JBC REVIEWS ASBMB AWARD ARTICLE

Toward the solution of the protein structure prediction problem

Received for publication, April 20, 2021, and in revised form, June 7, 2021 Published, Papers in Press, June 11, 2021,
https://doi.org/10.1016/j.jbc.2021.100870

Robin Pearce[1] **and Yang Zhang**[1,2,*]

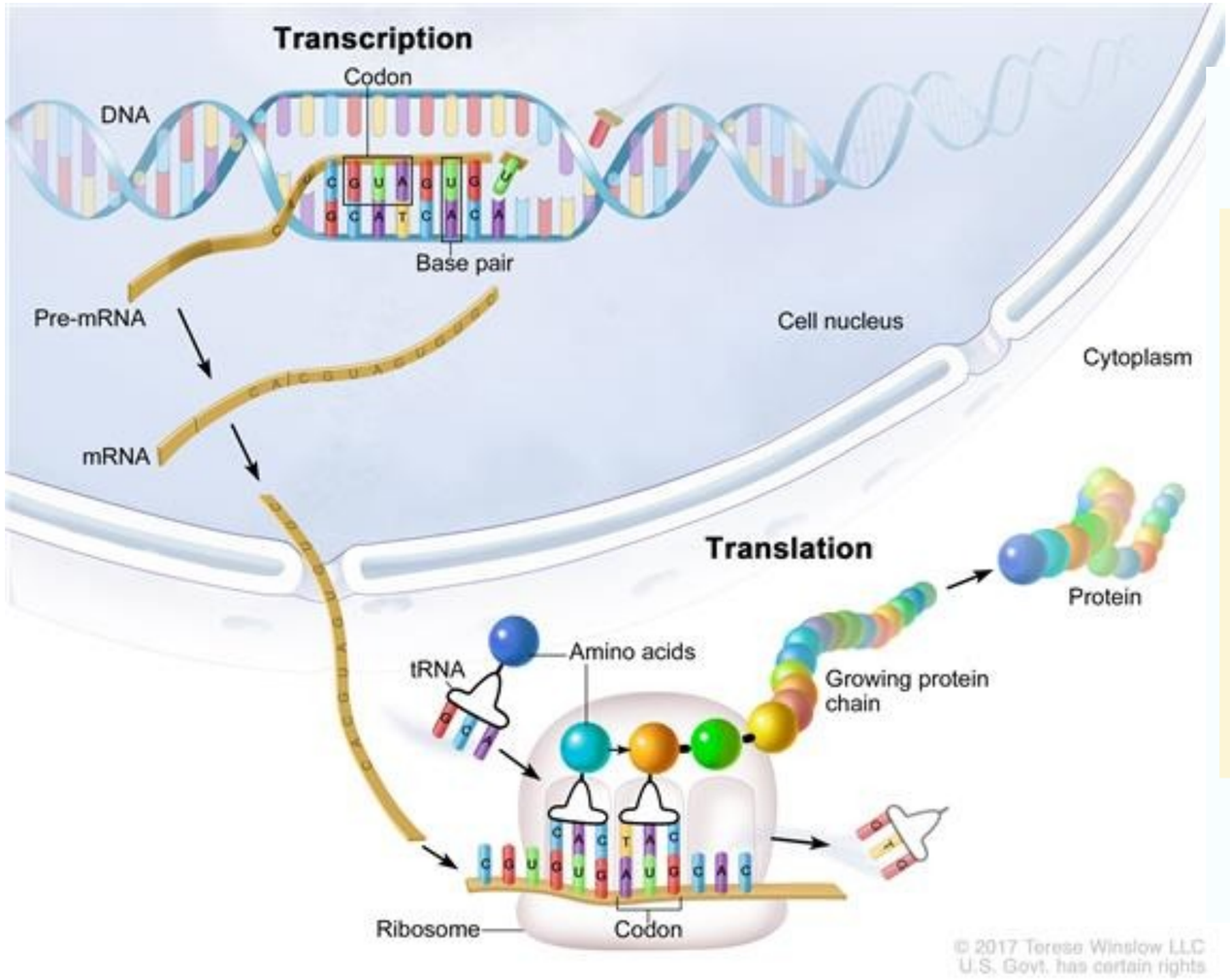
From the [1]*Department of Computational Medicine and Bioinformatics,* [2]*Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan, USA*

Edited by Wolfgang Peti

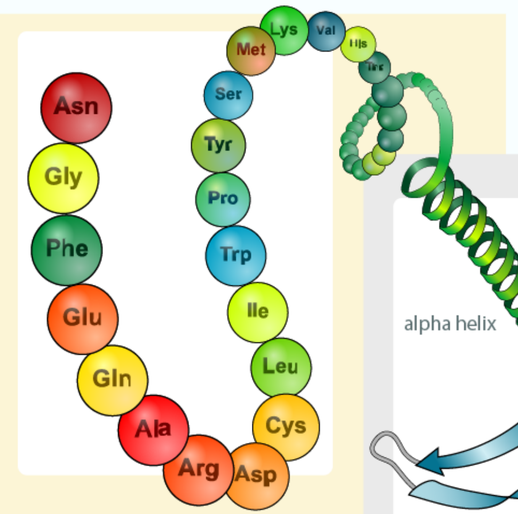
Structures predictions and molecular docking

Structural Bioinformatics

Proteins are encoded in DNA, produced by ribosomes as linear polymers of amino acids, and have a hierarchical structure.

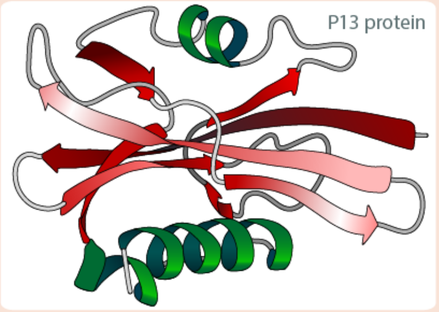


Structure prediction

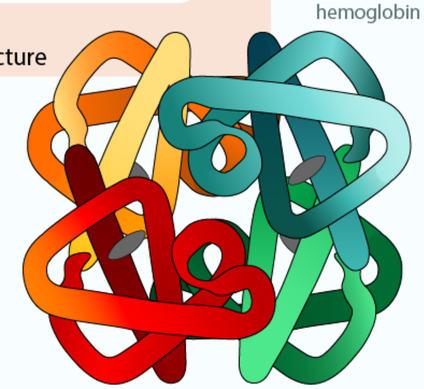


Primary structure
amino acid sequence

Secondary structure
regular sub-structures



Tertiary structure
three-dimensional structure



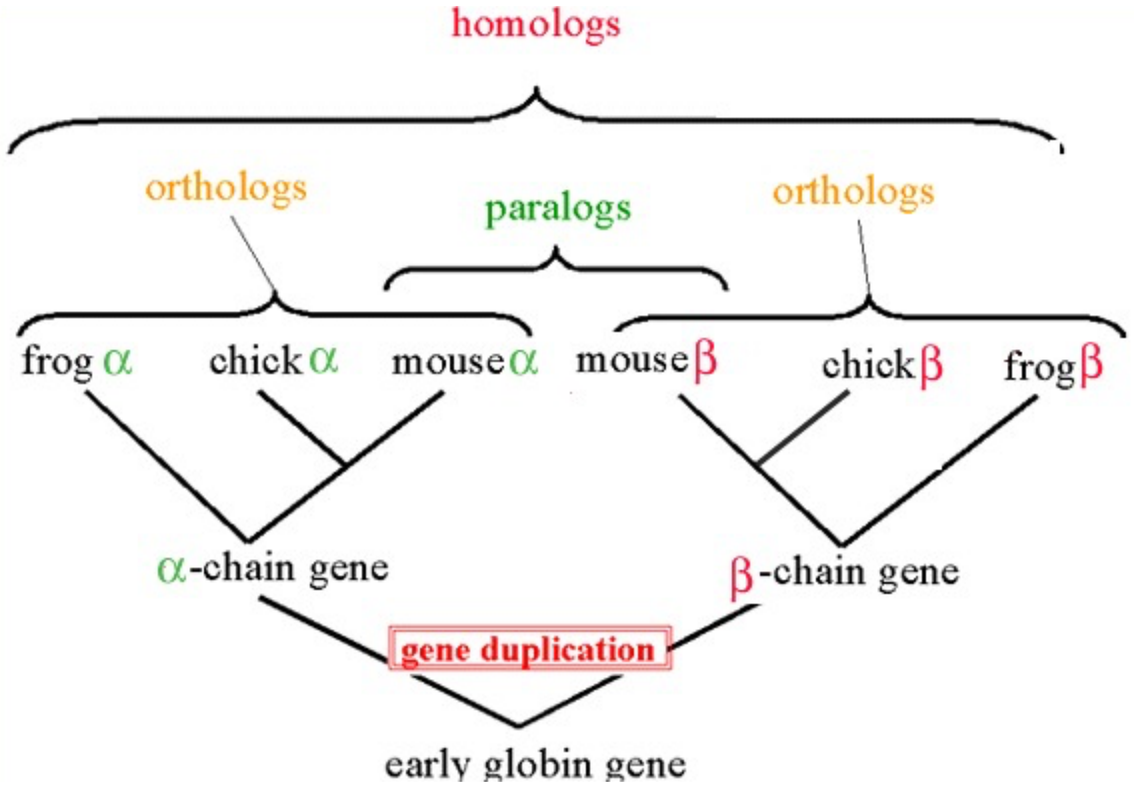
Quaternary structure
complex of protein molecules

Protein Design

Protein structure and function are intimately linked

Upon natural evolution, through sequence modifications, structural motifs performing specific functions have been reused and adapted to work in different contexts.

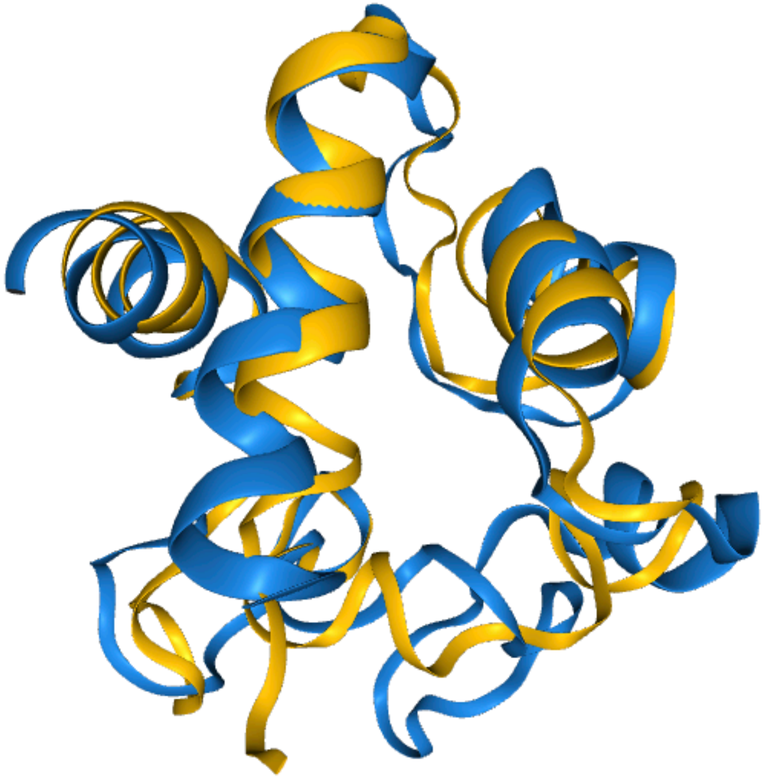
As a result if one takes proteins performing specific functions, even in very unrelated organisms, they still show high structural similarity and some sequence similarity.



Cytochrome C: *human* vs *thermus thermophilus* (23% seq identity)

```
Q    4 EKGKKIFIMKCSQCHTVE--KGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGII-WGEDTLMEYLENPKKYIPGTKM
       E+G+++F   C+ CH V          GP L     R +                GI+    + L  ++ +P    PG KM
T  219 ERGQQVFQQNCAACHGVARSMPPAV-IGPELGLWGNRTSL----------GA--GIVENTPFNI KAWTRDPAGMKPGVKM

Q   81 IFVGI--KKKEERADLIAYLKKATNE
          G     +E+   L+ YL+    E
T  285 --PGFPQLSEEDLDALVRYLEGLKVE
```



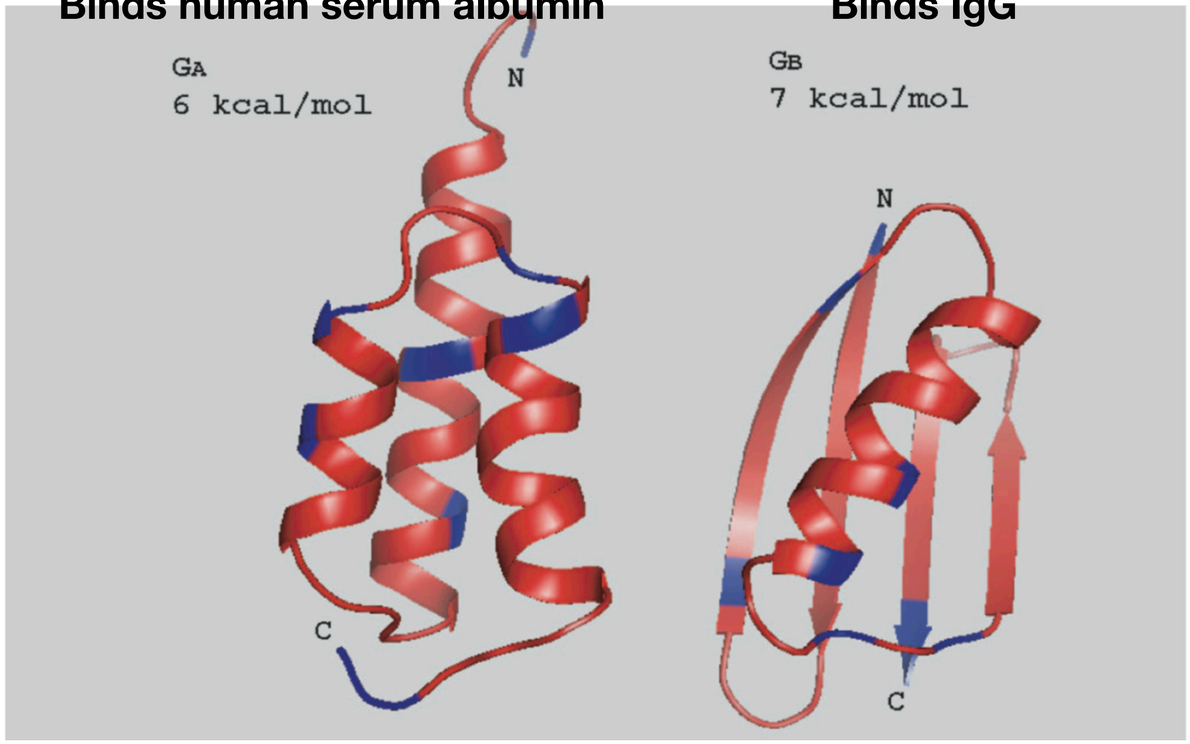
Ideally a structure superposition should allow us to observe how sequences have evolved, in practice since we do not know enough structure, we align sequences and when aligned sequences are similar we hypothesize structure and function similarity.

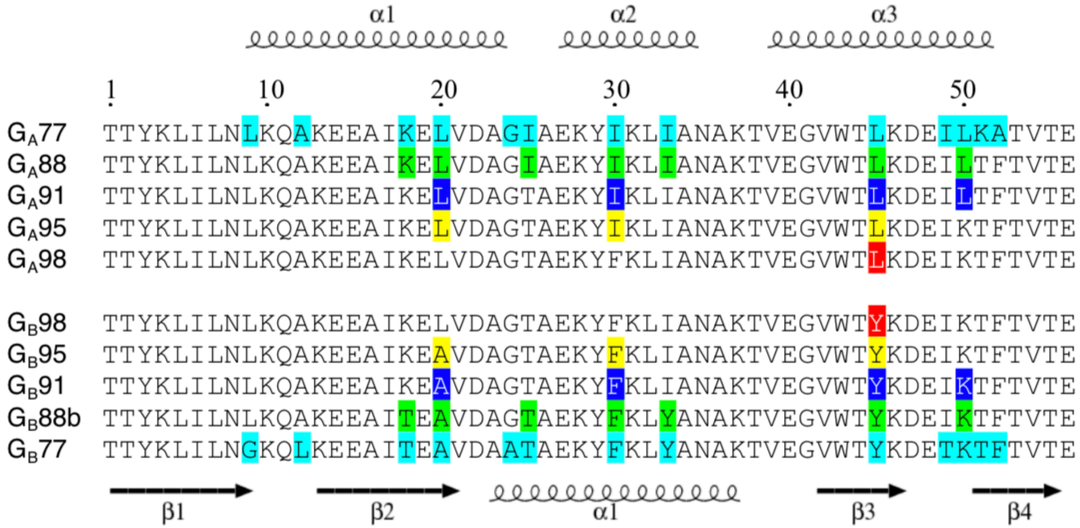
More Conserved -> Function <-> Structure <-> Sequence <- Less Conserved

...yet small sequence variations can lead to dramatic differences.

The correlation of protein structures and sequences in nature is not a necessary codition imposed by chemistry, indeed it is possible to design proteins with different structures and very similar sequences:

Binds human serum albumin **Binds IgG**





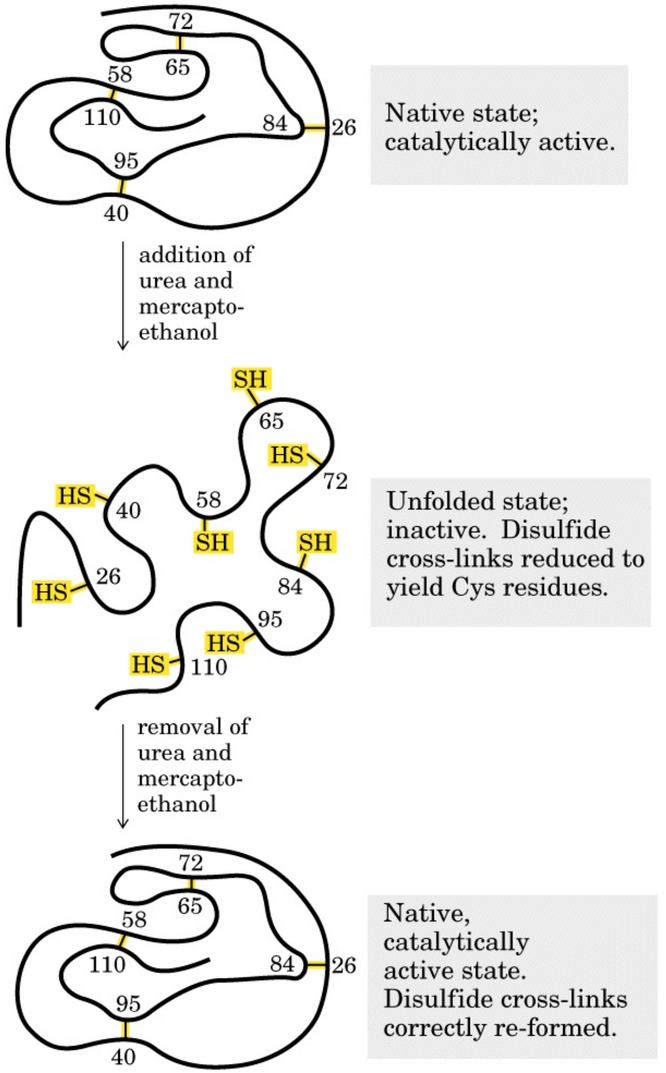
1AA is enough to determine the fold! And the function?

GA98 exhibits diminished affinity for HSA but has acquired affinity for IgG. GB98 binds tightly to IgG but not HSA.

1.Alexander, P. A., He, Y., Chen, Y., Orban, J. & Bryan, P. N. A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 21149 (2009).

Anfinsen: The native structure of a protein is dictated by its sequence

- Small protein can fold independently (Anfinsen's experiment with ribonuclease A)
- Large proteins fold with the assistance of molecular chaperons

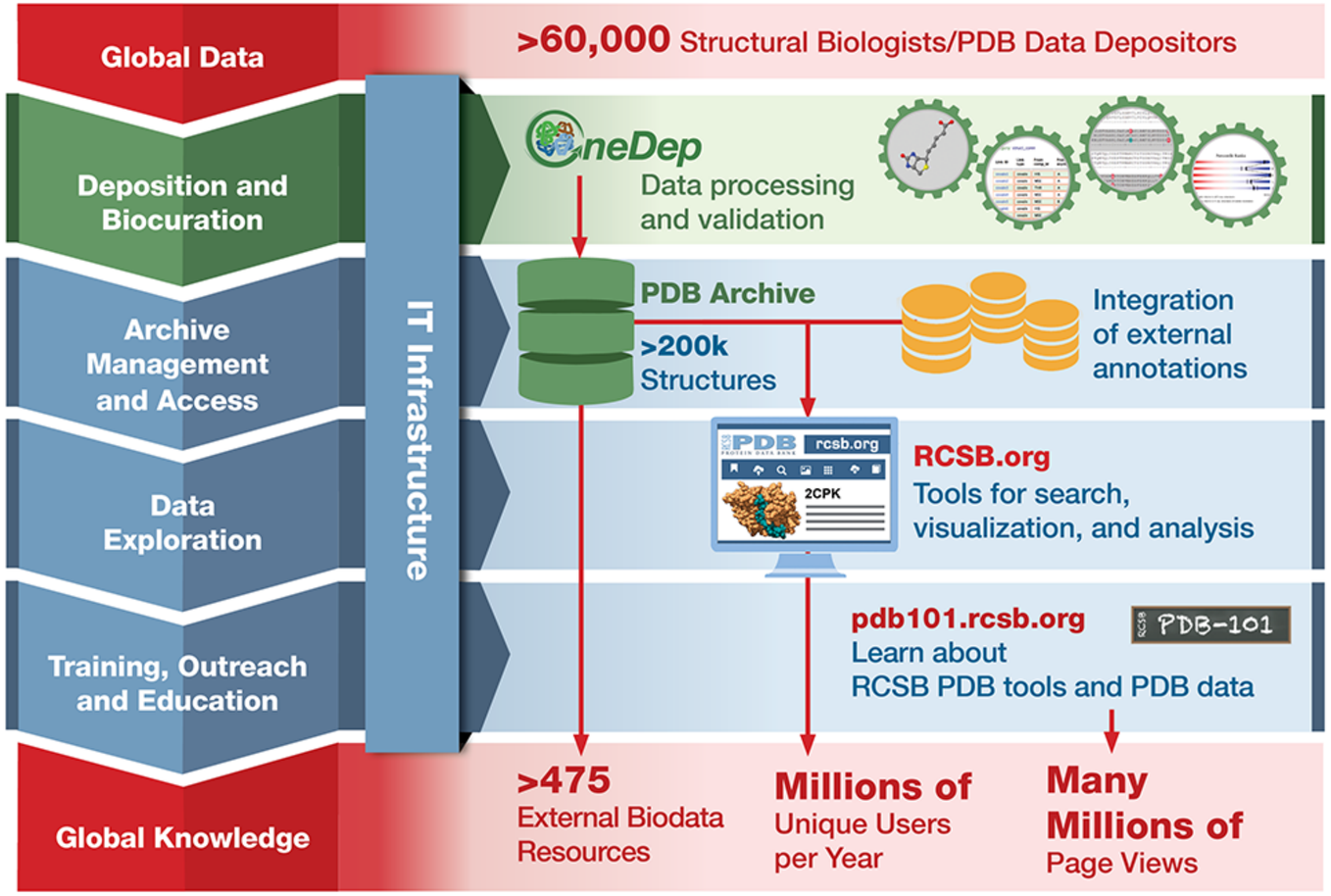


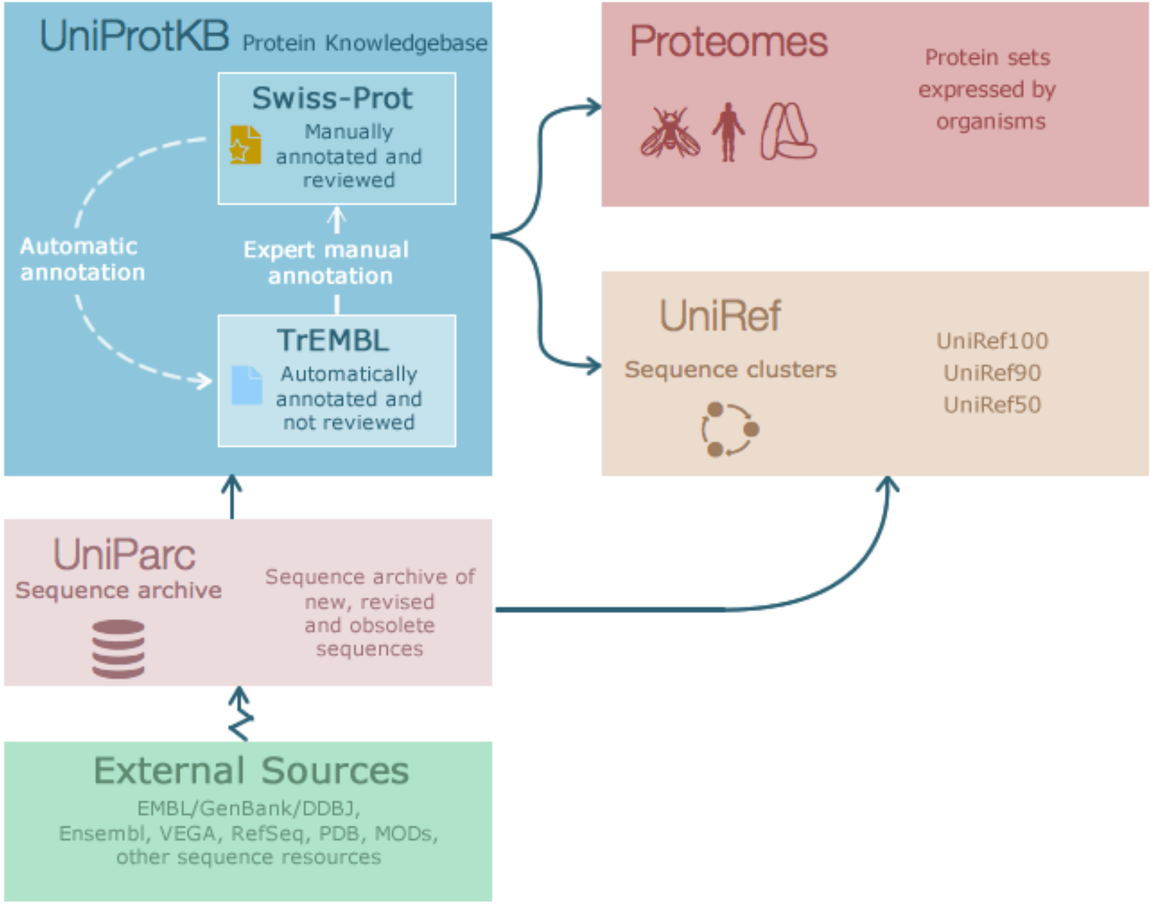
If proteins can refold spontanously in vitro, this suggest that all the information needed to fold is stored in their sequence.

Principles That Govern Protein Folding
Anfinsen C.B.
Science (1974) 181, 223-230.

Introduction to Proteins, Kessel and Ben-Tal, 2018

The data: public, curated databases

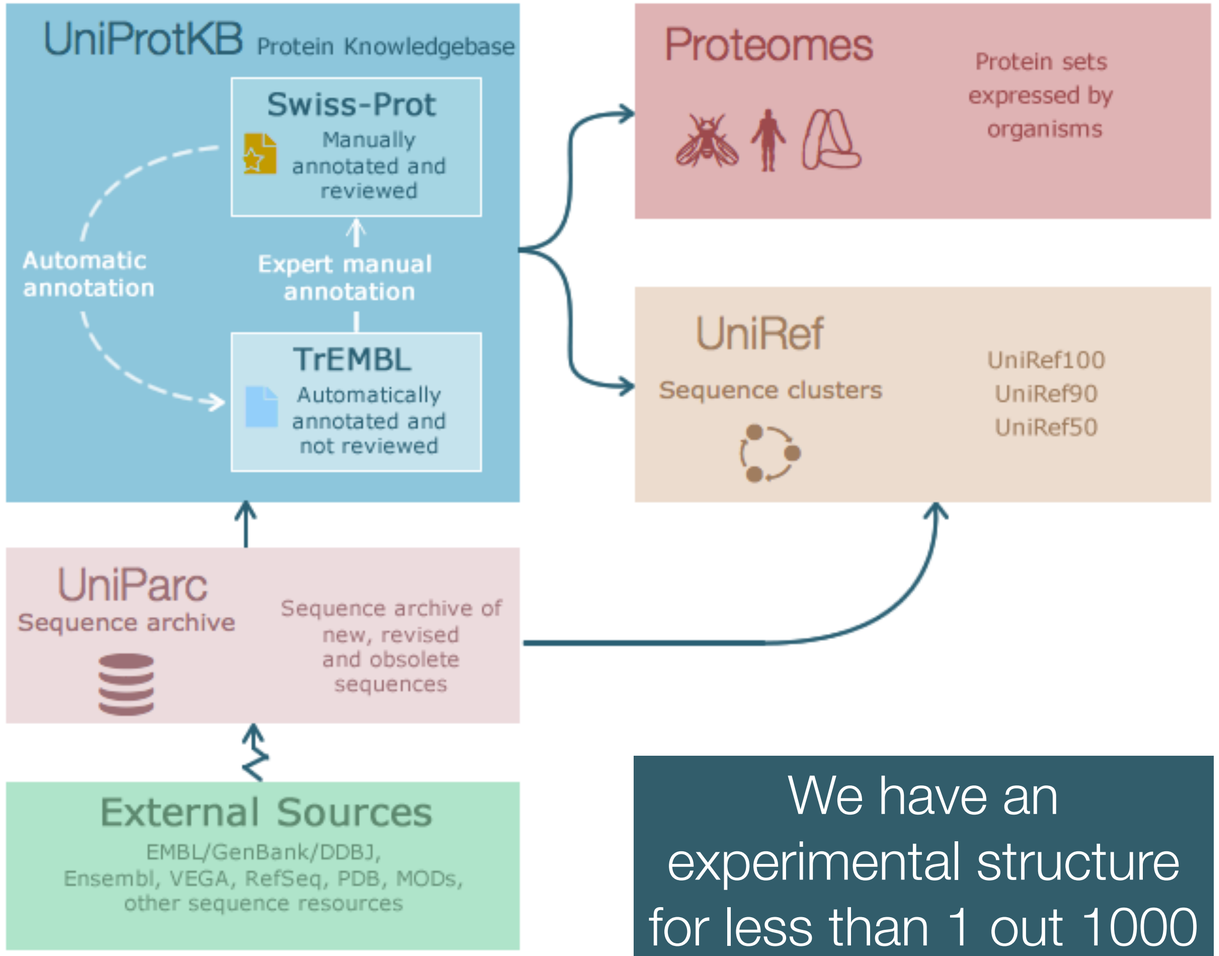




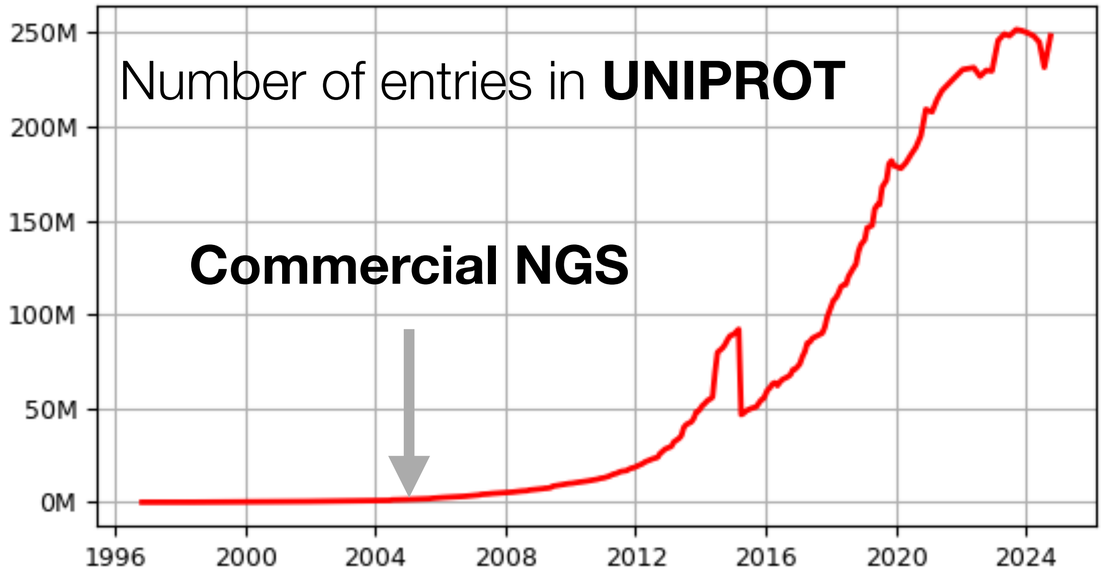
www.rcsb.org

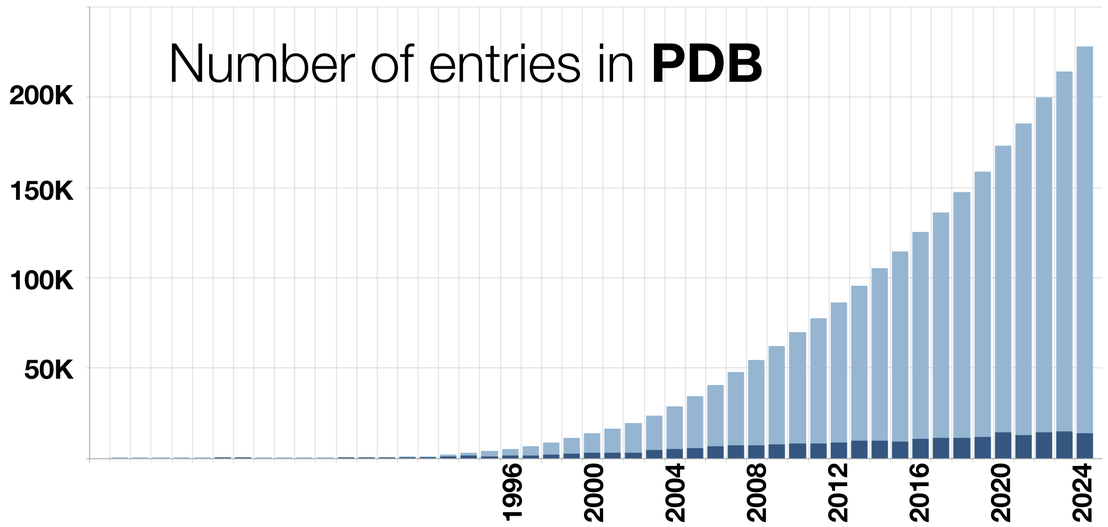
www.uniprot.org

UNIPROT, Next Generation Sequencing and the need for protein structure prediction



We have an experimental structure for less than 1 out 1000 sequences

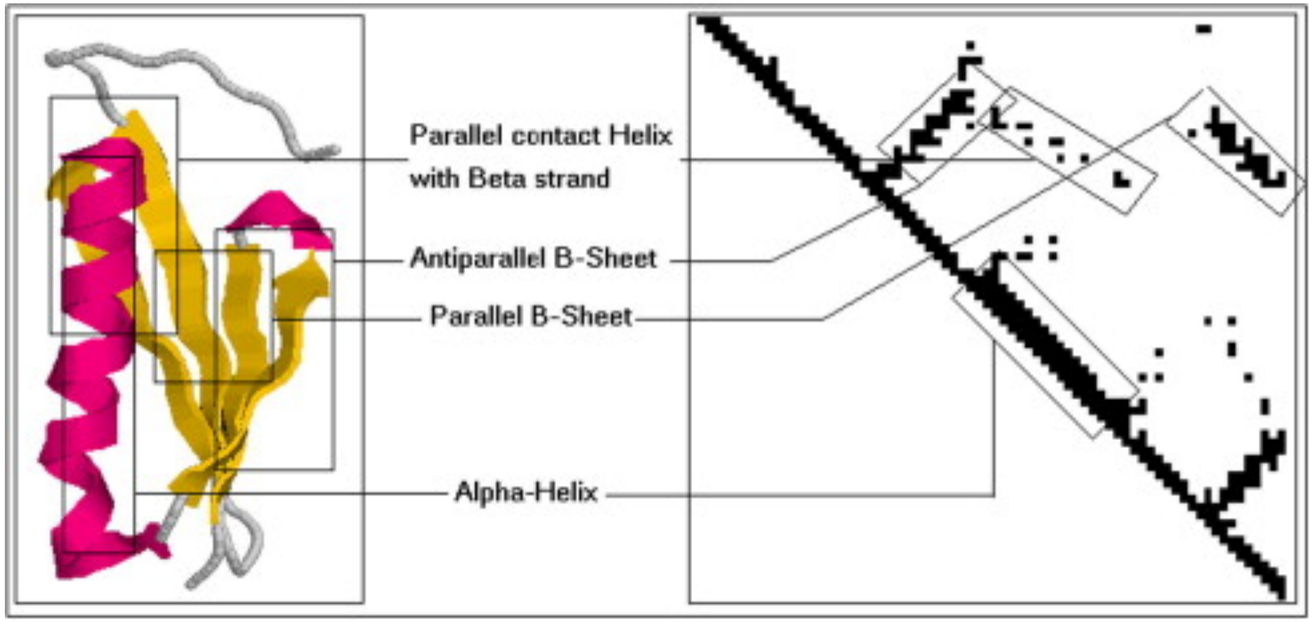




UNIVERSITÀ DEGLI STUDI DI MILANO

DBS

Data representation: sequence, secondary structure, contact maps, 3D structures




```
ATOM      1  N   MET A   1      12.969  18.506  30.954
ATOM      2  CA  MET A   1      13.935  18.529  29.843
ATOM      3  C   MET A   1      13.138  18.692  28.517
ATOM      4  O   MET A   1      12.007  18.222  28.397
ATOM      5  CB  MET A   1      14.733  17.216  29.882
ATOM      6  CG  MET A   1      15.742  16.983  28.738
ATOM      7  SD  MET A   1      17.378  17.025  29.359
ATOM      8  CE  MET A   1      17.166  16.055  30.819
ATOM      9  N   THR A   2      13.719  19.413  27.573
ATOM     10  CA  THR A   2      13.088  19.661  26.283
ATOM     11  C   THR A   2      13.561  18.631  25.300
ATOM     12  O   THR A   2      14.763  18.432  25.121
ATOM     13  CB  THR A   2      13.527  20.980  25.667
ATOM     14  OG1 THR A   2      13.307  22.020  26.627
ATOM     15  CG2 THR A   2      12.704  21.284  24.409
```

Contact or distance maps are matrices reporting either a 0/1 or a distance between two amino acids (usually considering the Cb carbon)

Secondary structure prediction is relatively easy but secondary structure can depend on tertiary and quaternary structure

Useful as:

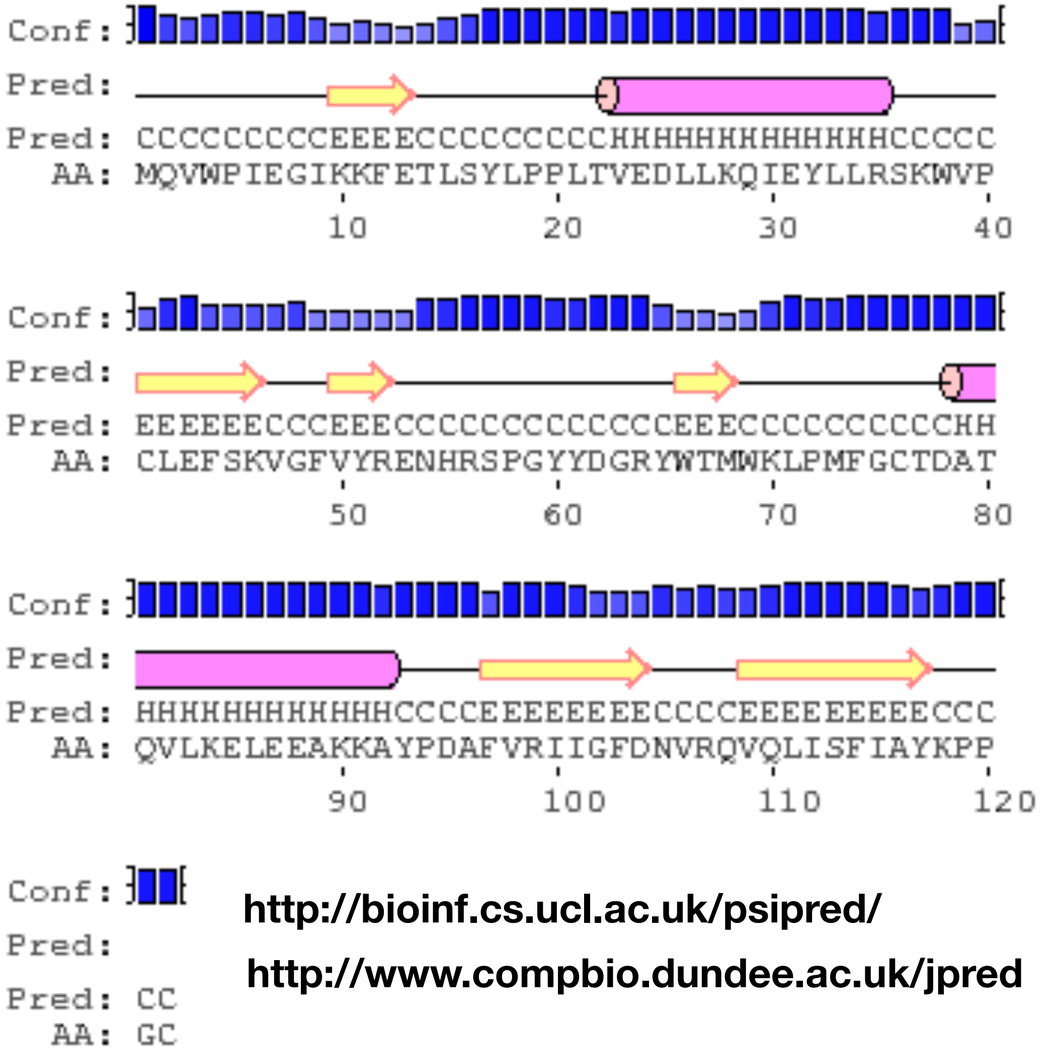
- **a step towards tertiary structure prediction**
- **Sequence alignment**
- **Structure determination at intermediate resolution**
- **Prediction of other properties (aggregation, trans-membrane, disorder)**

First generation of predictors (~1970) were based on single residues frequencies, second generation (~1990) took into account patterns for segments. Both biased by the small size of the PDB and the lack of systematic assignment.

Accuracy ~60%

Sequence alignment is performed (psi-blast) and neural networks are trained to assign secondary structures (from one letter sequence to one letter secondary structure)

Accuracy ~81%



http://bioinf.cs.ucl.ac.uk/psipred/

http://www.compbio.dundee.ac.uk/jpred

Phys-chem based methods like MD simulations can predict protein structures, but are limited by size and folding time scale.

Classical molecular dynamics simulations of a protein in explicit environment with an energy function that tries to approximate the interactions (Force Field):

first approx for vibrations (anharmonic potential can be used for more accurate vibrations)

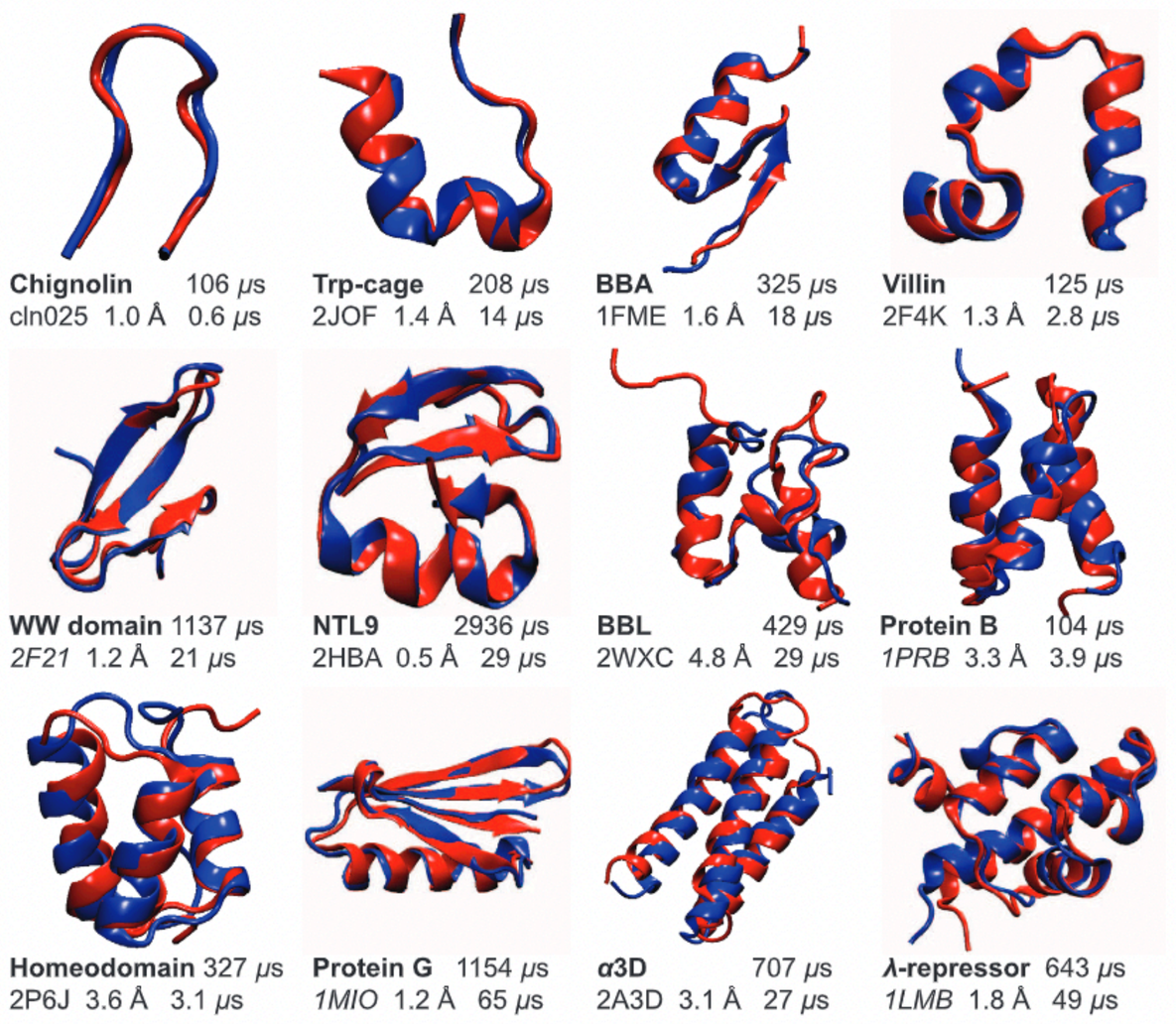
geometrical consideration pi-bonds, etc

$$V(r) = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{torsions} k_\phi[\cos(n\phi + \delta) + 1] + \sum_{\substack{nonbond \\ pairs}} \left[\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}} \right]$$

Dispersions (interactions of neutral molecules)

point charge Coulomb

excluded volume

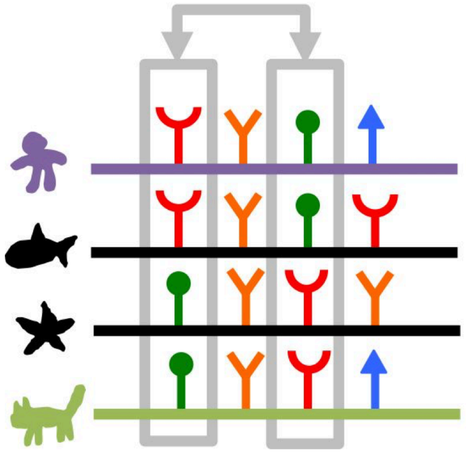


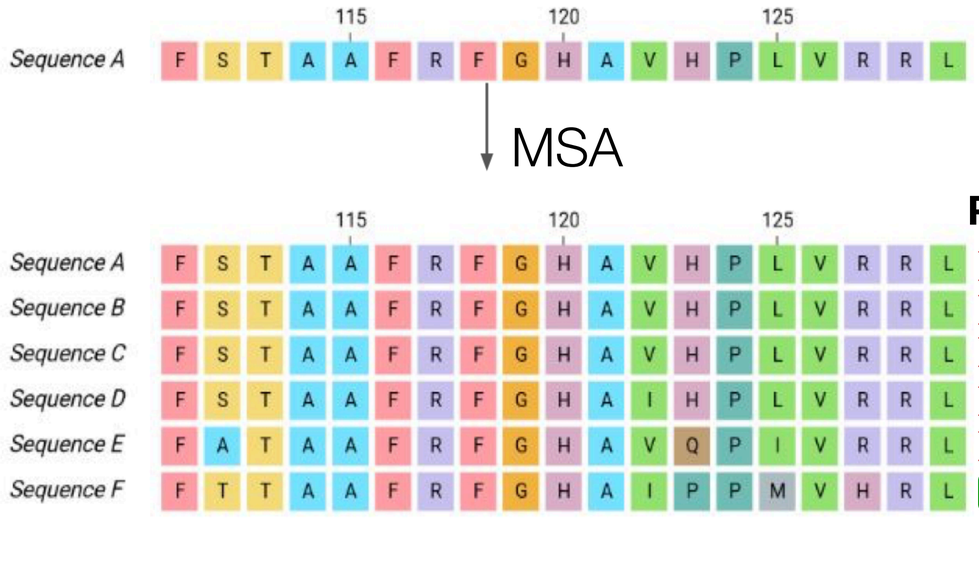
How fast-folding proteins fold. *Sci New York N Y* **334**, 517–20 (2011).

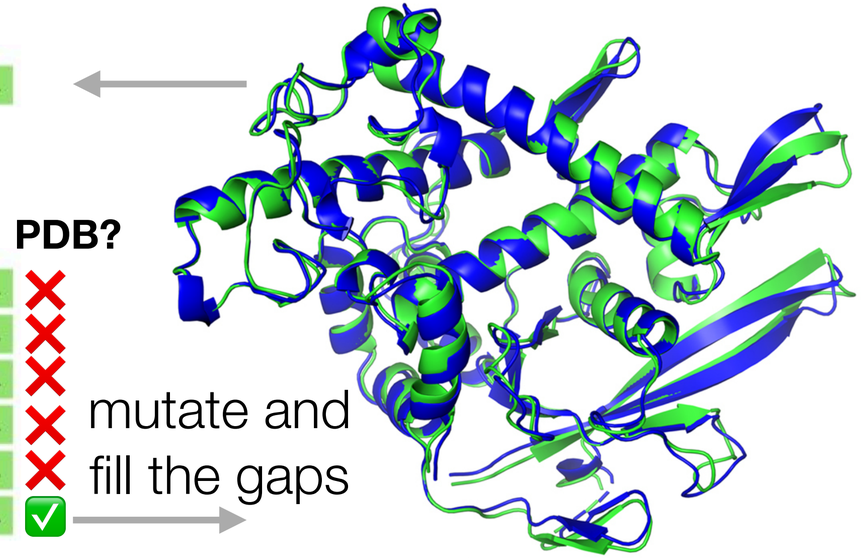
UNIVERSITÀ DEGLI STUDI DI MILANO

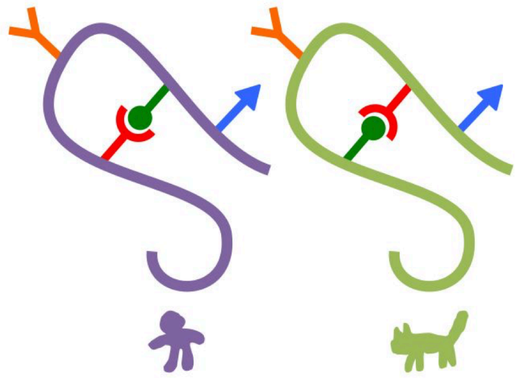
DBS

Structure prediction: homology modelling and coevolution analysis







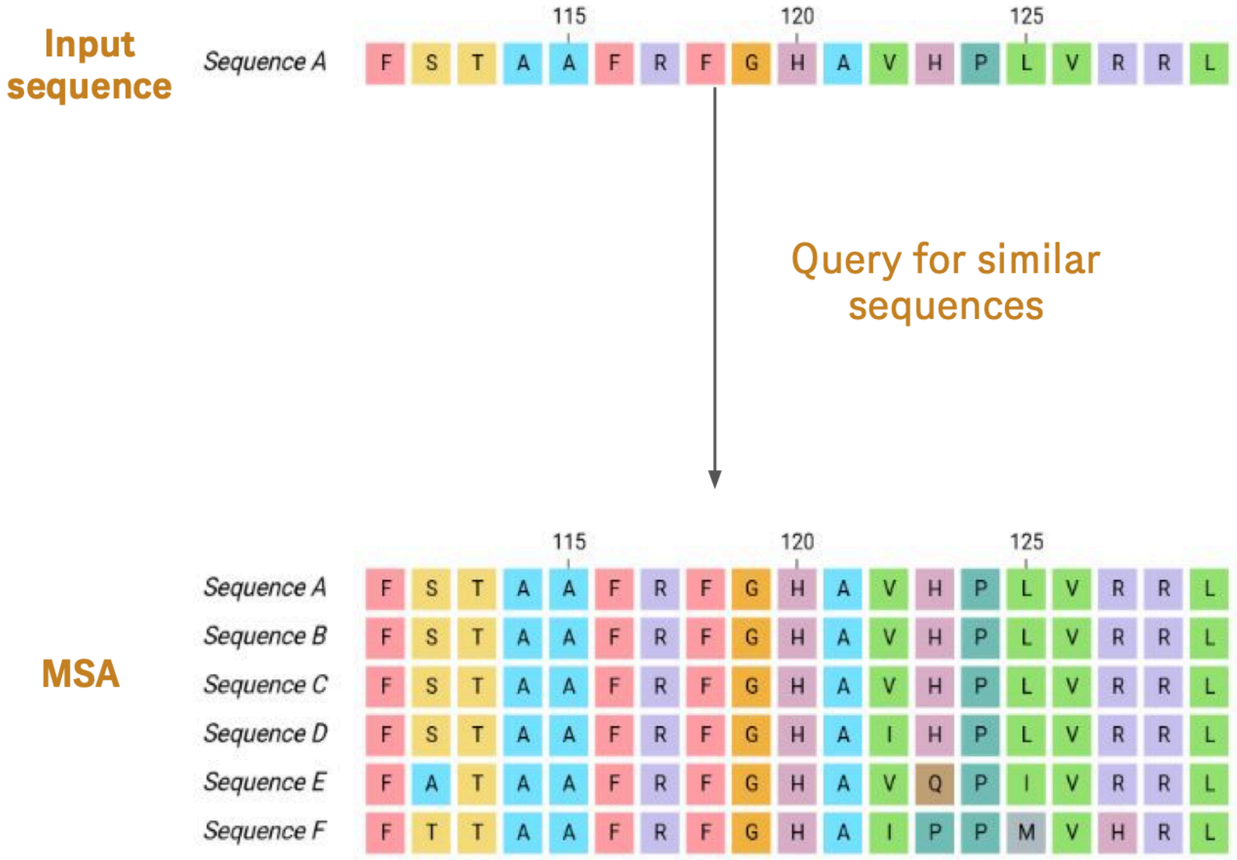


If two positions are in contact, then if one is mutated, the other also loses function.

Thus, statistically observed pairs of residues that "change together" across evolution often denotes contact.

For few cases with 10K or more sequences in the MSA

Search for patterns in protein sequences:



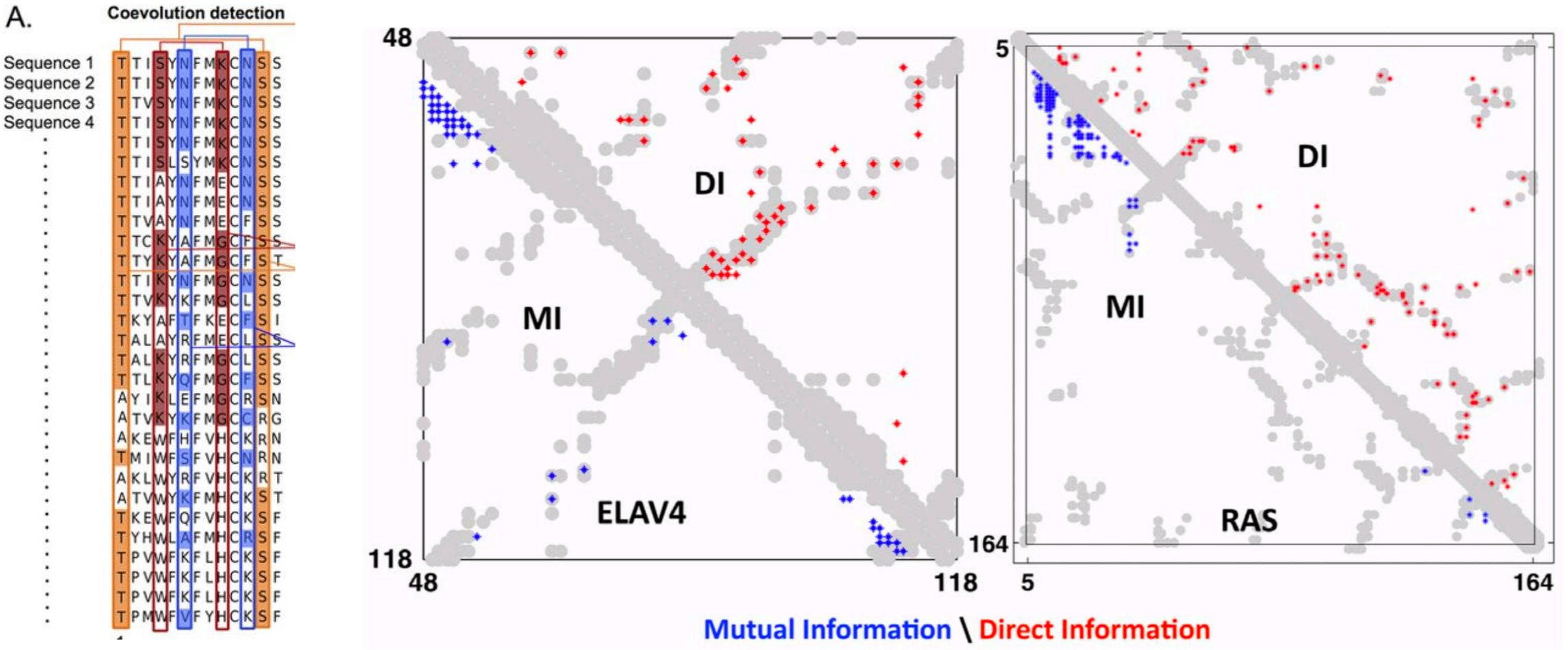
From an MSA, assuming that similar sequences are evolutionary related, we can learn a lot:

1. Sequence conservation: The frequency of each aminoacid in each position: P(F1), P(S2), …

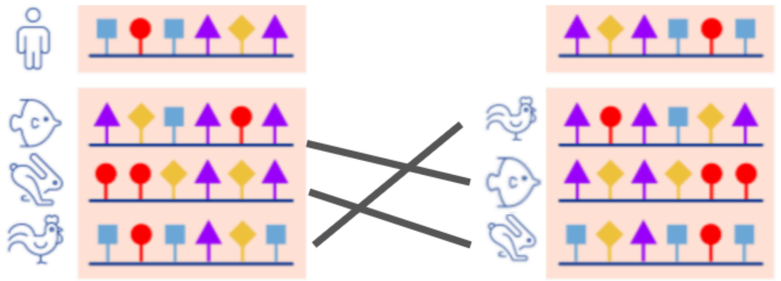
2. The frequency of pair of AA: P(F1, S2), P(F1, T3), P(F1, A4), …

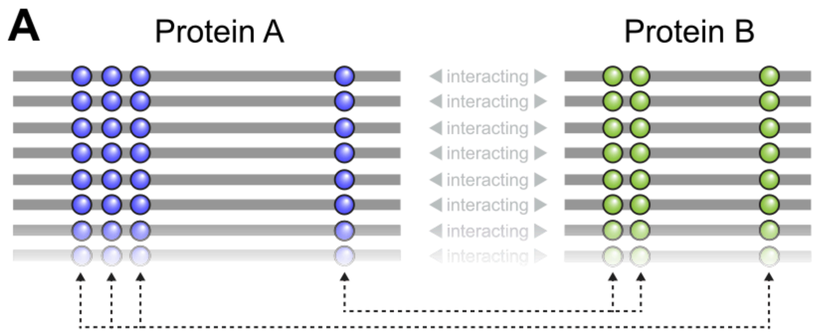
In principle we could get probability of triplets, etc, but already for pairs we need huge MSA.

A properly performed Coevolution analysis from LARGE MSA can provide reliable structural data

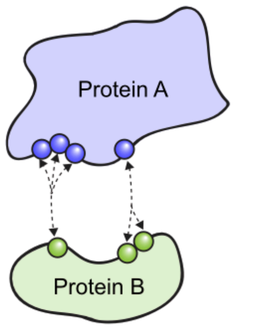


Coevolution also works for protein complexes

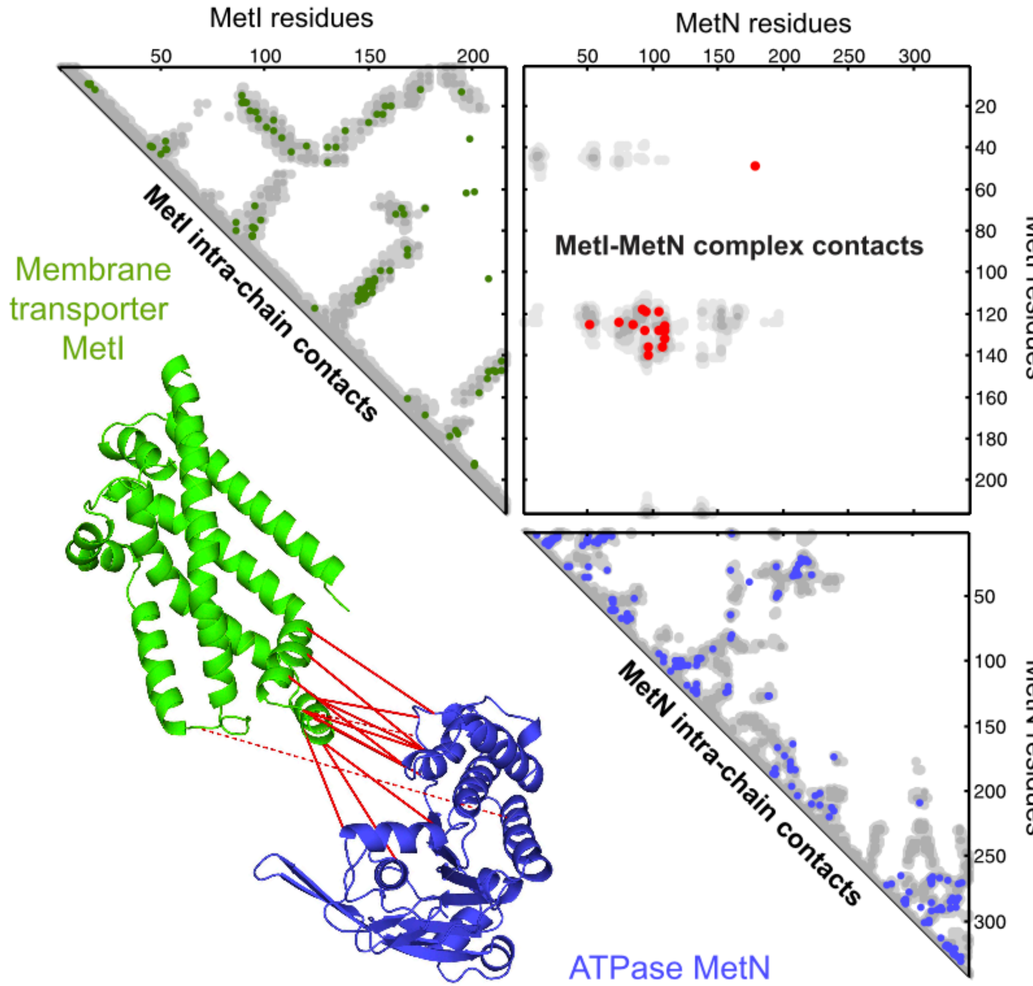




Statistical co-evolution analysis



Inter-protein ECs (evolutionary couplings)



Summarising:

Chemistry does not enforce sequence-structure similarity in proteins.

"Our" proteins are the result of natural evolution. Sequences have generally evolved under the constraint of preserving/repurposing/improving function. Function can be associated with structure. Thus, in our experience, there is a sequence-structure similarity relationship.

MSA is a way to analyze sequences and identify homology. From MSA it is possible to find patterns related to the structure/function of a system.

Keep in mind that 30% of proteins/protein regions are unstructured, so the above assumption does not hold because function is not associated with structure. Evolution worked differently there.

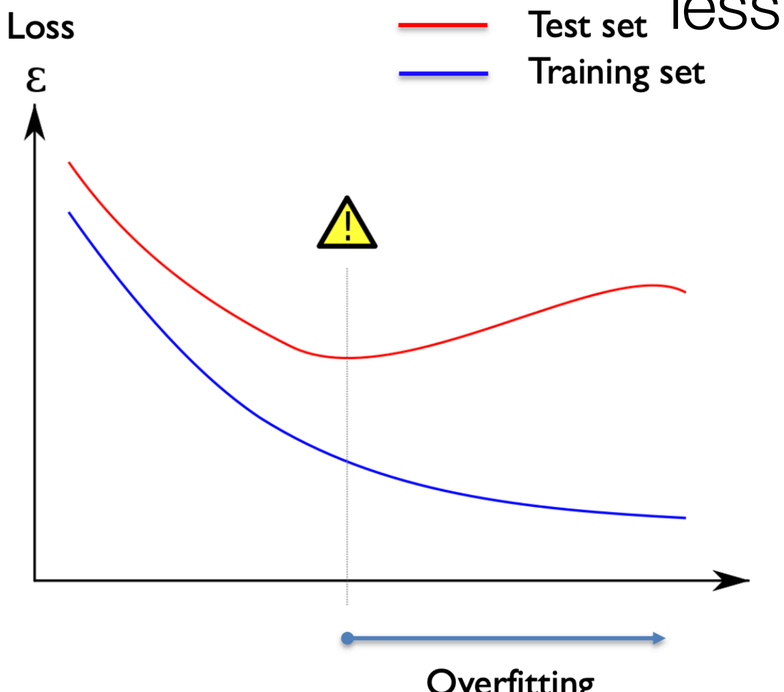
From COEVOLUTION to MACHINE LEARNING

Instead than solving the coevolution analysis case by case, that is very expensive and requires large MSA, would it be possible to use Machine Learning on many MSA and structures to train a single, reusable, network that is able to extract the structural information present in a MSA also when limited number of sequences are available?

Not all ML are the same

The architecture is the creative, trial-and-error, part. Not all architecture are equally good at learning from a limited amount of data. Architecture design imply making strong hypothesis on how we want to interpret the data.

The weights are obtained for a given architecture and are the result of a more or less costly computer optimisation.



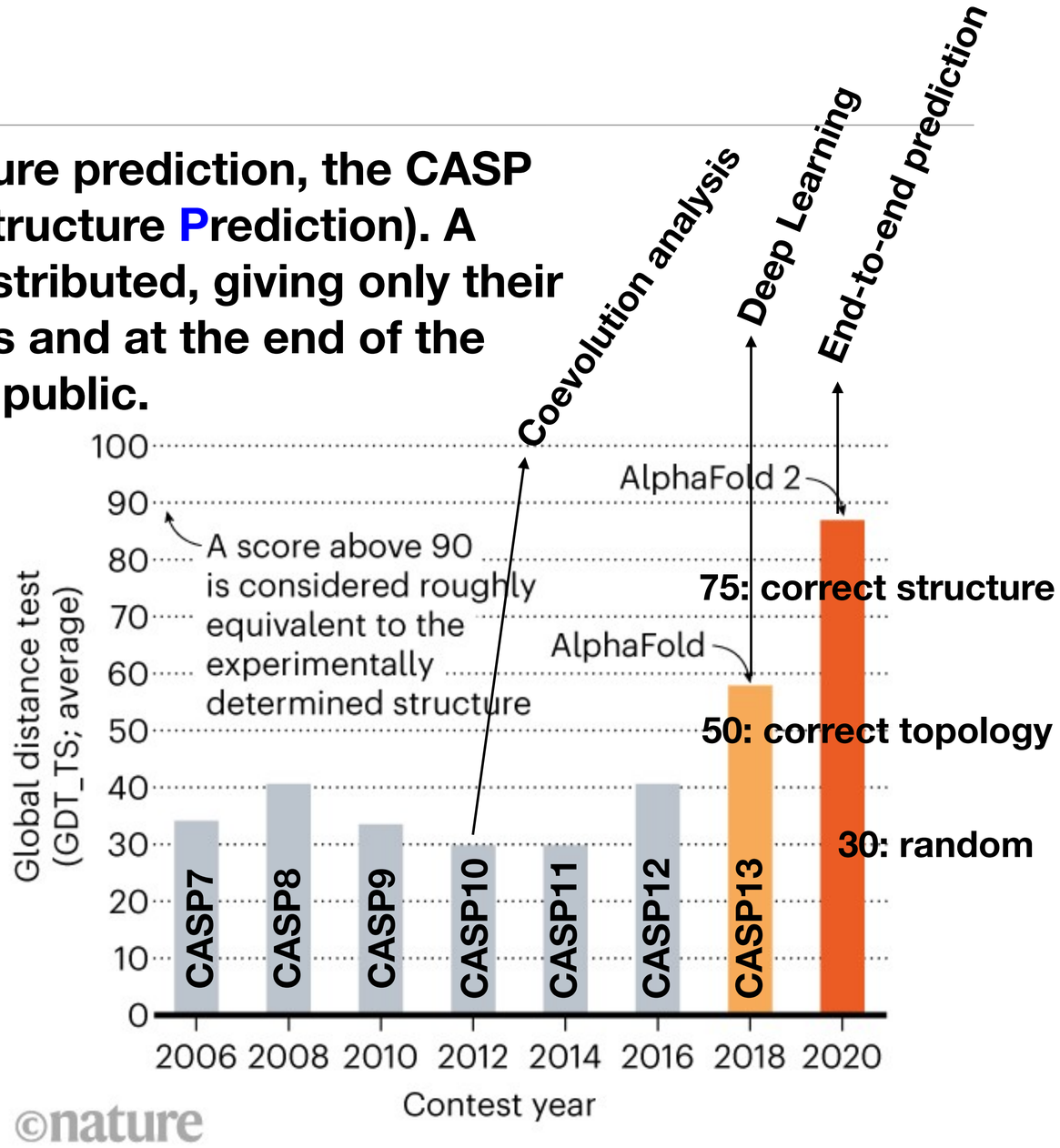
To train an architecture, that means to find the weights, you need a SCORING function that is a measure in some way of the distance of the output from the Training set.
It could be something as simple as sum[(output-training)^2]

Ab-initio protein structure predictions

From 1994 there is a competition for protein structure prediction, the CASP (Critical Assessment of techniques for protein Structure Prediction). A number of structures recently determined are not distributed, giving only their sequences. Participants can deposit their models and at the end of the competition the results are made public.

From the early 2000 the way to predict the structure of a protein has been to follow a "fragment replacement" strategy:

1. **Generate many configuration using local information like secondary structure prediction as well as contact restraint using physico-chemical principles.**
2. **Refine the final structure at atomic resolution and use some fancy "scoring function"**

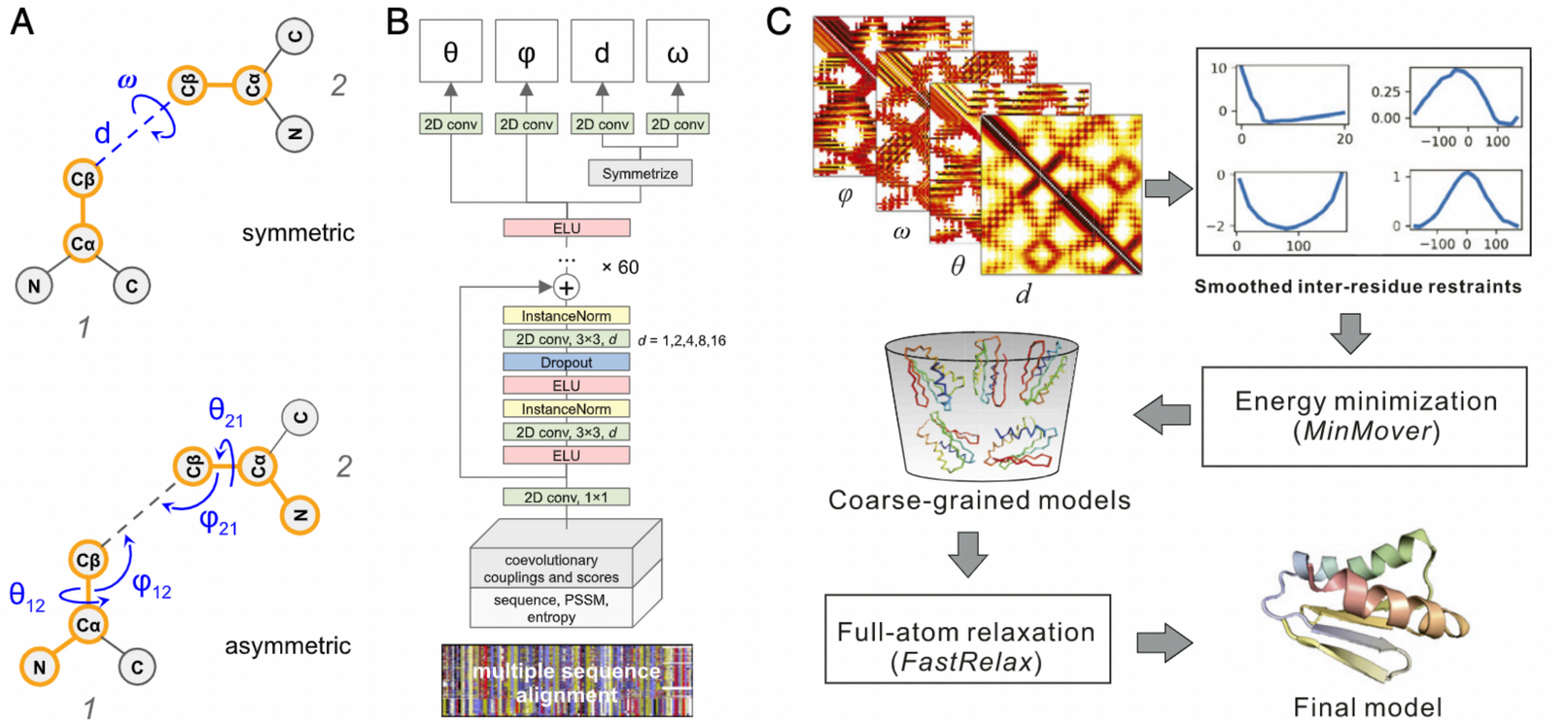


UNIVERSITÀ DEGLI STUDI DI MILANO

DBS

https://predictioncenter.org

CASP13 - AlphaFold, RaptorX and TrRosetta: deep-learning distances and orientations from coevolution features



~15,000 proteins structures PDB
Histograms for the distances among all couples of amino acids (distance distributions):

ALA-ALA: distances from all couples in all selected PDBs
ALA-CYS:…
ALA-ASP:…

All sequences are also associated to their relative MSA derived covariance matrix

A deep neural network is trained to go from the covariance matrix to the distance distribution

NOTE: we move from contact, binary, maps to distance distribution maps, this allow doing direct minimisation of the structure using the propagation property of the network

1.Yang, J. *et al*. Improved protein structure prediction using predicted interresidue orientations. *Proc National Acad Sci* **117**, 1496–1503 (2020).

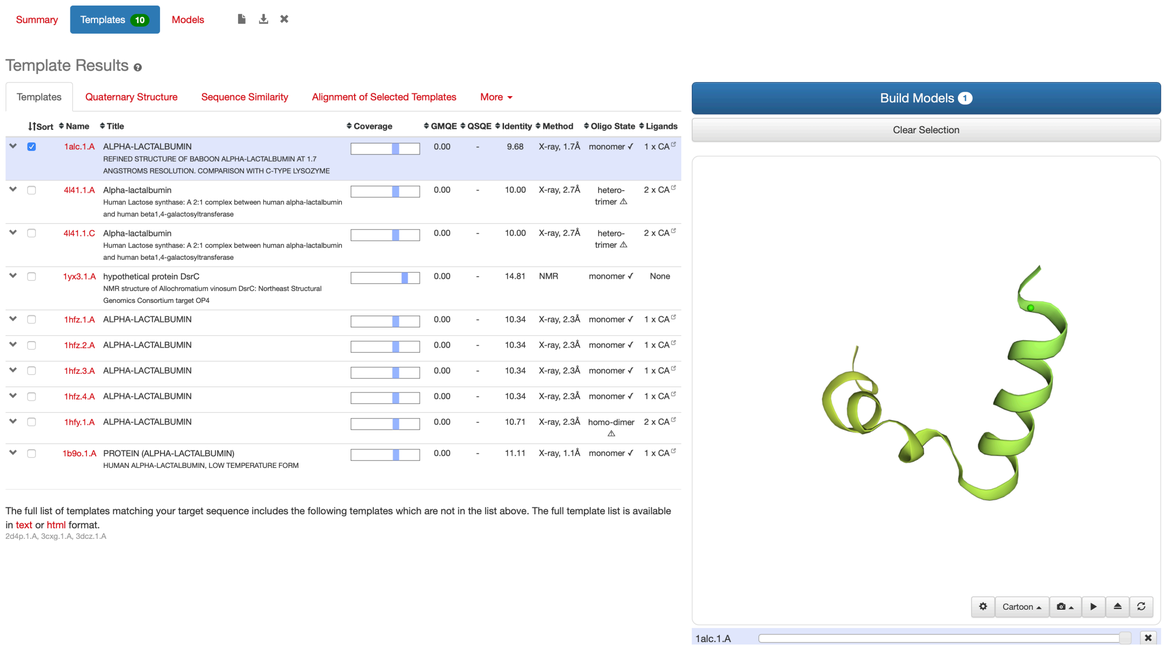
https://github.com/gjoni/trRosetta

MIZ1: an example from my own experience

MIZ1 is a protein that is studied here in the department for its functional role in plants were it plays a key role in root development. Its structure is unknown and we wanted to determine it, or at least to help in design a construct for further experiments, because the full length protein gets into the inclusion bodies in E. Coli.

```
MVPYQELTLQRSFSYNSRKINPVTSPARSSHVRSPSSSALIPSIPEHELFLVPCRRCSYV
PLSSSSSASHNIGKFHLKFSLLLRSFINIINIPACKMLSLPSPPSSSSSVSNQLISLVTG
GSSSLGRRVTGTLYGHKRGHVTFSVQYNQRSDPVLLLDLAMSTATLVKEMSSGLVRIALE
CEKRHRSGTKLFQEPKWTMYCNGRKCGYAVSRGGACTDTDWRVLNTVSRVTVGAGVIPTP
KTIDDVSGVGSGTELGELLYMRGKFERVVGSRDSEAFYMMNPDKNGGPELSIFLLRI
```

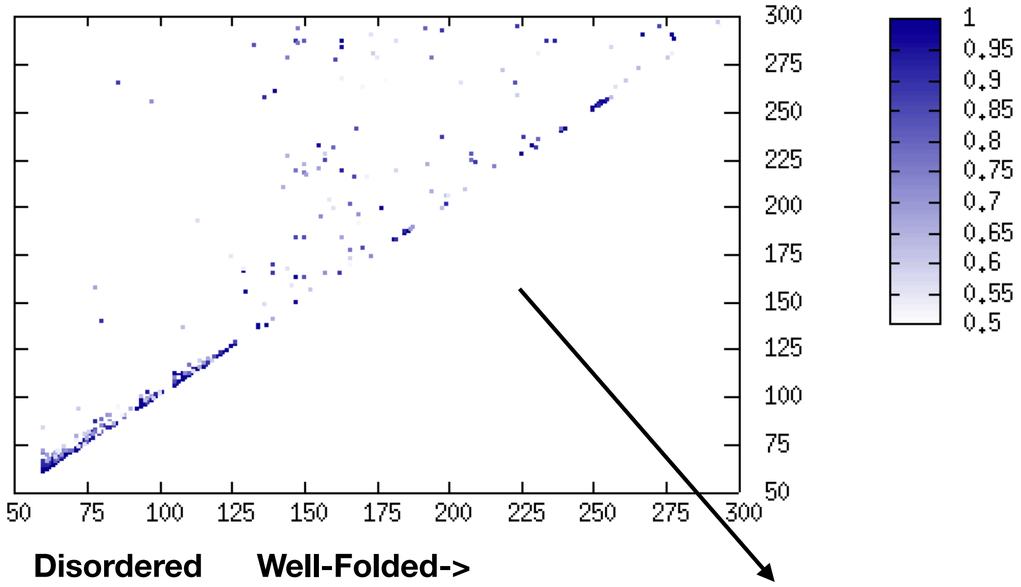
First test - homology modelling - but there are no homologues with known structure:



NOPE!

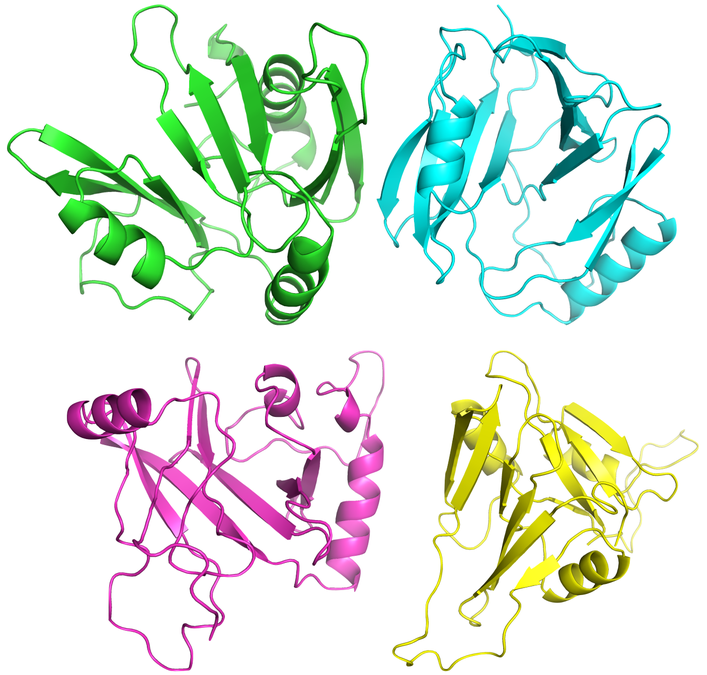
MIZ1: an example from my own experience

MSA and coevolution analysis, we found ~700 homologues sequences, that is a small number for pre-AF methods.



Yet, we used the map to design a construct to produce the protein that actually worked well

Robetta + ev-couplings - ‘old’ ab-initio structure prediction:



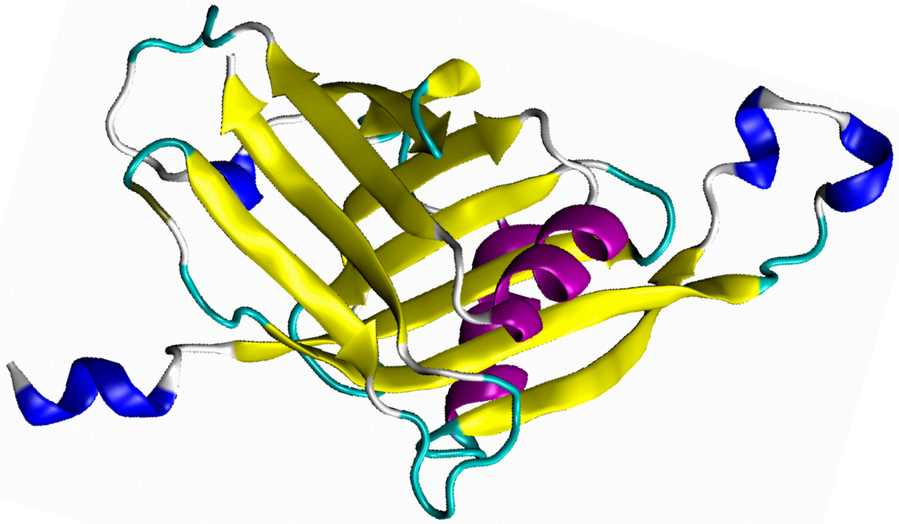
Four structures with nothing in common… nope.

MIZ1: an example from my own experience

RaptorX
(first time of ev-couplings+deep learning):

Five structures with something in common, but when looked in the details they are unphysical

trRosetta
(post AF ev-couplings+deep learning):

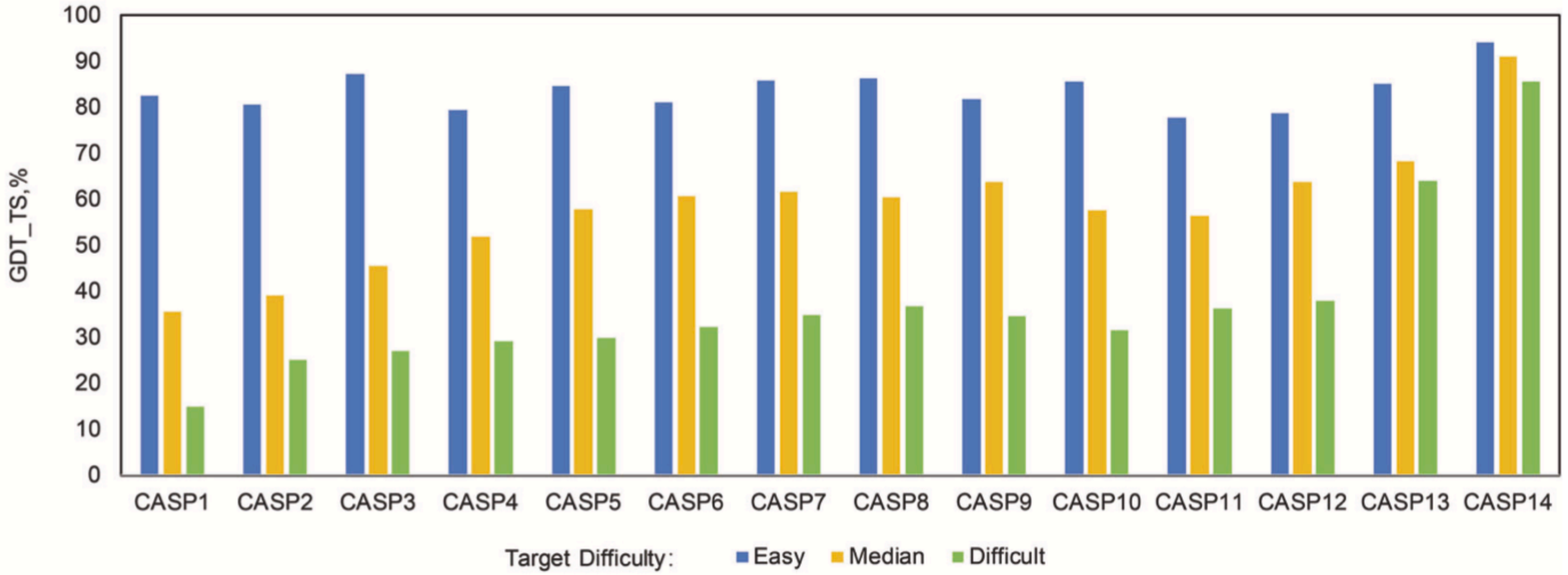


Well converged, well-defined unique structural model with perfectly reasonable features.

NOTE 1: this is are case with ~700 hundred homologue sequences, this is still a relatively large number

NOTE 2: I am talking about trRosetta and not AF, because the first AlphaFold as never been made publicly available

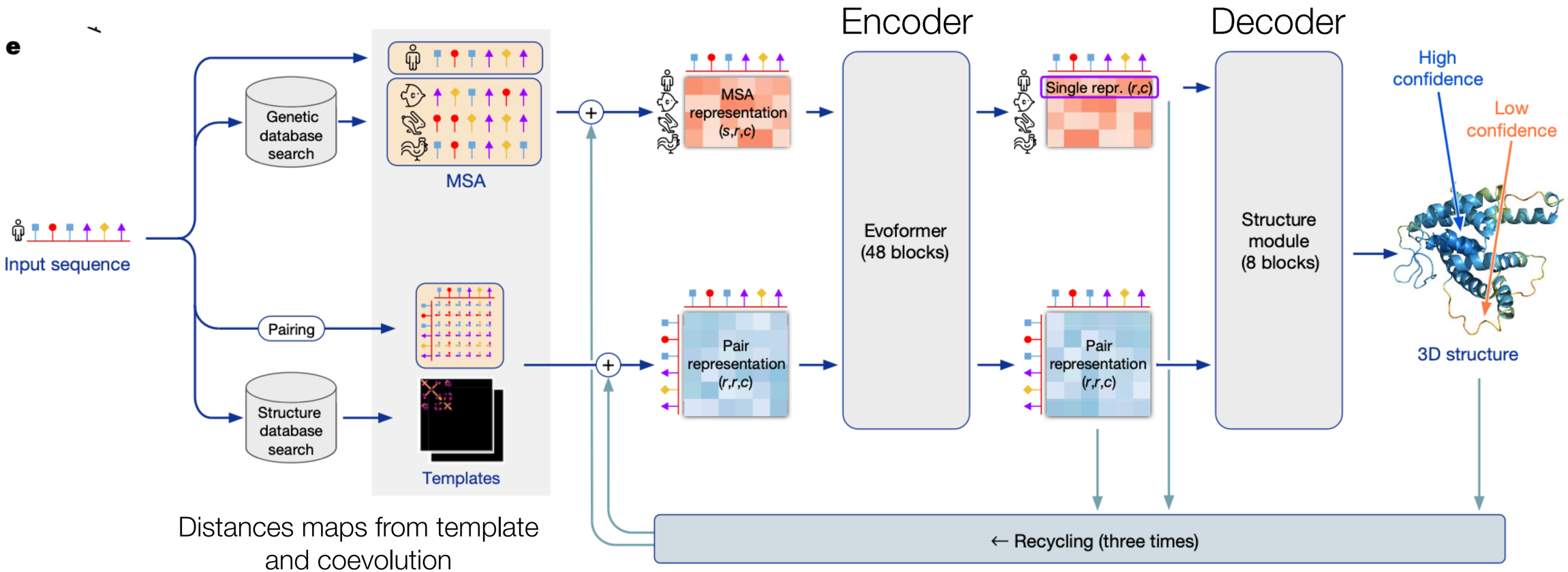
CASP14 - AlphaFold2: not just a technical improvement but a completely ad hoc newly designed network architecture



How to improve the quality of the prediction when <u>less than one hundred homologue sequences</u> are available?

AlphaFold2: some ideas on the architecture

In pre AF2 approaches the network allowed to increase the amount of information extracted from the MSA. **AF2 first addition is to improve the quality of the MSA itself. The second addition is to have an end-to-end network** instead than using the network to generate the distance distribution map and then applying it to a conventional fragment replacement method.

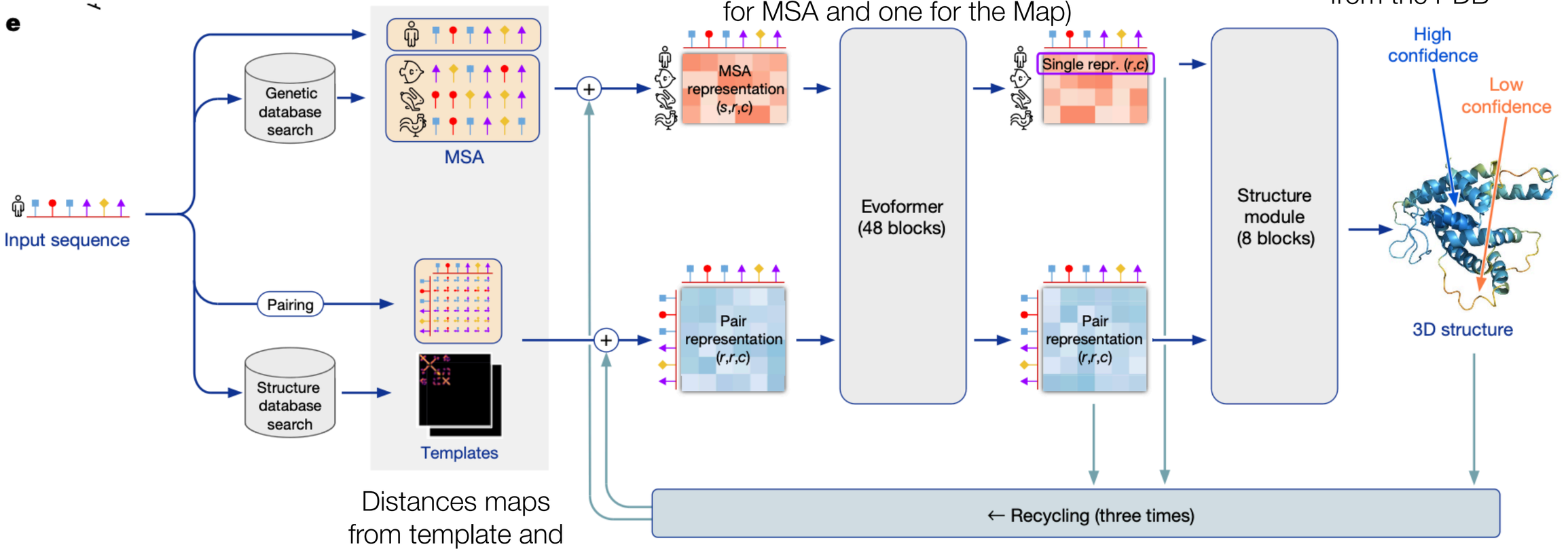


Distances maps from template and coevolution

AlphaFold2: some ideas on the architecture

Evoformer: improve the MSA (better alignment and remove useless/wrong sequences) and the distances map (Two Transformers (Attention) one for MSA and one for the Map)

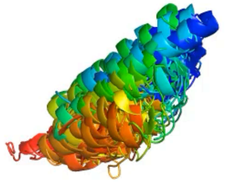
Get a structure using physical and geometrical constraints learned from the PDB



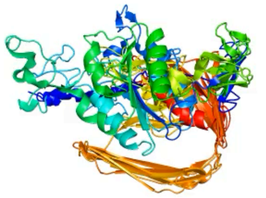
Distances maps from template and coevolution

Use the resulting structure to improve again MSA and distances map

Examples:



Recycling iteration 0, block 01
Secondary structure assigned from the final prediction



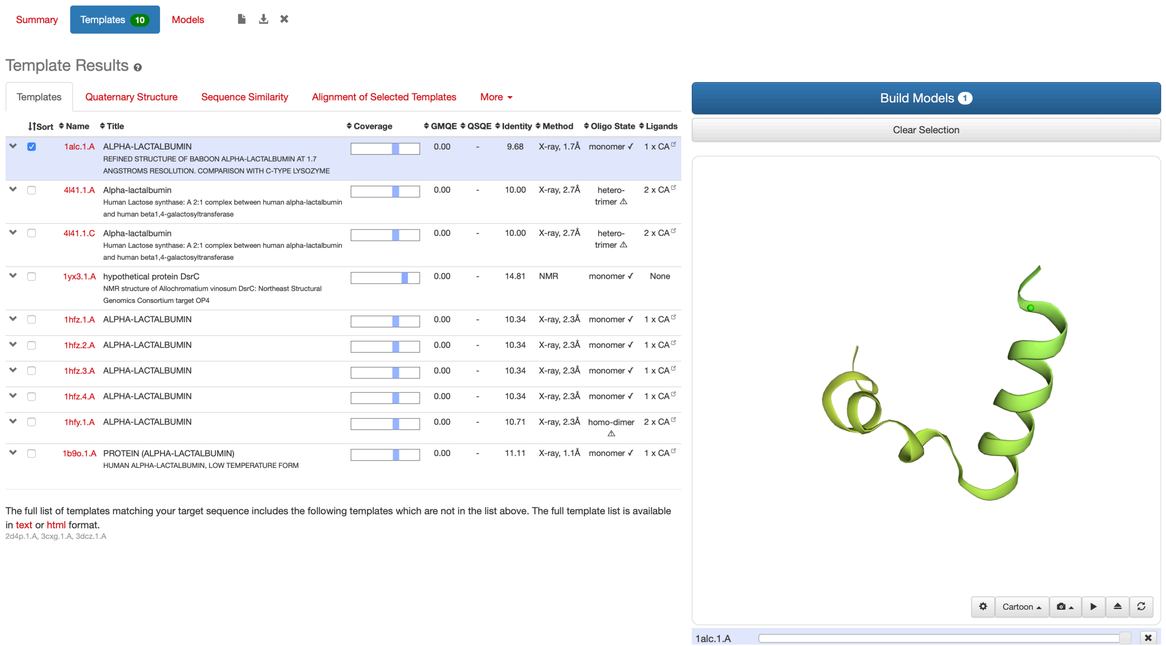
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

MIZ1: an example from my own experience

MIZ1 is a protein that is studied here in the department for its functional role in plants were it plays a key role in root development. Its structure is unknown and we wanted to determine it, or at least to help in design a construct for further experiments, because the full length protein gets into the inclusion bodies in E. Coli.

```
MVPYQELTLQRSFSYNSRKINPVTSPARSSHVRSPSSSALIPSIPEHELFLVPCRRCSYV
PLSSSSSASHNIGKFHLKFSLLLRSFINIINIPACKMLSLPSPPSSSSSVSNQLISLVTG
GSSSLGRRVTGTLYGHKRGHVTFSVQYNQRSDPVLLLDLAMSTATLVKEMSSGLVRIALE
CEKRHRSGTKLFQEPKWTMYCNGRKCGYAVSRGGACTDTDWRVLNTVSRVTVGAGVIPTP
KTIDDVSGVGSGTELGELLYMRGKFERVVGSRDSEAFYMMNPDKNGGPELSIFLLRI
```

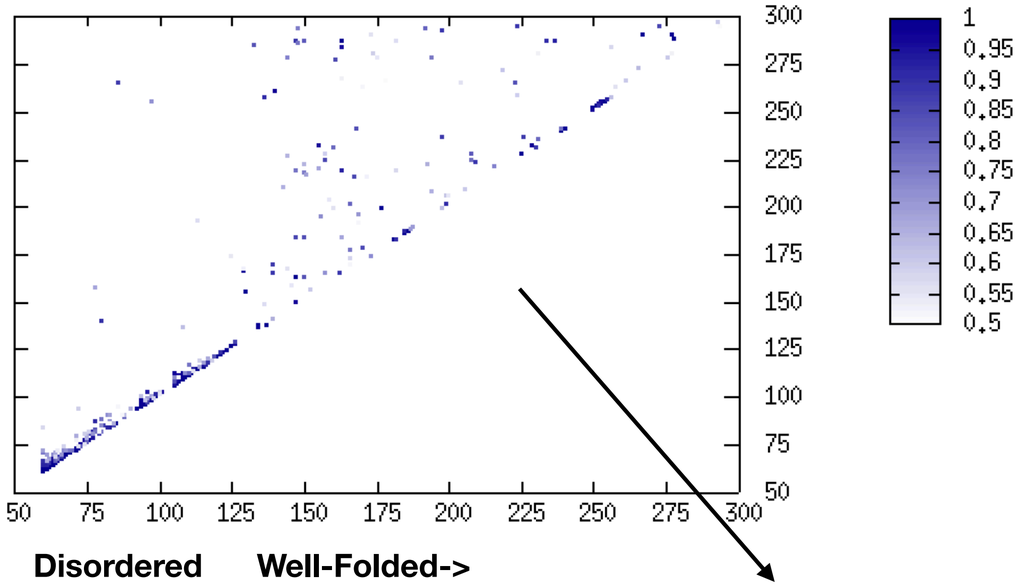
First test - homology modelling - but there are no homologues with known structure:



NOPE!

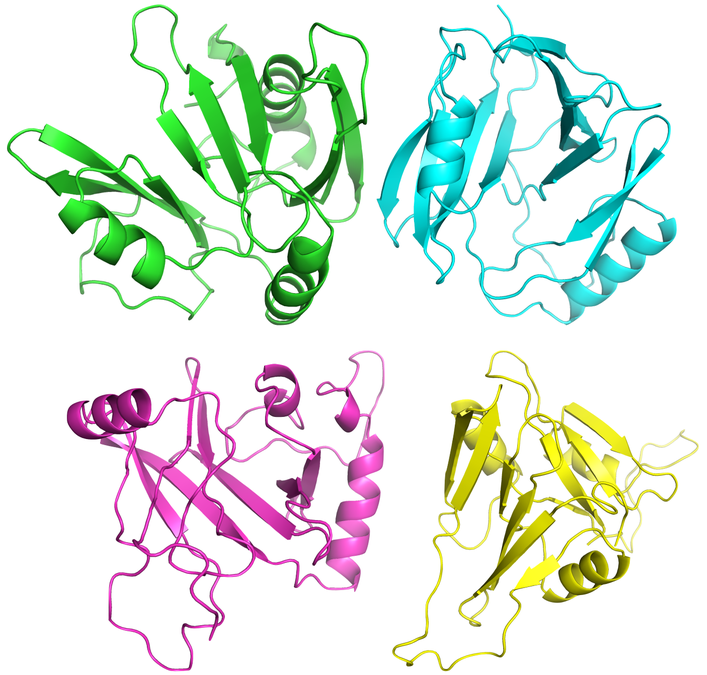
MIZ1: an example from my own experience

MSA and coevolution analysis, we found ~700 homologues sequences, that is a small number for pre-AF methods.



Yet, we used the map to design a construct to produce the protein that actually worked well

Robetta + ev-couplings - ‘old’ ab-initio structure prediction:



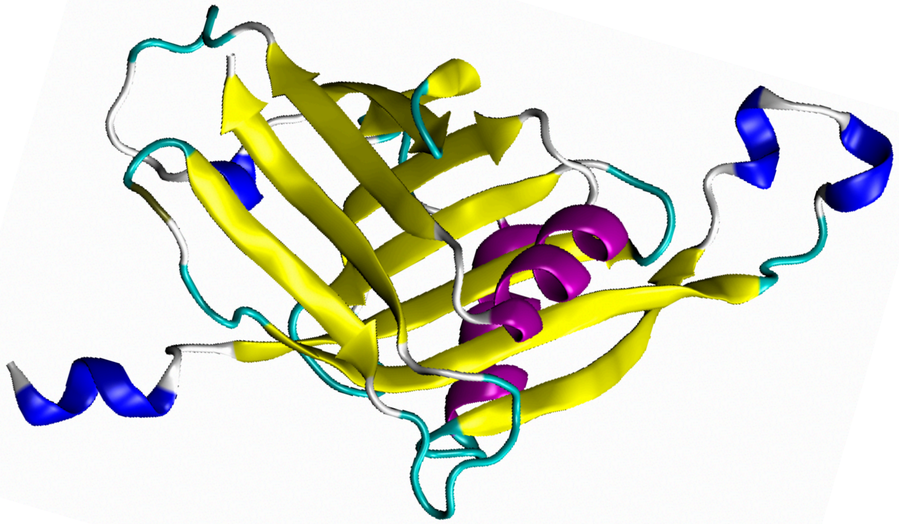
Four structures with nothing in common… nope.

MIZ1: an example from my own experience

RaptorX
(first time of ev-couplings+deep learning):

Five structures with something in common, but when looked in the details they are unphysical

trRosetta
(post AF ev-couplings+deep learning):



Well converged, well-defined unique structural model with perfectly reasonable features.

NOTE 1: this is are case with ~700 hundred homologue sequences, this is still a relatively large number

NOTE 2: I am talking about trRosetta and not AF, because the first AlphaFold as never been made publicly available

AlphaFold Protein Structure Database

Developed by Google DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism or sequence search BETA

Search

Examples: MENFQKVEKIGEGTYGV... Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli

See search help Go to online course

See our updates – September 2024

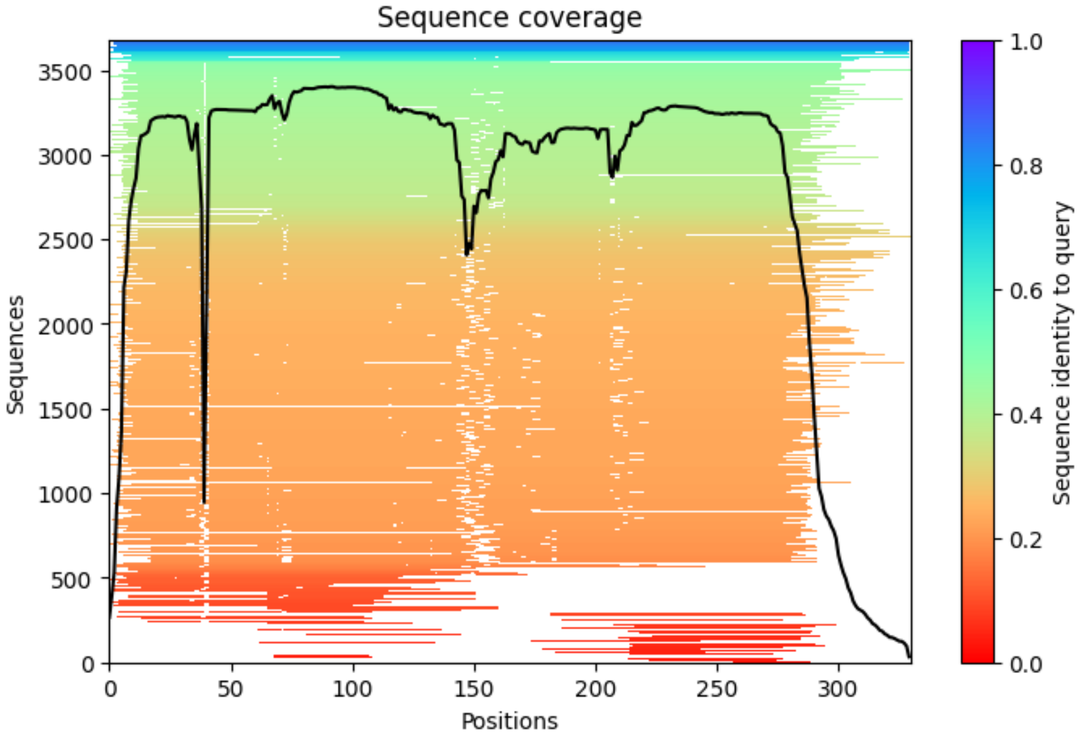
AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

Interpreting an AF2 prediction: from COLABFOLD

It is possible to obtain AF2 predictions either precalculated from the AFDB or running from COLABFOLD (and references therein)

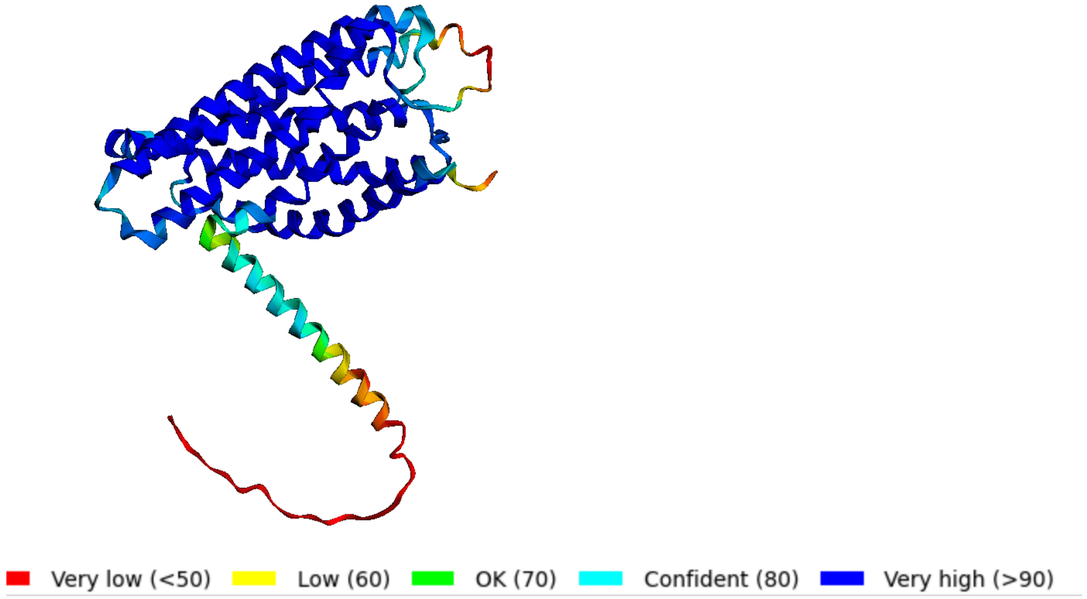
The first data obtained are about the MSA: in this case I have many many sequences mostly with high identity and coverage: very good.

As an example I took from uniprot “A0A2K5XT84” the free fatty acid receptor 2 of “*mandrillous leucophaeus*” and I ran it on COLABFOLD



Interpreting an AF2 prediction: from COLABFOLD

Once the run is done I look at the results: all the predicted structures look the same. This is a very good indication about prediction robusteness, if AF2 is not convinced about the prediction why I should be?

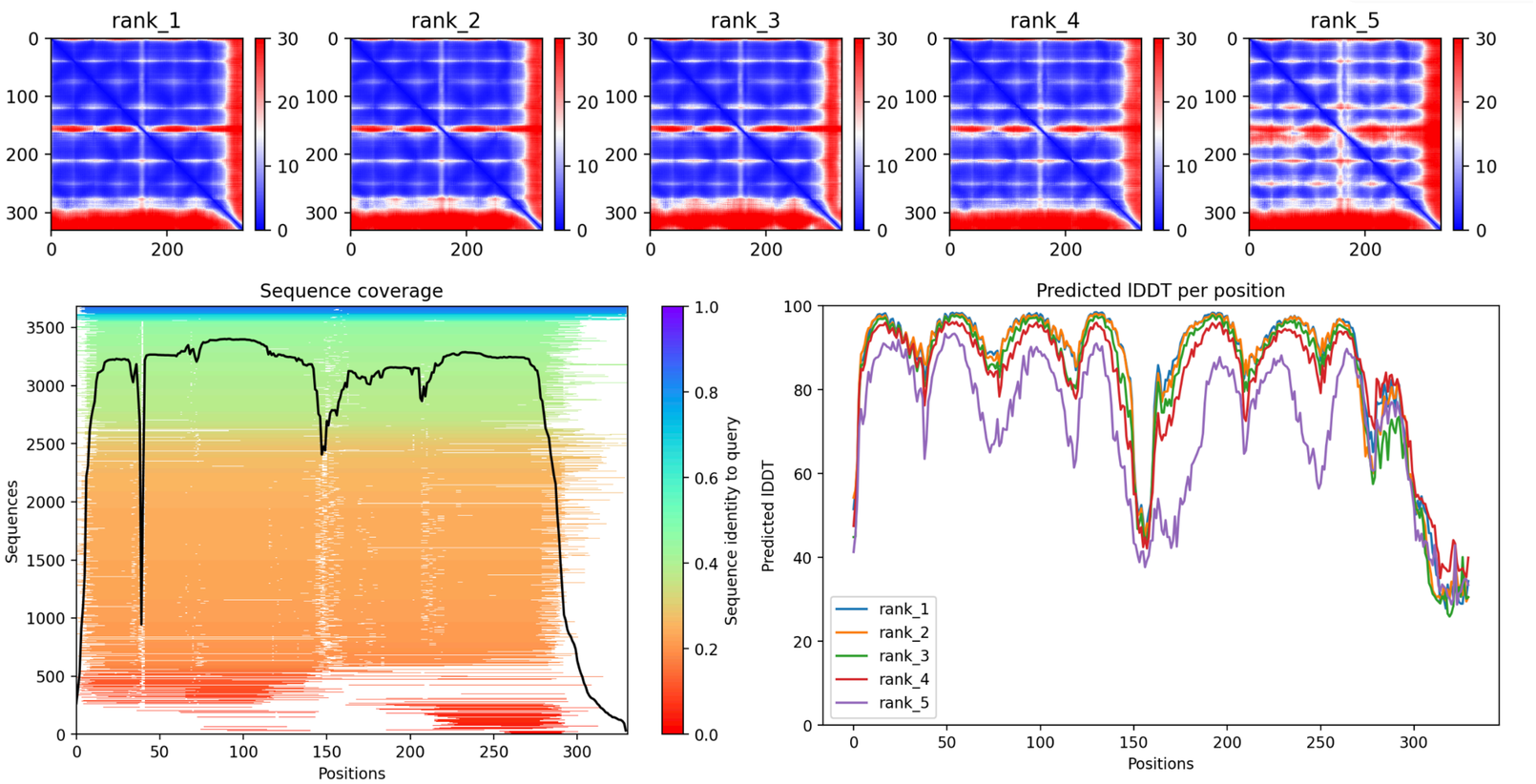


The I can look at the **PLDDT** (that is used to colour the strucutre). This the **KEY INDICATOR** of prediction quality for **TERTIARY STRUCTURE**.

Interpreting an AF2 prediction: from COLABFOLD

On TOP we have the PAE indicator that is relevant for super tertiary/quaternary structure and does report about the quality of relative organisation of tertiary units.

pLDDt is correlated with MSA, but runnign a prediction is a random process, so one structure is less good than the other even in this optimal case

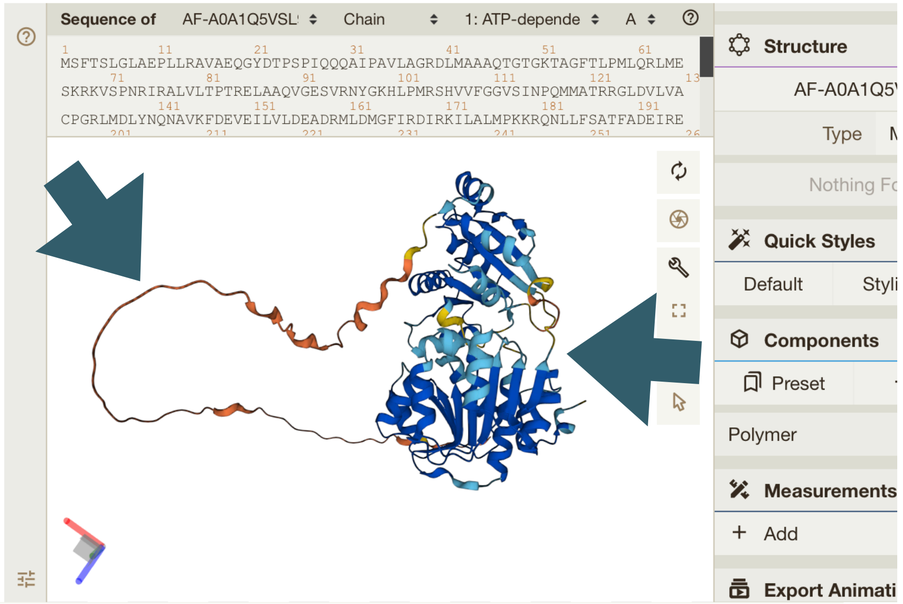


Interpreting an AF2 prediction: from AFDB

AFDB does not provide informations about the MSA and the consistency of the prediction: so we need to FOCUS on **PLDDT** and PAE

Lets' look at "Q5VSL9" from AFDB that is a prediction for the "ATP-dependent RNA helicase RhlE" from "*Aeromonas allosaccharophila*":

PLDDT is HIGH everywhere but in the DISORDERED REGIONS



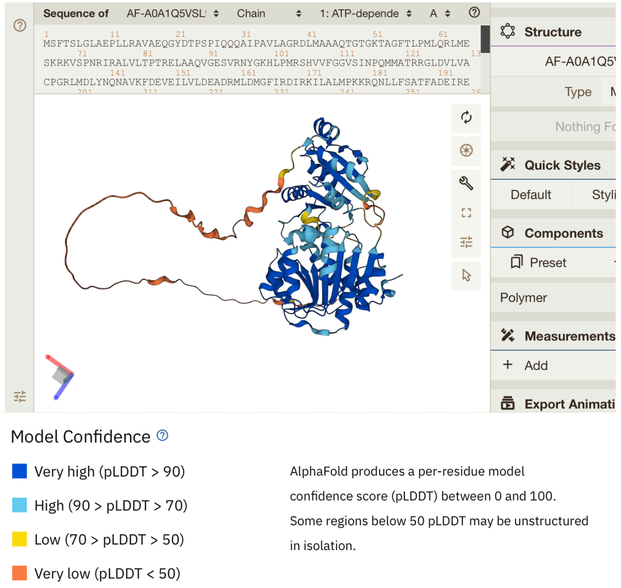
Model Confidence

- Very high (pLDDT > 90)
- High (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

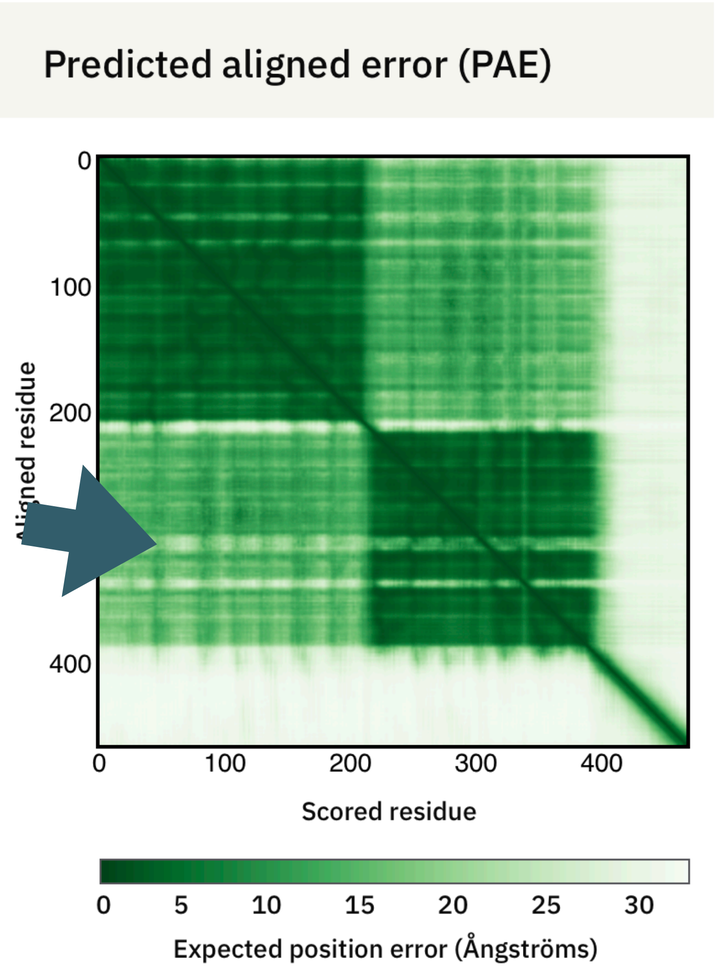
AlphaFold produces a per-residue model confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Interpreting an AF2 prediction: from AFDB

The low **PLDDT** in the region linking the two domains is reflected in the lower confidence reported for their relative orientation



HINT: when a protein has long disordered regions is better to cut them out from the prediction. Long stretch of IDP regions hamper the prediction.



UNIVERSITÀ DEGLI STUDI DI MILANO

DBS

Summary

AF2 and similar methods can provide PREDICTIONS of structure and CONFIDENCE metrics. The latter are very important to evaluate the quality (local and global) of the prediction. Even for structure in AFDB, you can run it again to be sure and get more INFO.

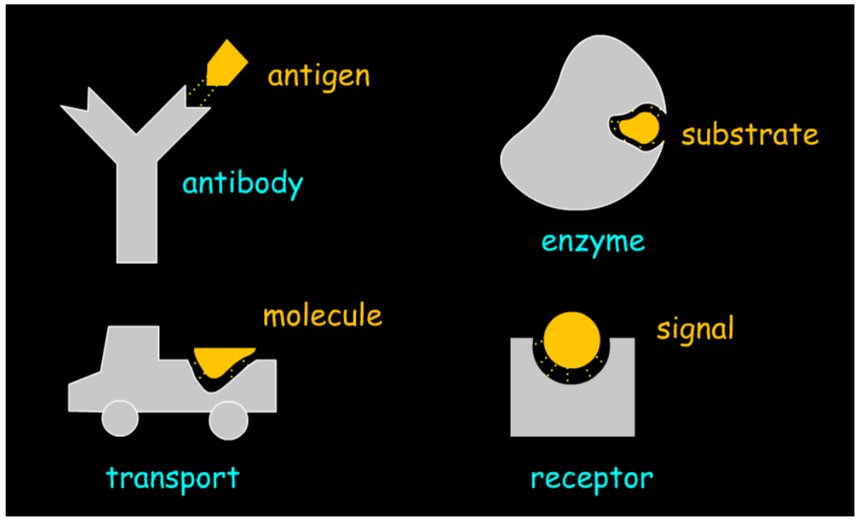
DISORDERED REGIONS will not be predicted realistically and may interfere with the rest of the prediction.

Limited MSA may be reflected in a poor PLDDT score. Remember that AF2 and related methods are designed to make predictions based on MSA.

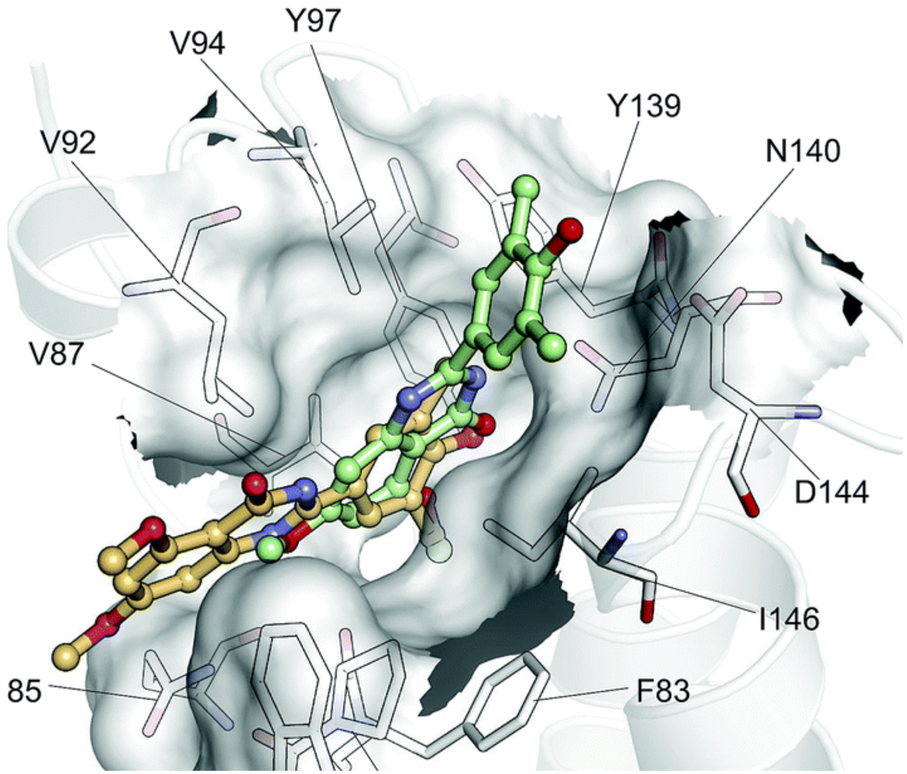
Molecular Recognition and Docking

Molecular recognition is the ability of biomolecules to recognize other biomolecules and selectively interact with them in order to promote fundamental biological events such as transcription, translation, signal transduction, transport, regulation, enzymatic catalysis, viral and bacterial infection and immune response.

Molecular docking is the process that involves placing molecules in appropriate configurations to interact with a receptor. Molecular docking is a natural process which occurs within seconds in a cell.



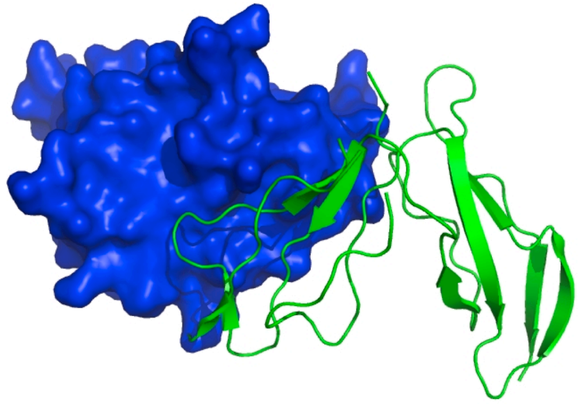
Molecular Docking: binding pose and affinity



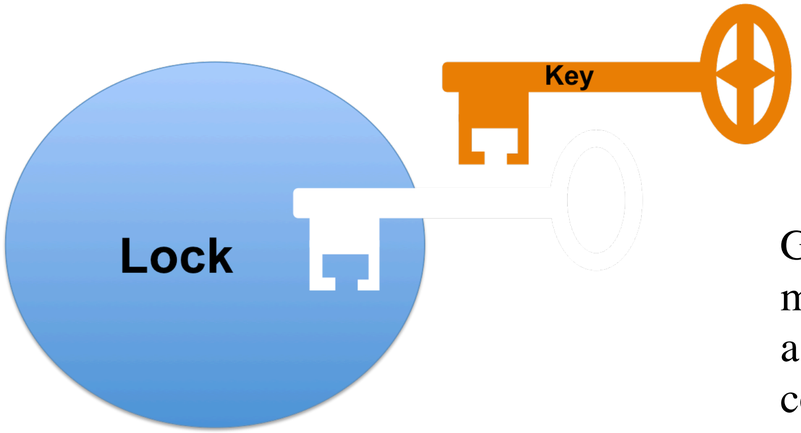
Prediction of the <u>bound structure and binding strength</u> for

- **Small molecules (virtual screening, drug discovery, ..)**
- **Macromolecular complexes (protein-protein, protein-nucleic, …)**

Complexation can be associated with conformational changes



Lock and Key



Emil Fischer (1894)

Specificity in enzyme-substrate recognition

Generally speaking the idea is that specificity in molecular recognition is the result of rigid surface that as a consequence can bind together only when exactly complementary (**Lock and Key model**)

Docking algorithm: (0 level)

- analyse surface of the two molecules
- find possible regions of binding (surface compatibility -> shape and electrostatic interactions)
- minimise the energy.

Actually in many cases recognition mechanisms are more complex, with in principle flexibility on both the receptor and the ligand **(Flexible Docking)**

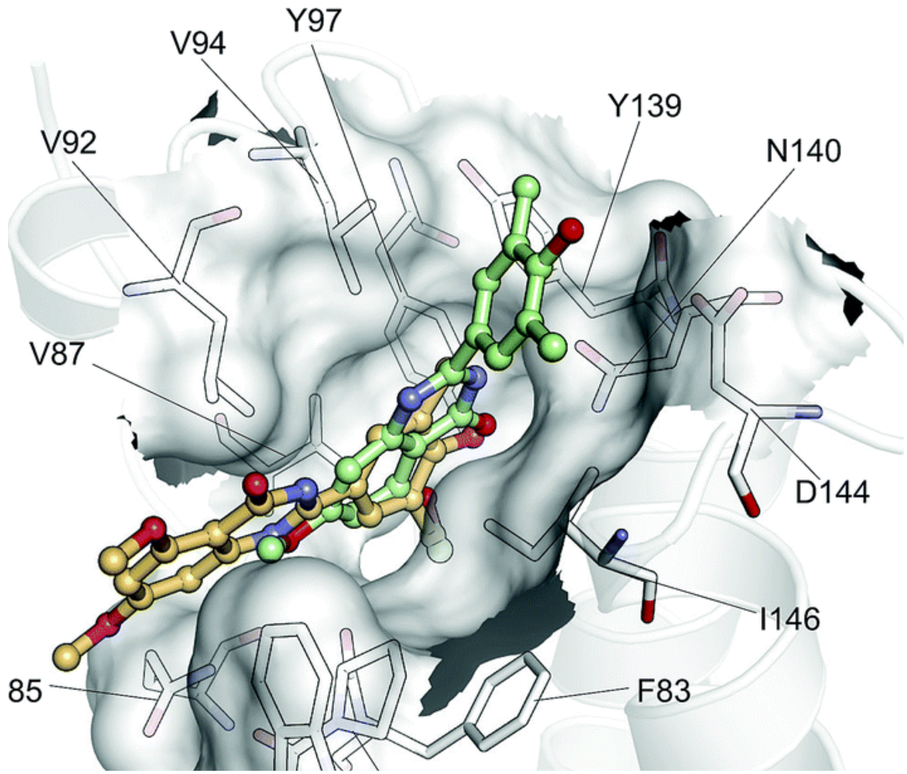
Docking mechanisms

Lock-and-Key: this is simple, the assumption is that both the receptor and the ligand are rigid and their structure is ideal for binding. The next step is to accept that the ligand (this is more specific for the binding of small molecules) can be flexible, while the receptor is rigid.

Induced Fit: In 1958 Daniel Koshland introduced the "induced-fit theory". The basic idea is that in the recognition process, both ligand and target mutually adapt to each other through small conformational changes, until an optimal fit is achieved. This is traduced in what is usually called the flexible-docking where both active site region and the ligand are allowed a certain degree of flexibility.

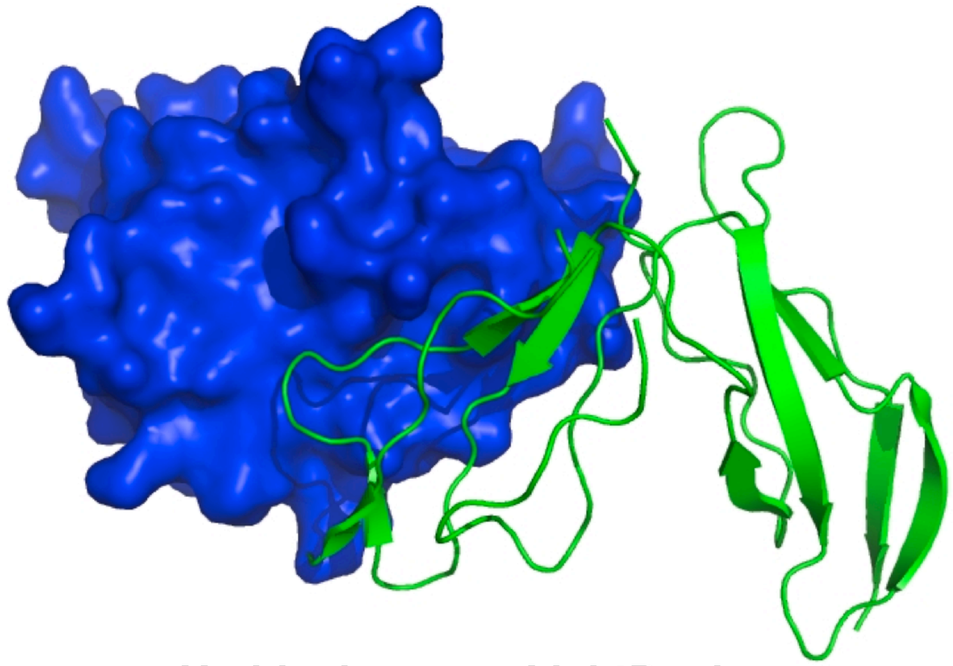
Conformational Selection: this recognises that both the ligand and the receptor can be in many many different configurations, and that binding can only occur when relatively optimal configurations are populated. This is traduced in the so called ensemble docking, where multiple configurations of the receptor are used at the same time each can also be considered locally flexible to account for additional induced fit.

Molecular Docking: softwares



Autodock vina
UCSF dock
Glide
...

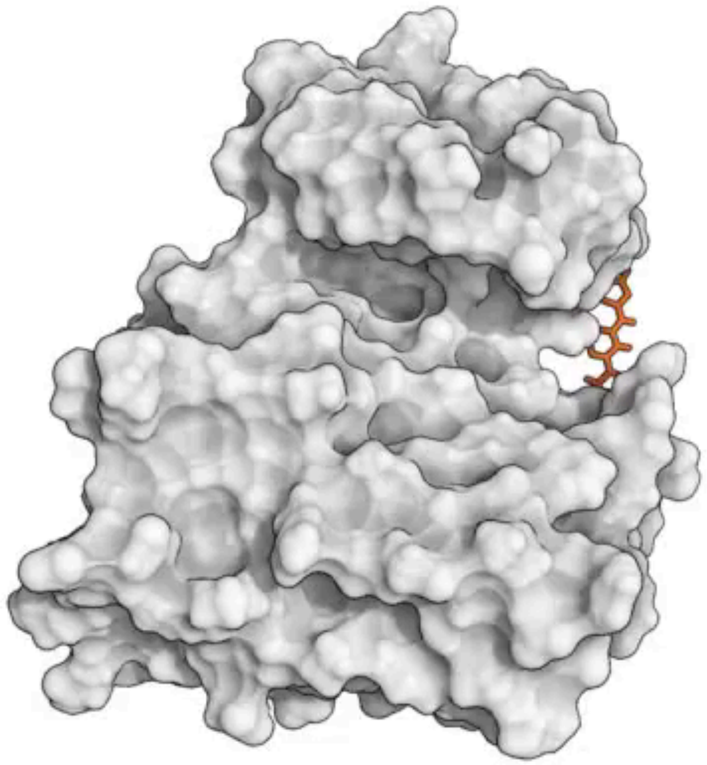
DiffDock
...



Haddock
Rosetta Dock
ClusPro
...

LightDock
AFMultime
RosettaFold2
...

MD simulations



In principle MD simulations can be used to search for ligand binding, but the are problems:

1) Parametrize force-fields for all the possible chemicals is hard (this is more relevant for ligand binding than for protein complexes)
2) The binding time scale can be slow in particular when associated with conformational changes (the problem of sampling). This is even more relevant for protein complexes where the systems become large and so the simulations become slow.

1.Shan, Y., Kim, E. T., Eastwood, M. P., Dror, R. O., Seeliger, M. A. & Shaw, D. E. How does a drug molecule find its target binding site? *J Am Chem Soc* **133,** 9181–3 (2011).

Rigid or semi-flexible docking: a simplified simulation approach

There are two problems:

1. **How to generate quickly as many ‘reasonable’ poses as possible (including or not flexibility)**
2. **How to distinguish (‘score’) a good pose from a bad one (will the true complex score better than all other possible complexes?)**

A number of poses can be generated at random by rotation and translation of a molecule around the other (using Monte Carlo or genetic algorithms) and then they can be scored for example using a physicochemical inspired scoring function (a simplified force-field). Protein Flexibility can be introduced using multiple structures, while ligand flexibility can be introduced allowing the molecule to rotate around “rotatable” bonds.

VINA scoring function

This is what is usually called a “scoring function” that is a mathematical object that given some numbers return a single number, e.g. all the x, y, z coordinates of a given configuration put into the scoring function return a single number that is the score for that configuration.

$$\Delta G_{binding} = \Delta G_{gauss} + \Delta G_{repulsion} + \Delta G_{hbond} + \Delta G_{hydrophobic} + \Delta G_{tors}$$

ΔG_{gauss}
Attractive term for dispersion, two gaussian functions

$\Delta G_{repulsion}$
Square of the distance if closer than a threshold value

ΔG_{hbond}
Ramp function - also used for interactions with metal ions

$\Delta G_{hydrophobic}$
Ramp function

ΔG_{tors}
Proportional to the number of rotatable bonds

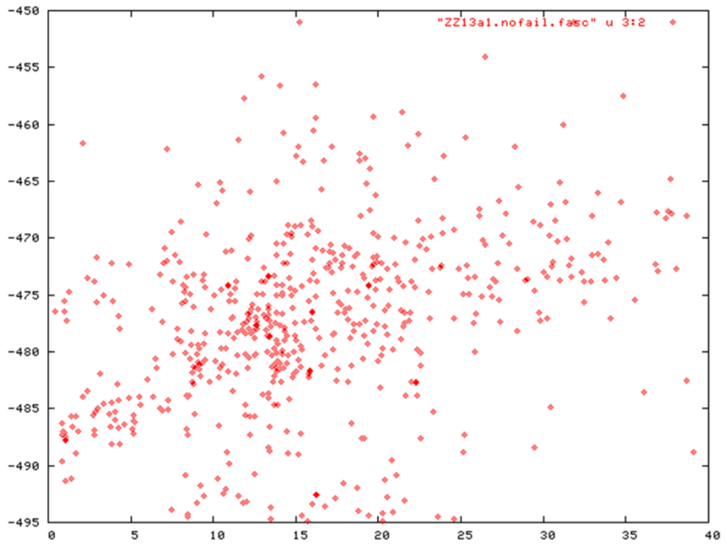
These are all related to intermolecular interactions. The hydrophobic term can be seen as the one accounting for solvation.

This is then only source of internal energy

http://www.pdbbind.org.cn/index.asp

Scoring function and convergence

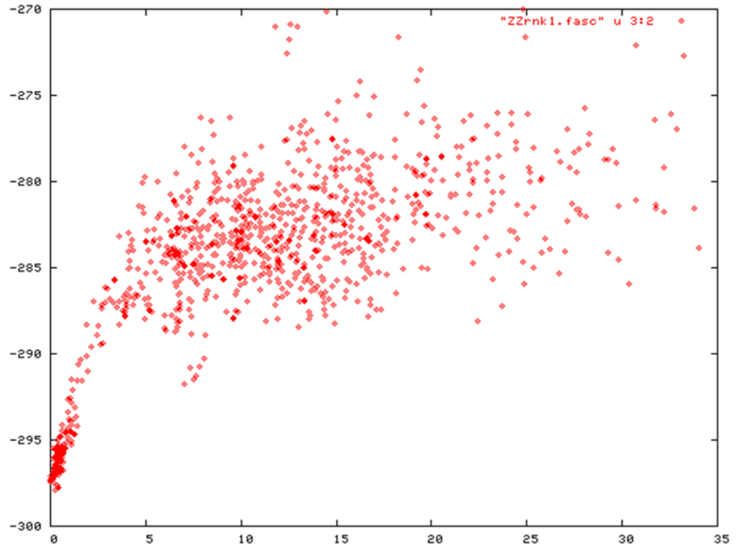
Energy



RMSD

Bad

Energy



RMSD

Good

Monte Carlo sampling

In principle one could generate configurations at random, evaluate their scoring function, and keep the lowest energy one. The limit of this approach is that most of the time will be spent evaluating scoring functions for bad scoring configurations.

Metropolis Monte-Carlo (but often refereed simply as Monte Carlo) is an algorithm to evaluate a probability distribution.

$$P(x) = \frac{\exp\left[\frac{-U(x)}{k_B T}\right]}{\sum \exp\left[\frac{-U(x)}{k_B T}\right]}$$

Calculating P(x) would mean calculating it for all possible configurations, but the probability is simply proportional to:

$$f(x) = \exp\left[\frac{-U(x)}{k_B T}\right]$$

$$f(x) \propto P(x)$$

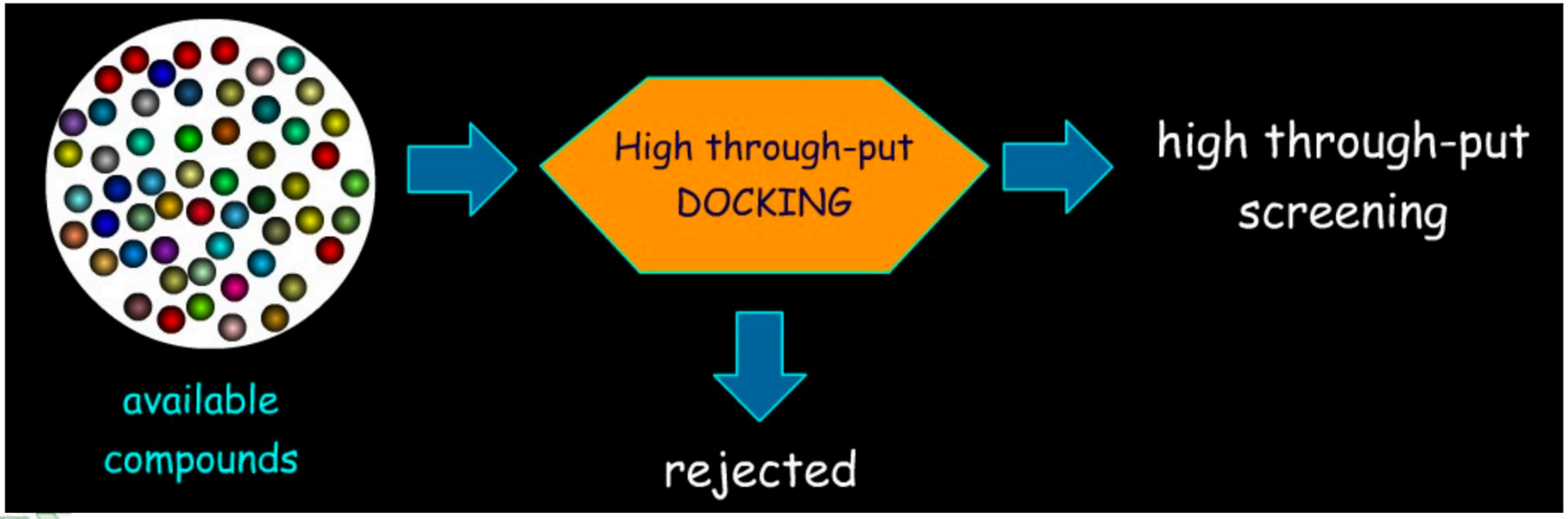
Starting from a configuration we can calculate its energy or score, then we can generate a second configuration and get its score and we need to decide if to keep it or not. How? Their relative probability is:

$$\frac{f(x_{new})}{f(x)} = \exp\left(\frac{-U(x_{new}) + U(x)}{k_B T}\right)$$

Do we keep the new one or not? We generate a random number (0..1), if it is smaller than the above ratio we keep the new configuration, otherwise the old.

Virtual Screening

- When the goal of docking is to dock all the compounds of a library (the molecules being available or not yet synthesized), the process is called virtual screening or high throughput docking
- Virtual screening identifies active compounds in a large database and ranks them by their affinity to the receptor
- The method is not used to recognize active molecules but to eliminate those that are likely to be inactive



Complementary approaches

- Pocket identification algorithms: what are the features of a good binding site? (P2rank, …)
- Ligand design methods (which chemistry is optimal to bind in a given pocket?) here generative AI can help a lot (DiffSBDD, …)
- Structure prediction methods are trying to implement also the prediction of chemicals: Alphafold3, RosettaFold-AllAtom, …
- Prediction of ADME/PK (Absorption, Distribution, Metabolism, Excretion and Pharmacokinetics) refers to the complete set of properties of a drug molecule describing its entry into the body, residence time within the body, distribution to organs, metabolic transformations involved in its clearance, and routes of elimination