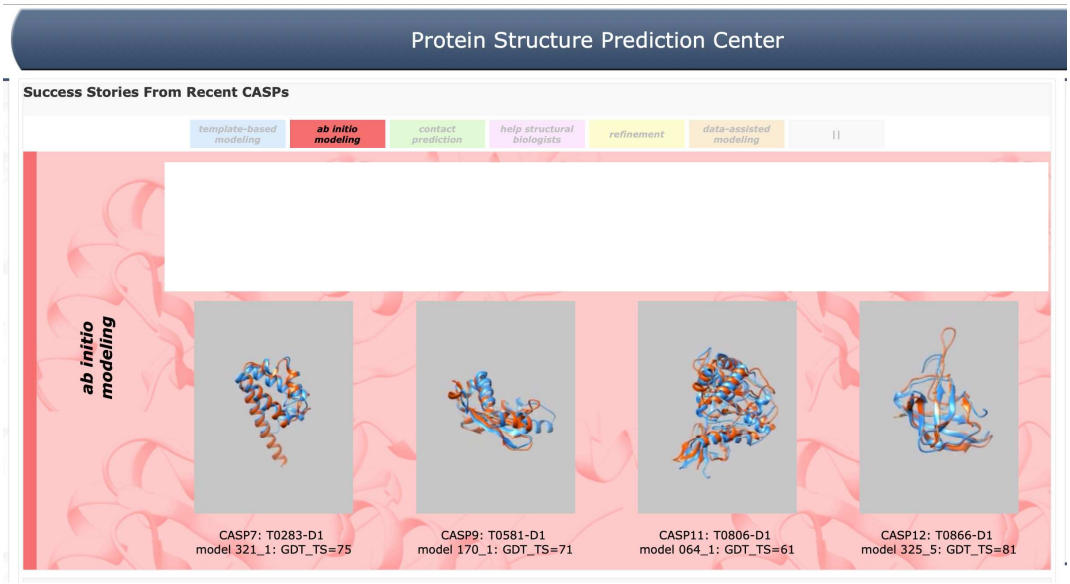
Other considerations: modeling and resolutions based on your need

As with any experiment, keep your intended application in mind…

- do you want to examine a ligand binding site (high resolution)
- or maybe a residue neighborhood (medium resolution)
- or domain boundary definition or even topology classification (low resolution)

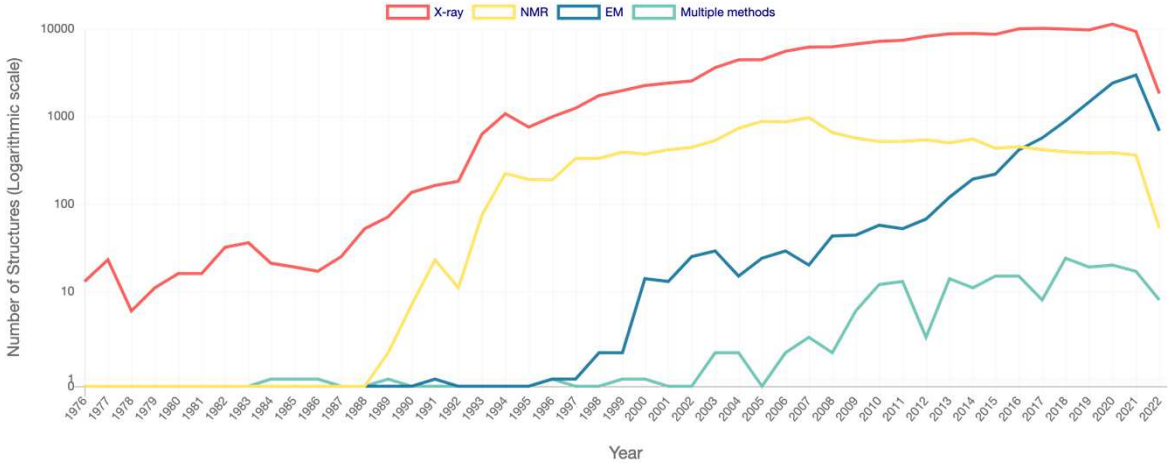
5

CASP competition to evaluate structure prediction



6

Comparative Modeling: some numbers

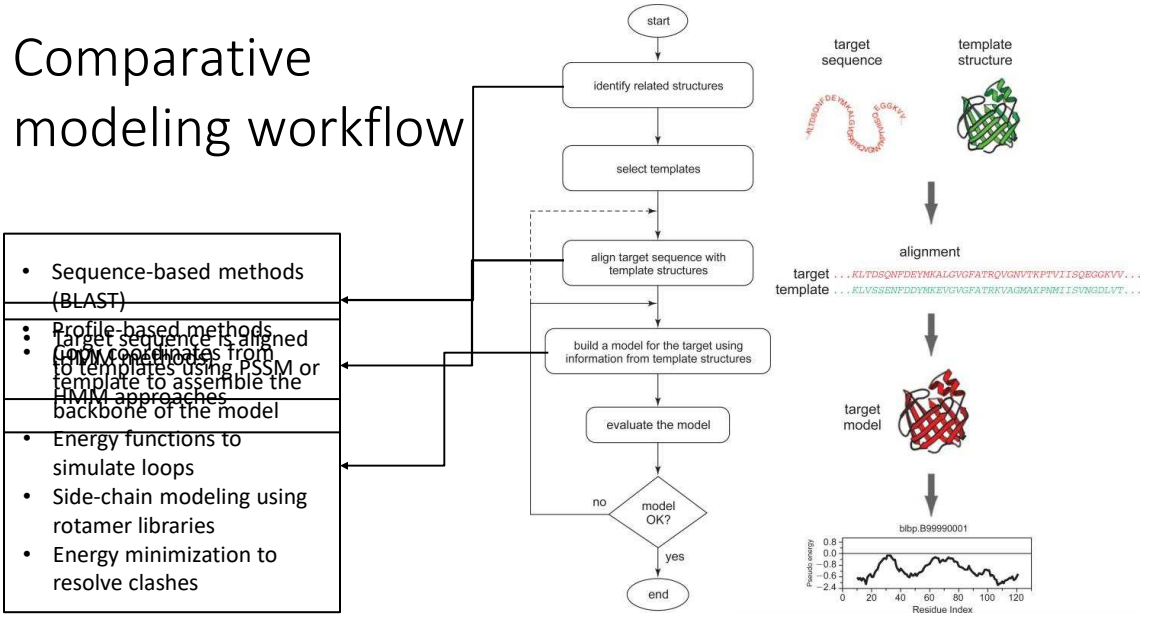


7

Comparative modeling workflow

- Sequence-based methods (BLAST)
- Profile-based methods (PSSM methods)
- HMM methods

- Energy functions to simulate loops
- Side-chain modeling using rotamer libraries
- Energy minimization to resolve clashes



Target sequence is aligned to templates using PSSM or HMM approaches

Coordinates from template to assemble the backbone of the model

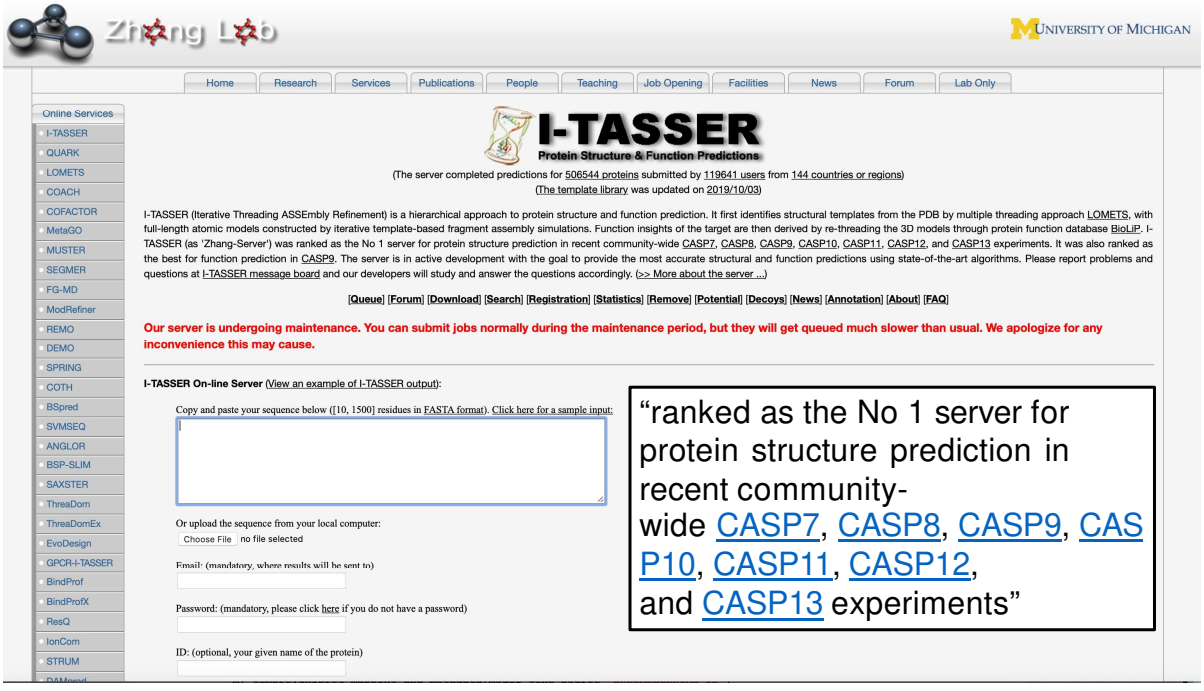
Eswar, N., Webb, B., Marti-Renom, M. A., ... & Sali, A. (2006). *Current protocols in bioinformatics.*

8

I-TASSER for comparative modeling

Zhang, Y. (2007). *Proteins: Structure, Function, and Bioinformatics.*

9



10

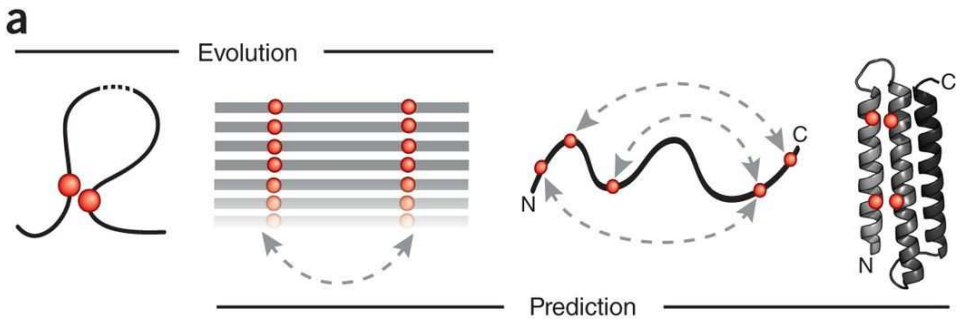
Comparative modeling limitations

Continued reading on comparative modeling

- Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., ... & Sali, A. (2007). Comparative protein structure modeling using MODELLER. *Current protocols in protein science*, *50*(1), 2-9.
- Song, Y., DiMaio, F., Wang, R. Y. R., Kim, D., Miles, C., Brunette, T. J., ... & Baker, D. (2013). High-resolution comparative modeling with RosettaCM. *Structure*, *21*(10), 1735-1742.
- Cavasotto, C. N., & Phatak, S. S. (2009). Homology modeling in drug discovery: current trends and applications. *Drug discovery today*, *14*(13-14), 676-683.
- Larsson, P., Wallner, B., Lindahl, E., & Elofsson, A. (2008). Using multiple templates to improve quality of homology models in automated homology modeling. *Protein Science*, *17*(6), 990-1002.
- Bower, M. J., Cohen, F. E., & Dunbrack Jr, R. L. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *Journal of molecular biology*, *267*(5), 1268-1282.
- Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, *39*(suppl_2), W29-W37.

11

Evolutionary coupling-based modeling



"The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated."

- Sewall Wright

Marks, D. S., Hopf, T.A., & Sander, C. (2012). *Nature biotechnology.*

12

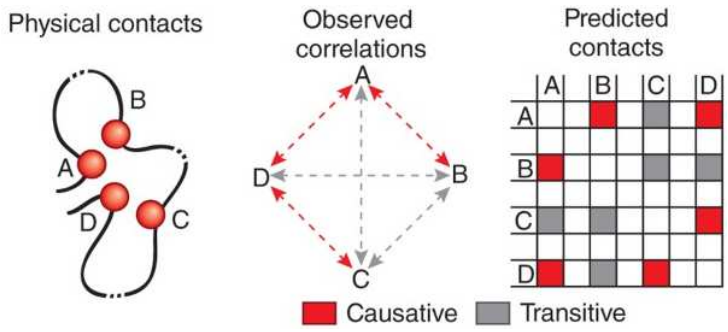
Finding evolutionary covariation

Local statistical models i.e. mutual information

- Assume pairs of residues are statistically independent of other pairs of residues
- Issue: cooperative interaction exist and are crucial in folding

Global statistical models i.e. maximum entropy

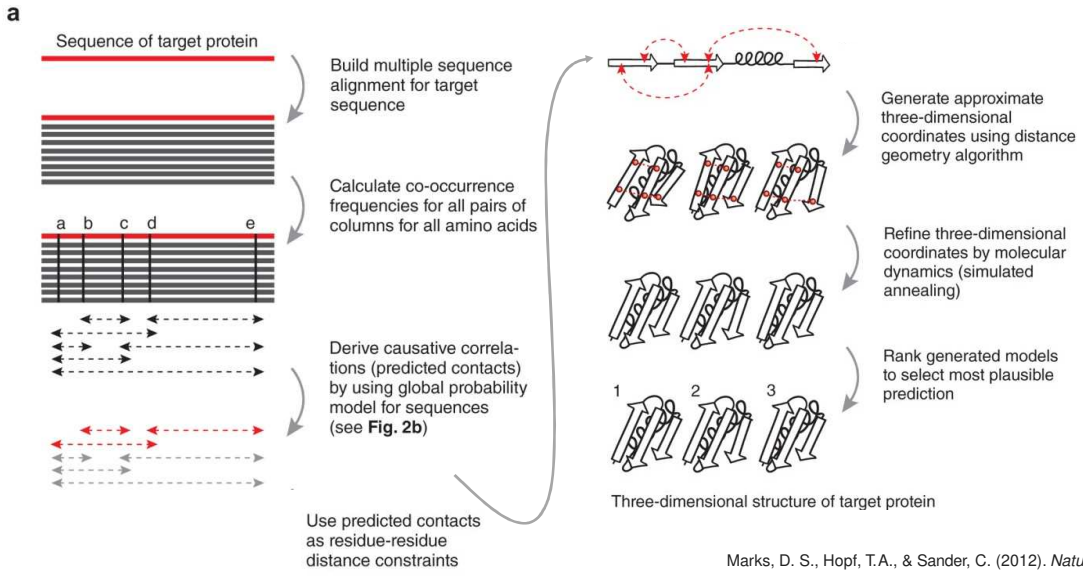
- Correlated pairs of residues are dependent on each other
- Given all pair correlations, which best explain all the others (going from correlation to causation)



Marks, D. S., Hopf, T.A., & Sander, C. (2012). *Nature biotechnology.*

13

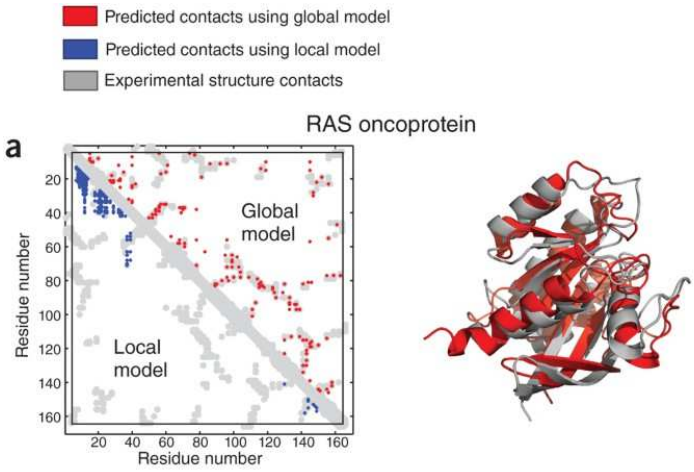
Detecting evolutionary couplings



Marks, D. S., Hopf, T.A., & Sander, C. (2012). *Nature biotechnology.*

14

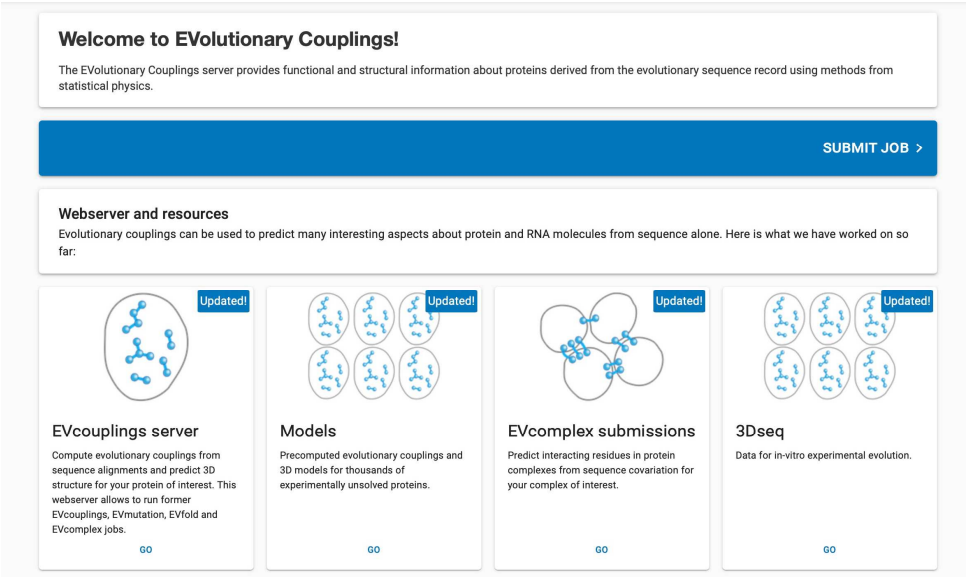
Maximum entropy models for selecting global residue contacts



Marks, D. S., Hopf, T.A., & Sander, C. (2012). *Nature biotechnology*.

15

EVcouplings



16

Evolutionary modeling limitations

Continued reading on EC modeling

- Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., & Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS one*, *6*(12), e28766.
- Tang, Y., Huang, Y. J., Hopf, T. A., Sander, C., Marks, D. S., & Montelione, G. T. (2015). Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nature methods*, *12*(8), 751-754.
- Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., & Marks, D. S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, *149*(7), 1607-1621.
- Kamisetty, H., Ovchinnikov, S., & Baker, D. (2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences*, *110*(39), 15674-15679.
- Ovchinnikov, S., Park, H., Varghese, N., Huang, P. S., Pavlopoulos, G. A., Kim, D. E., ... & Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, *355*(6322), 294-298.

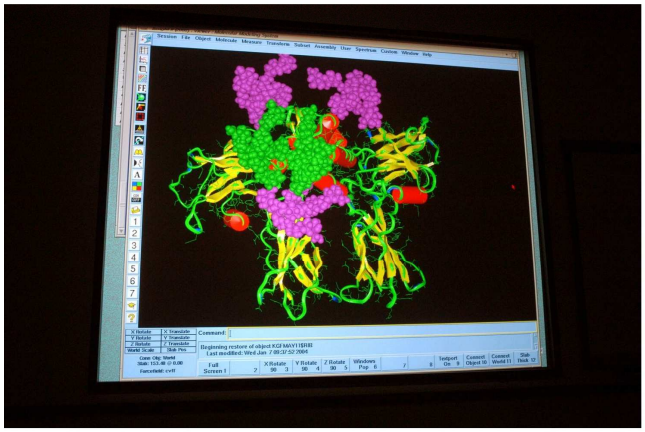
17

How one scientist coped when AI beat him at his life's work

A Harvard biologist on his journey from melancholy to acceptance.

By Sigal Samuel | Feb 15, 2019, 4:10pm EST

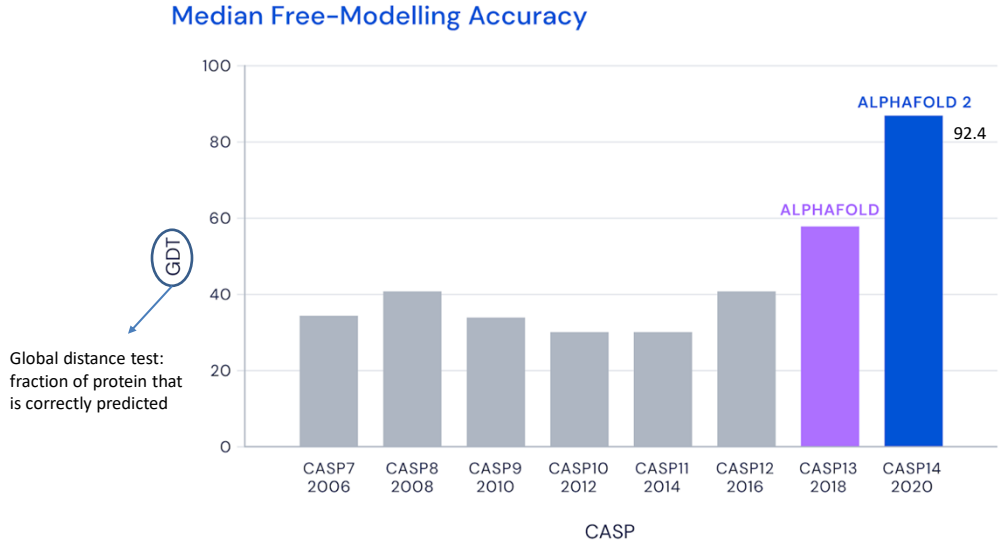
SHARE



A protein bound to a receptor is shown in the Molecular Graphics Lab in California. | Ann Johansson/Corbis via Getty Images

18

Best team in CASP performance over the years



Global distance test: fraction of protein that is correctly predicted

https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

19

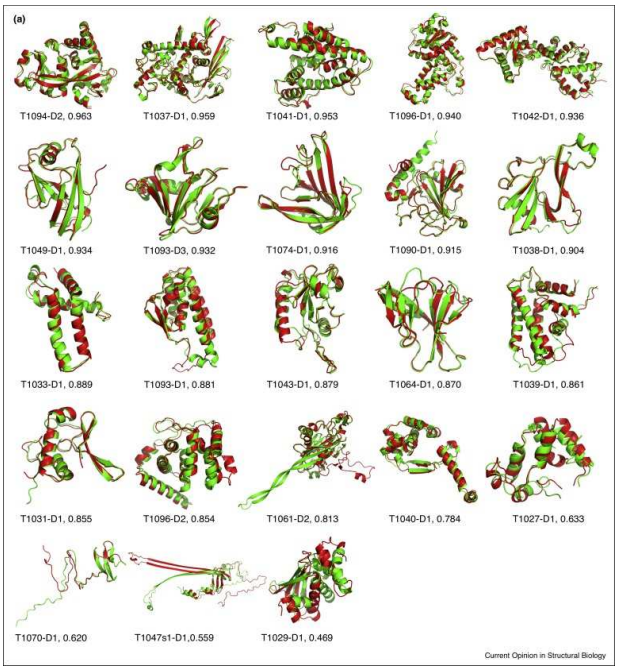
How much better did AF2 do over other methods?

97 total targets

https://moalquraishi.wordpress.com

20

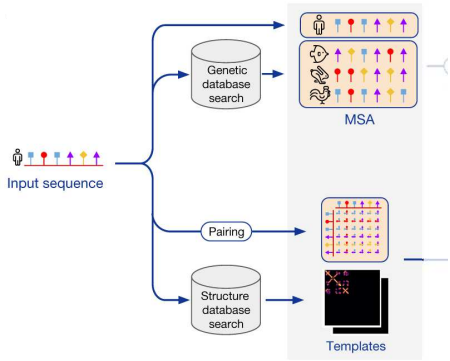
AlphaFold2 at CASP14



Pearce, R., & Zhang, Y. (2021). *Current Opinion in Structural Biology*.

21

This is exciting but how does it work?!



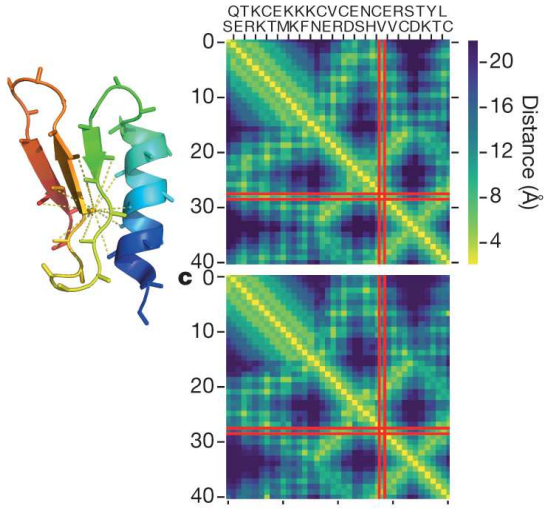
Input sequence is searched against genetic databases to construct a MSA and structure databases to find template structures

This is then used to build a pair representation and an MSA representation

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* (2021).

22

But what is a pair representation?!?



- Distogram is used to map 2D pairwise distances
- Distogram are independent of translations and rotations, so no need to align structures (mush faster)

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... & Hassabis, D. (2020). *Nature*.

23

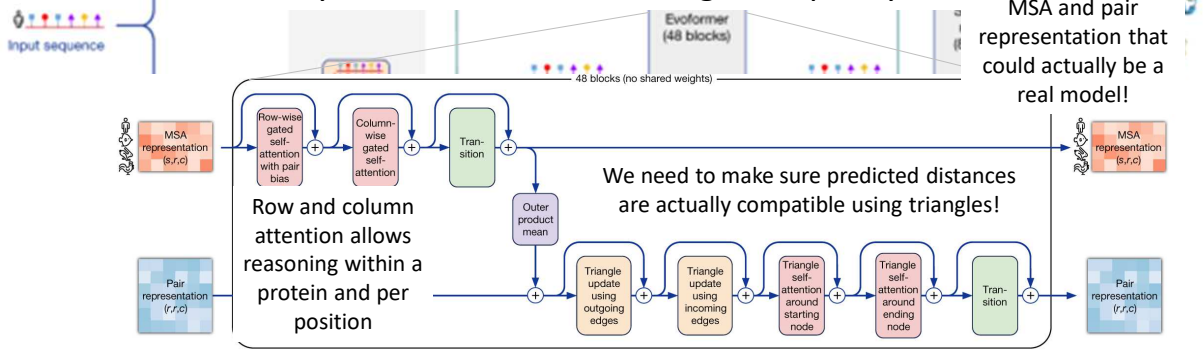
This is exciting, but how does the Evoformer work????

Notice the input and output of the Evoformer are the same

In this process the MSA and pair representations are refined through exchanging information with each other

Does anyone remember the triangle inequality the

The end result is an MSA and pair representation that could actually be a real model!



Row and column attention allows reasoning within a protein and per position

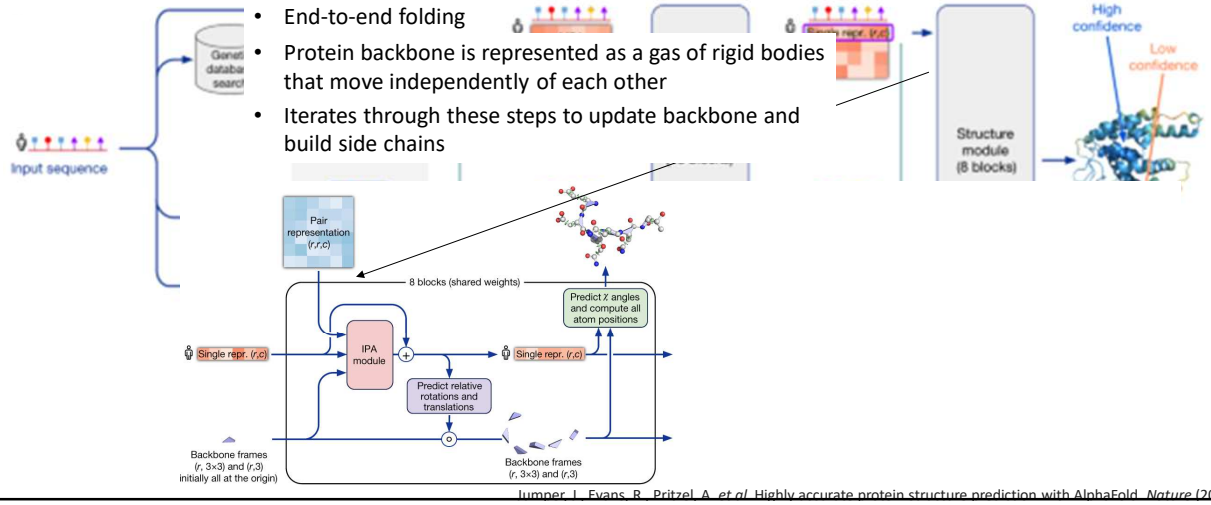
We need to make sure predicted distances are actually compatible using triangles!

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* (2021).

24

This is exciting, but how does the structure module work?!

- End-to-end folding
- Protein backbone is represented as a gas of rigid bodies that move independently of each other
- Iterates through these steps to update backbone and build side chains



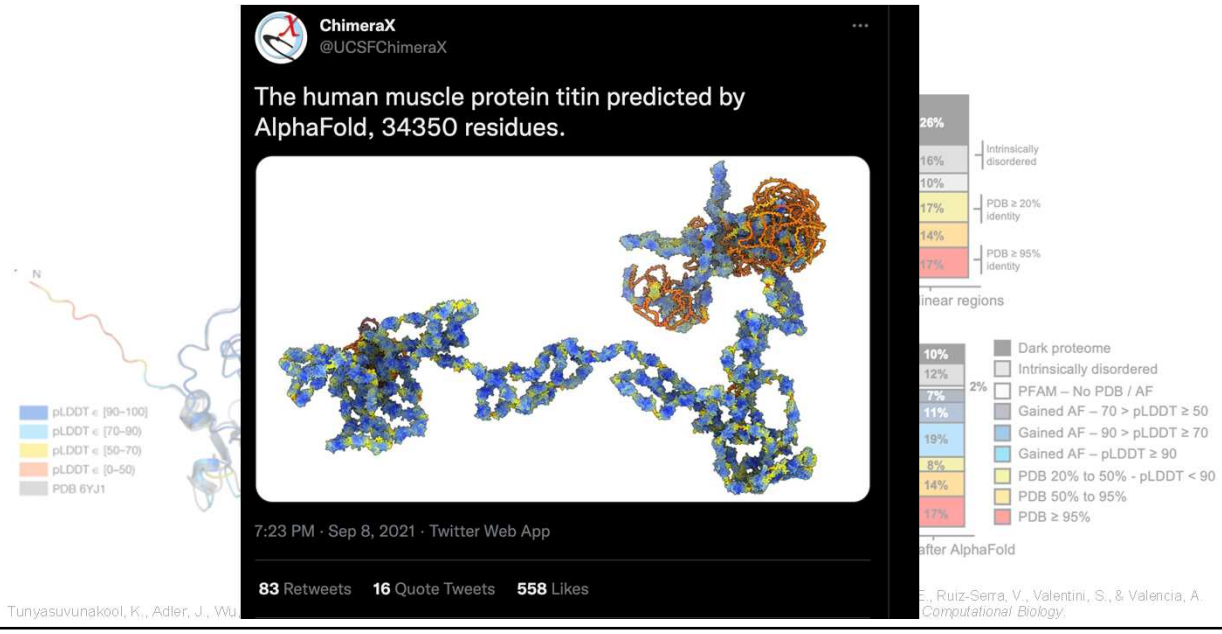
Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* (2021).

25

Okay we can predict protein structure with high accuracy… now what?

26

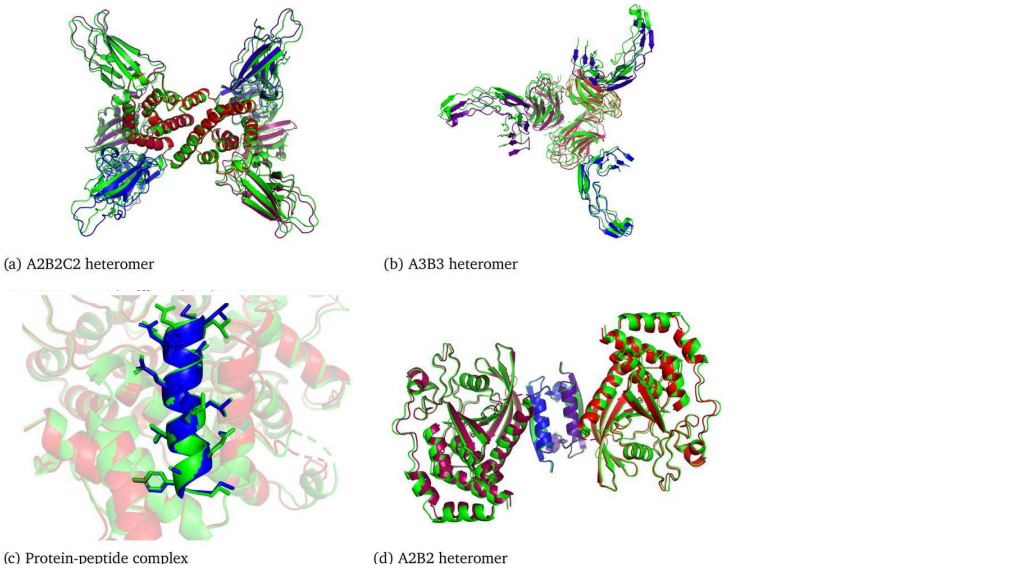
Proteome-wide structure prediction



Tunyasuvunakool, K., Adler, J., Wu ... E., Ruiz-Serra, V., Valentini, S., & Valencia, A. *Computational Biology.*

27

AlphaFold Multimer for protein complexes



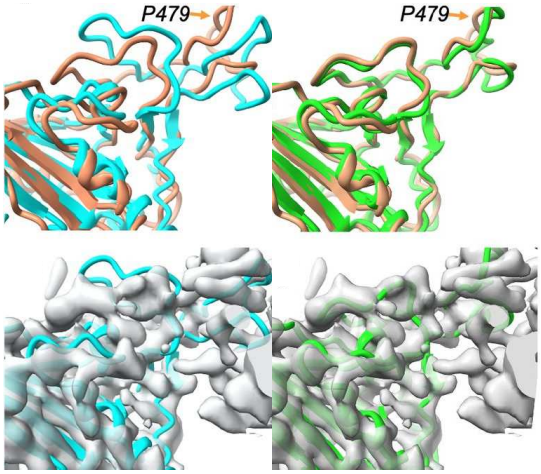
(a) A2B2C2 heteromer (b) A3B3 heteromer

(c) Protein-peptide complex (d) A2B2 heteromer

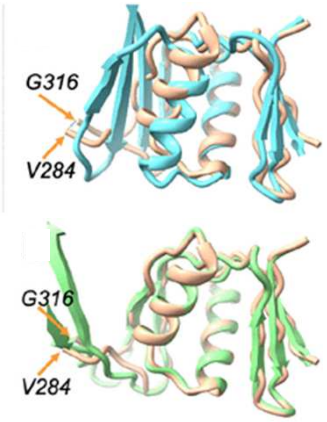
Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A. W., Green, T., ... & Hassabis, D. (2021). *BioRxiv.*

28

Combining structure prediction with other structure determination methods



SARS-CoV-2 spike protein receptor



Flagellar basal body

Terwilliger, T. C., Poon, B. K., Afonine, P., Schlicksup, C. J., Croll, T. I., Millan-Nebot, C., ... & Adams, P. D. (2022). *bioRxiv*.

29

Ab initio NN modeling limitations

Continued reading on *Ab initio* modeling

- AlQuraishi, M. (2019). End-to-end differentiable learning of protein structure. *Cell systems*, *8*(4), 292-301.
- Pearce, R., & Zhang, Y. (2021). Deep learning techniques have significantly impacted protein structure prediction and protein design. *Current Opinion in Structural Biology*, *68*, 194-207.
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706-710.
- AlQuraishi, M. (2019). AlphaFold at CASP13. *Bioinformatics*, *35*(22), 4862-4865.
- Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, *80*(7), 1715-1735.

30

Modeling servers

- Comparative modeling
  - https://zhanglab.ccmb.med.umich.edu/I-TASSER/
- Evolutionary modeling
  - https://evcouplings.org/job
- AlphaFold2 modeling
  - https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/beta/AlphaFold2_advanced.ipynb

31

Thank you!

Questions?

32