Ab Initio Protein Structure Prediction: AlphaFold

Ab initio Protein Structure Prediction

Ab initio prediction before AlphaFold

Ab initio prediction: Predicting Contacts

AlphaFold 1

AlphaFold 2

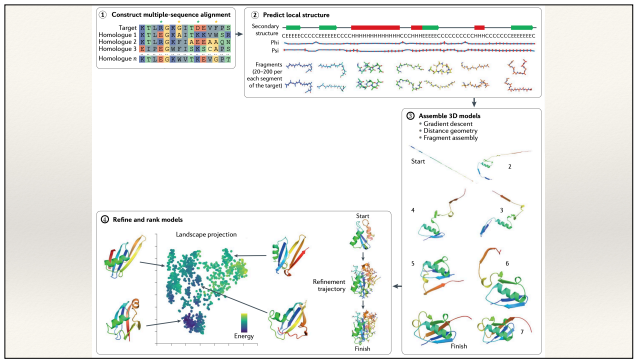
Ab initio Protein Structure Prediction

Ab initio prediction before AlphaFold

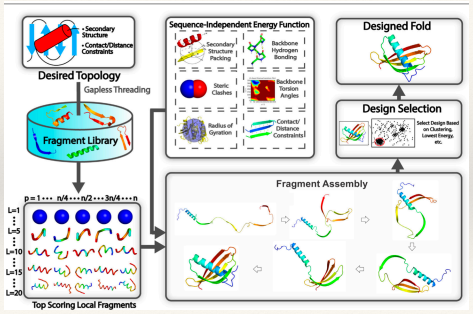
Ab initio prediction: Predicting Contacts

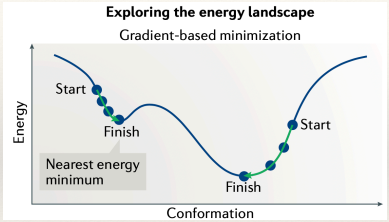
AlphaFold 1

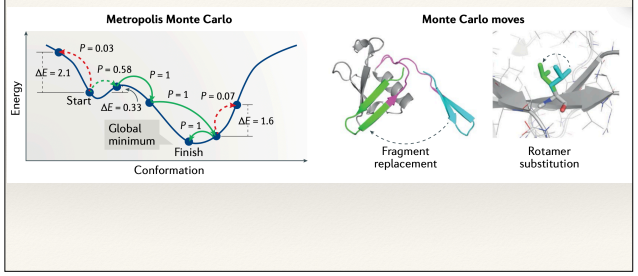
AlphaFold 2



Fragment based methods







Ab initio Protein Structure Prediction

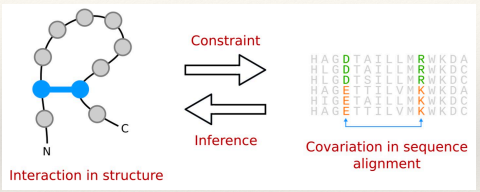
Ab initio prediction before AlphaFold

Ab initio prediction: Predicting Contacts

AlphaFold 1

AlphaFold 2

Predicting residue contacts



Predicting residue contacts

1. Given a multiple sequence alignment (MSA):

X_1 H A G D T A I L L M R W K D A
H L G D T A I L L M R W K D C
H L G D T S I L L M R W K D C
H A G E T T I L V M K W K D A
H I G E T A I L L M K W K D C
X_N H A G E T T I L V M K W K D C

2. Compute "mean" sequence and covariance matrix:

$$\overline{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$$

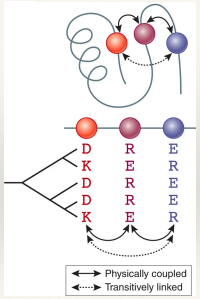
$$\overline{C} = C(MSA, \overline{X}) = \frac{1}{N}\sum_{n=1}^{N} (X_n - \overline{X})^T (X_n - \overline{X})$$

3. Compute contact J(i,j)

J(i, j) = C(i, j)?

Predicting residue contacts

No! We need to pay attention to indirect effects:



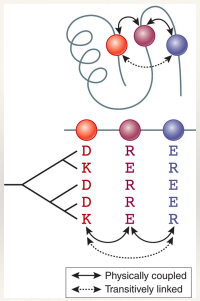
Predicting residue contacts

No! We need to pay attention to indirect effects:

Gaussian model:

Each sequence X_i in the MSA is drawn from a multivariate Gaussian distribution characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ, with the probability:

$$P(X_n|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{D}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(X_n - \boldsymbol{\mu})^T \Sigma^{-1} (X_n - \boldsymbol{\mu})\right]$$



Predicting residue contacts

No! We need to pay attention to indirect effects:

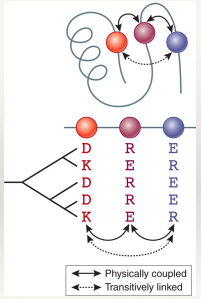
Gaussian model:

Each sequence X_i in the MSA is drawn from a multivariate Gaussian distribution characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ, with the probability:

$$P(X_n|\boldsymbol{\mu},\Sigma) = (2\pi)^{-\frac{q}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(X_n-\boldsymbol{\mu})^T\Sigma^{-1}(X_n-\boldsymbol{\mu})\right]$$

Assuming that the N sequences in the MSA are statistical independent, the probability, or likelihood of the data under this model is given by

$$P(MSA|\boldsymbol{\mu},\Sigma) = \prod_{n=1}^{N} P(X_n|\boldsymbol{\mu},\Sigma)$$



Predicting residue contacts

No! We need to pay attention to indirect effects:

Gaussian model:

Each sequence X_i in the MSA is drawn from a multivariate Gaussian distribution characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ, with the probability:

$$P(X_n|\boldsymbol{\mu},\Sigma) = (2\pi)^{-\frac{q}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(X_n-\boldsymbol{\mu})^T\Sigma^{-1}(X_n-\boldsymbol{\mu})\right]$$

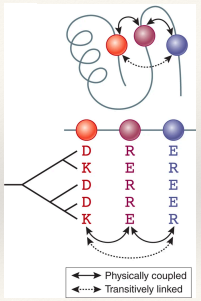
Assuming that the N sequences in the MSA are statistical independent, the probability, or likelihood of the data under this model is given by

$$P(MSA|\boldsymbol{\mu},\Sigma) = \prod_{n=1}^{N} P(X_n|\boldsymbol{\mu},\Sigma)$$

Using the maximum likelihood estimator for this probability

$$\boldsymbol{\mu} = \overline{X}$$

$$\Sigma = \overline{C} = C(MSA,\overline{X})$$



Predicting residue contacts

No! We need to pay attention to indirect effects:

Gaussian model:

$$P(X_n|\boldsymbol{\mu},\Sigma) = (2\pi)^{-\frac{q}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(X_n-\boldsymbol{\mu})^T\Sigma^{-1}(X_n-\boldsymbol{\mu})\right]$$

$$\boldsymbol{\mu} = \overline{X} \qquad \Sigma = \overline{C} = C(MSA,\overline{X})$$

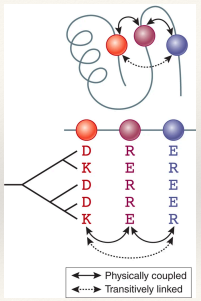
Note that:

$$(X_n-\boldsymbol{\mu})^T\Sigma^{-1}(X_n-\boldsymbol{\mu}) = \sum_{k=1}^{N}\sum_{l=1}^{N}(X_k-\mu_k)(\Sigma^{-1})(k,l)(X_l-\mu_l)$$

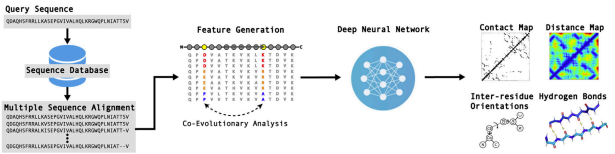
This shows that $(\Sigma^{-1})(k,l)$ serves as a coupling between positions k and l in the MSA.

Therefore:

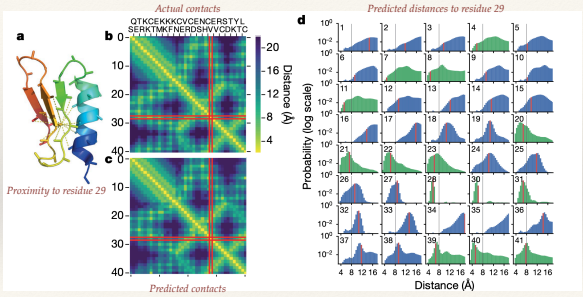
$$J = \Sigma = (C(MSA,\overline{X}))^{-1}$$



Predicting residue contacts



Predicting residue contacts: How well does it work?



Ab initio Protein Structure Prediction

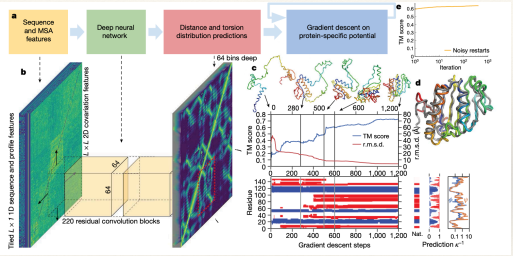
Ab initio prediction before AlphaFold

Ab initio prediction: Predicting Contacts

AlphaFold 1

AlphaFold 2

AlphaFold 1



AlphaFold1

AlphaFold 1

Reminder:

To compare two sets of points (atoms) A=[a_1, a_2, …a_N] and B=[b_1, b_2, …b_N]:

-Define a 1-to-1 correspondence between A and B

for example, a_i corresponds to b_i for all i in [1,N]

-Compute RMS as:

$$RMS(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d(a_i,b_i)^2}$$

Compute TM score:

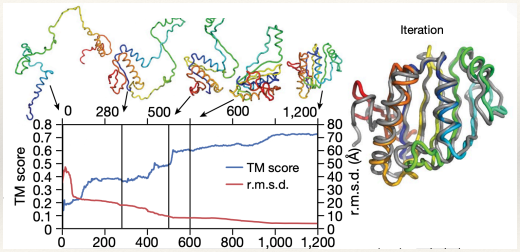
$$TM(A,B) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1+\left(\frac{d(a_i,b_i)}{d_0(N)}\right)^2} \quad \text{with } d_0(N) = 1.24\sqrt[3]{N-15} - 1.8$$

$d(a_i,b_i)$ is the Euclidian distance between a_i and b_i after optimal alignment of B onto A

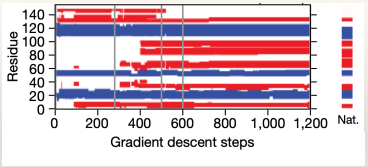
RMS: the lower, the better

TM: between [0,1]; the higher the better

AlphaFold 1

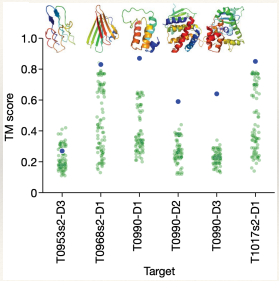


AlphaFold 1



Helix in blue, strand in red

AlphaFold 1: Success



Ab initio Protein Structure Prediction

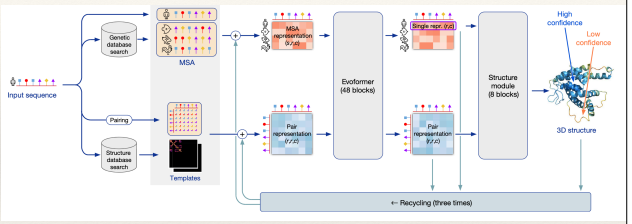
Ab initio prediction before AlphaFold

Ab initio prediction: Predicting Contacts

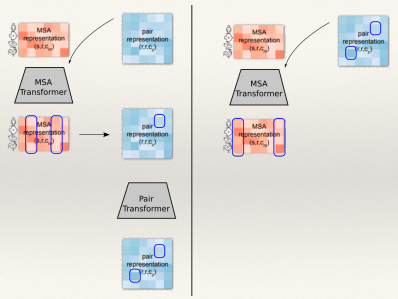
AlphaFold 1

AlphaFold 2

AlphaFold 2

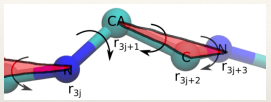


AlphaFold 2: some intuition



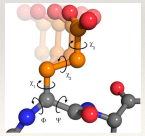
AlphaFold 2: the structure module

Predicting backbone:
the residues form a gas soup of triangles whose relative positions are characterized by affine transformation

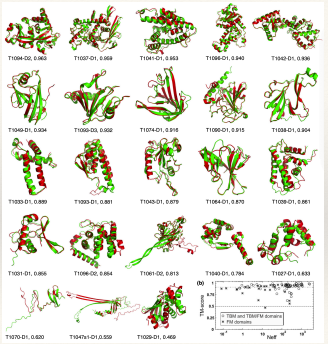


$$\mathbf{M} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Predicting side chains:

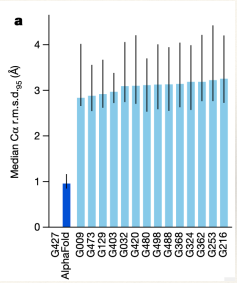


Successes at CASP14



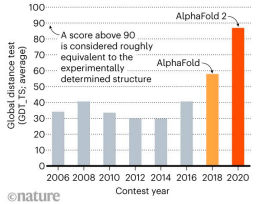
TBM: template-based modeling
FM: free modeling

Successes at CASP14



STRUCTURE SOLVER

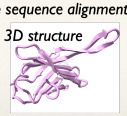
DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



©nature

Training

- *Sequence*
- *Multiple sequence alignment*
- *3D structure*



21 million parameters

Prediction

- *Sequence*
- *Multiple sequence alignment*

21 million parameters

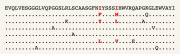
Focus attention on important relationships

3D prediction

Confidence estimates

Credit: Tom Terwilliger, Los Alamos NL

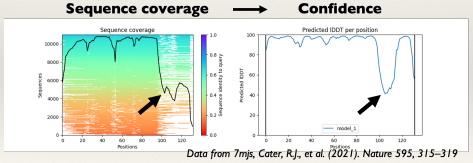
Multiple sequence alignment



*Residues that **co-vary** are probably close in 3D structure*

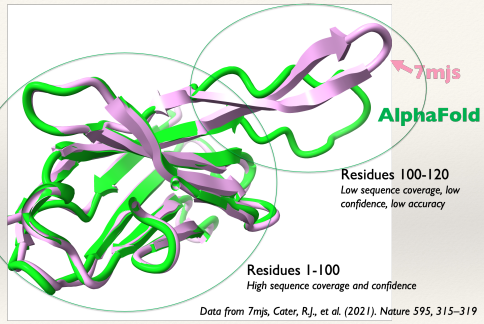
All sequences in alignment should be compatible with the right structure

Sequence coverage → Confidence



Data from 7mjs, Cater, R.J., et al. (2021). Nature 595, 315–319

Credit: Tom Terwilliger, Los Alamos NL



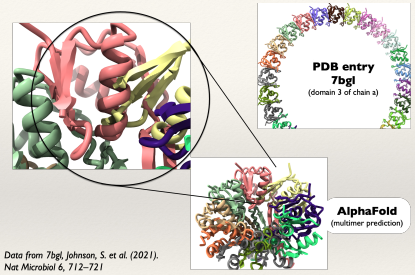
Residues 100-120
Low sequence coverage, low confidence, low accuracy

Residues 1-100
High sequence coverage and confidence

Data from 7mjs, Cater, R.J., et al. (2021). Nature 595, 315–319

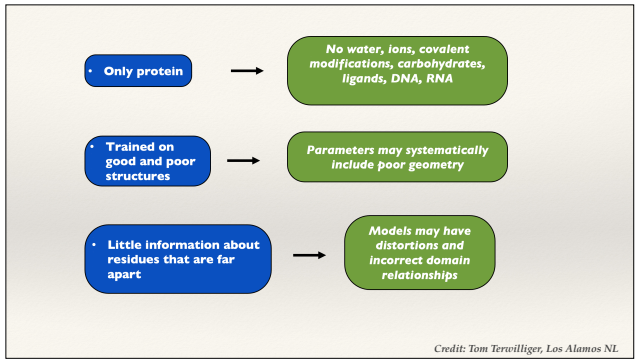
Credit: Tom Terwilliger, Los Alamos NL

Multimeric proteins



Data from 7bgl, Johnson, S. et al. (2021). Nat Microbiol 6, 712–721

Credit: Tom Terwilliger, Los Alamos NL


• Only protein
No water, ions, covalent modifications, carbohydrates, ligands, DNA, RNA
• Trained on good and poor structures
Parameters may systematically include poor geometry
• Little information about residues that are far apart
Models may have distortions and incorrect domain relationships
Credit: Tom Terwilliger, Los Alamos NL