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
Exploring the energy landscape

Gradient-based minimization

Start → Nearest energy minimum → Finish → Start
Ab initio Protein Structure Prediction

Ab initio prediction before AlphaFold

Ab initio prediction: Predicting Contacts

AlphaFold 1

AlphaFold 2
Predicting residue contacts

Interaction in structure

Constraint

Inference

Covariation in sequence alignment
Predicting residue contacts

1. Given a multiple sequence alignment (MSA):

\[
X_1 = \begin{array}{cccccccc}
H & A & G & D & T & A & I & L & L & M & R & W & K & D & A \\
H & L & G & D & T & A & I & L & L & M & R & W & K & D & C \\
H & L & G & D & T & S & I & L & L & M & R & W & K & D & C \\
H & A & G & E & T & T & I & L & V & M & K & W & K & D & A \\
H & I & G & E & T & T & I & L & M & K & W & K & D & C \\
H & A & G & E & T & T & I & L & V & M & K & W & K & D & C
\end{array}
\]

\[
X_N
\]

2. Compute “mean” sequence and covariance matrix:

\[
\bar{X} = \frac{1}{N} \sum_{n=1}^{N} X_n
\]

\[
C = C(\text{MSA}, \bar{X}) = \frac{1}{N} \sum_{n=1}^{N} (X_n - \bar{X})^T (X_n - \bar{X})
\]

3. Compute contact \( J(i,j) \)

\[
J(i,j) = C(i,j)?
\]
No! We need to pay attention to indirect effects:
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 $\mu$ and a covariance matrix $\Sigma$, with the probability:

$$P(X_n | \mu, \Sigma) = (2\pi)^{-\frac{L}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} (X_n - \mu)^T \Sigma^{-1} (X_n - \mu) \right]$$
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 $\mu$ and a covariance matrix $\Sigma$, with the probability:

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

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

$$P(\text{MSA} | \mu, \Sigma) = \prod_{n=1}^{N} P(X_n | \mu, \Sigma)$$
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 $\mu$ and a covariance matrix $\Sigma$, with the probability:

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

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

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

Using the maximum likelihood estimator for this probability

$$\mu = \bar{X}$$

$$\Sigma = \bar{C} = C(MSA, \bar{X})$$
Predicting residue contacts

No! We need to pay attention to indirect effects:

Gaussian model:

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

\[ \mu = \bar{X} \quad \Sigma = \bar{C} = C(MSA, \bar{X}) \]

Note that:

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

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

Therefore:

\[ J = \Sigma = (C(MSA, \bar{X}))^{-1} \]
Predicting residue contacts
Predicting residue contacts: How well does it work?

Actual contacts

Predicted contacts

Proximity to residue 29

Predicted distances to residue 29

CASP target T0995
Ab initio Protein Structure Prediction

Ab initio prediction before AlphaFold

Ab initio prediction: Predicting Contacts

AlphaFold 1

AlphaFold 2
AlphaFold 1

(a) Sequence and MSA features → Deep neural network → Distance and torsion distribution predictions → Gradient descent on protein-specific potential

(b) Tiled L x 1D sequence and profile features → L x L covariation features → 220 residual convolution blocks

(c) TM score and r.m.s.d. plots over gradient descent steps

(d) Predicted protein structures

(e) Noisy restarts vs. iteration plot
Reminder:

To compare two sets of points (atoms) \( A = \{a_1, a_2, \ldots, a_N\} \) and \( B = \{b_1, b_2, \ldots, b_N\} \):

- 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:

\[
TM(A, B) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{1 + \left( \frac{d(a_i, b_i)}{d_0(N)} \right)^2}
\]

with \( d_0(N) = 1.24\sqrt[3]{\frac{N}{N-15}} - 1.8 \)

\( d(a_i, b_i) \) 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

The image displays a graph showing the progress of AlphaFold 1 over iterations. The graph includes a plot of TM score and r.m.s.d. values against iteration number. As the iterations progress from 0 to 1,200, both TM score and r.m.s.d. values are monitored, illustrating the improvement in model accuracy with each iteration.
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

Input sequence

Genetic database search

Pairing

Structure database search

MSA

Template

MSA representation (r,s,c)

Evoformer (48 blocks)

Single repr. (r,c)

Structure module (8 blocks)

Pair representation (r,s,c)

Pair representation (r,s,c)

Recycling (three times)

High confidence

Low confidence

3D structure
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

Predicting side chains:
Successes at CASP14

TBM: template-based modeling
FM: free modeling
Successes at CASP14

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.
Training

- Sequence
- Multiple sequence alignment
  - 3D structure

Prediction

- Sequence
- Multiple sequence alignment

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**  \rightarrow  **Confidence**

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

Credit: Tom Terwilliger, Los Alamos NL
Residues 1-100
High sequence coverage and confidence

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

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

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

PDB entry 7bgl
(domain 3 of chain a)

AlphaFold
(multimer prediction)

Credit: Tom Terwilliger, Los Alamos NL
- Only protein

- Trained on good and poor structures

- Little information about residues that are far apart

- No water, ions, covalent modifications, carbohydrates, ligands, DNA, RNA

- Parameters may systematically include poor geometry

- Models may have distortions and incorrect domain relationships