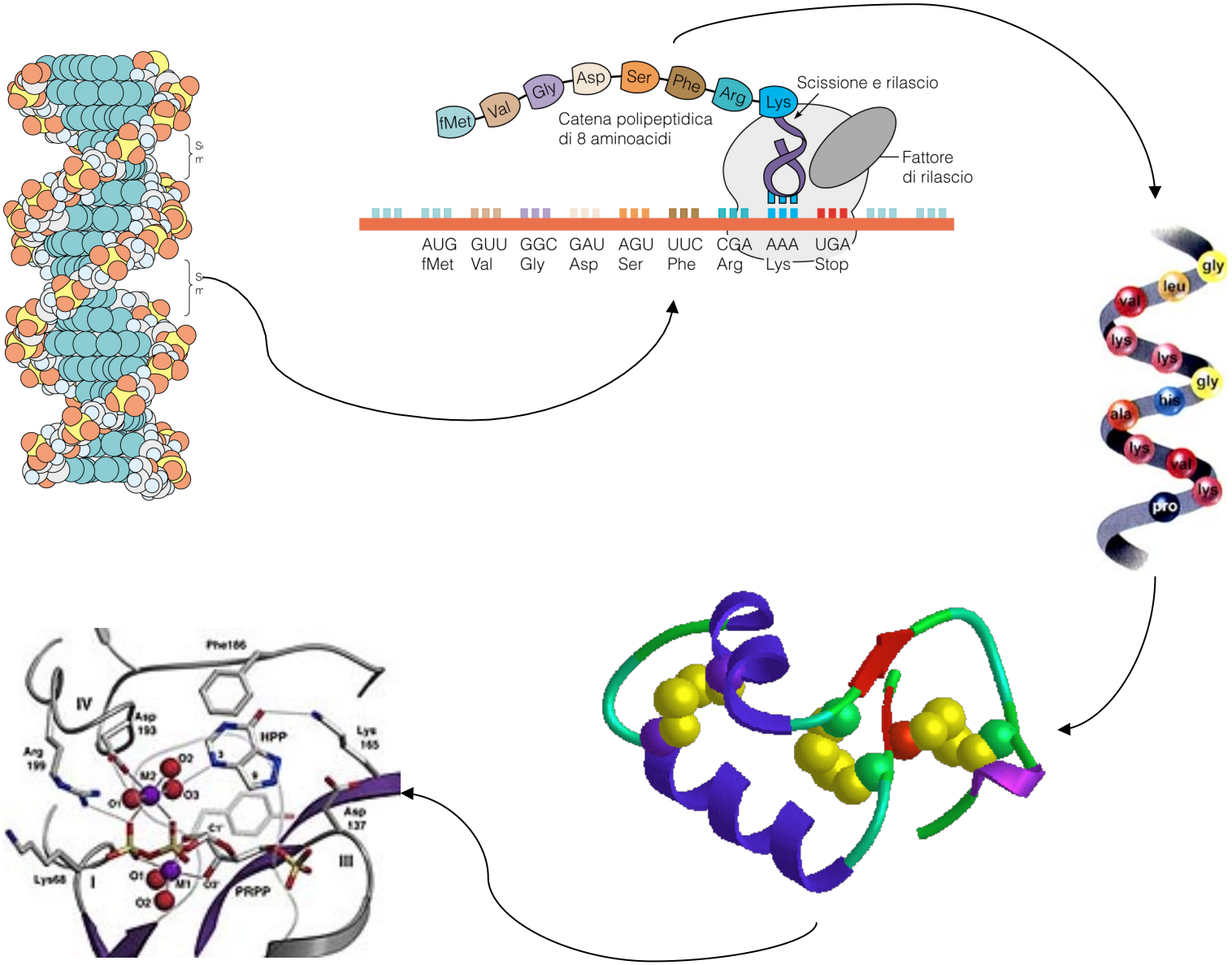
Lesson 11.
Content

1. The protein folding problem.

2. Comparative (Homology) modelling. Choice of template and alignment.

3. Comparative(Homology) modelling. Modelling of the protein core and loops

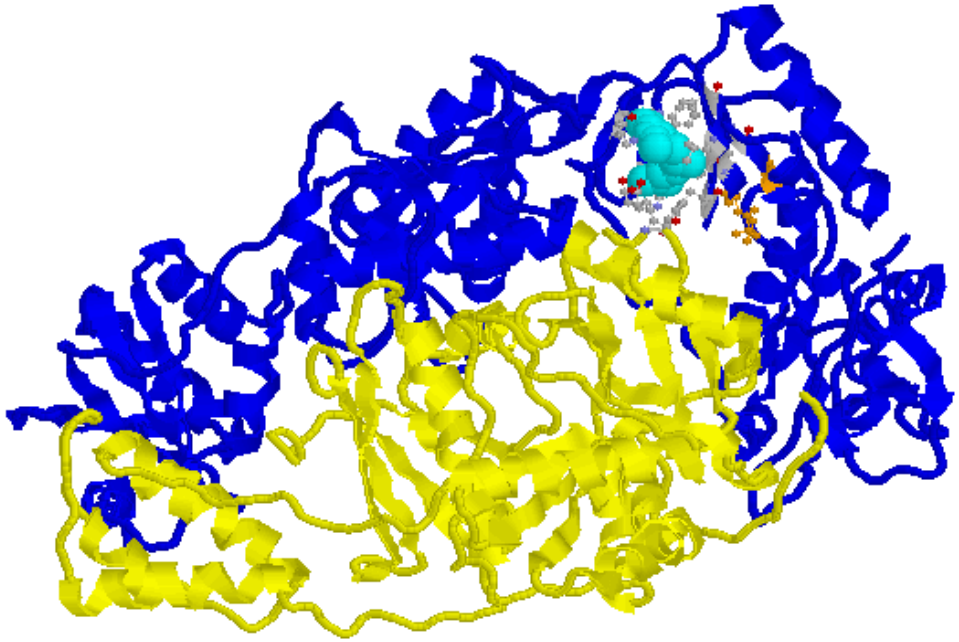
4. Comparative(Homology) modelling. Optimization and quality check.


fMet
Val
Gly
Asp
Ser
Phe
Arg
Lys
Scissione e rilascio
Catena polipeptidica
di 8 aminoacidi
Fattore
di rilascio
AUG GUU GGC GAU AGU UUC CGA AAA UGA
fMet Val Gly Asp Ser Phe Arg Lys Stop
gly
leu
val
lys
lys
gly
his
ala
lys
val
lys
pro
Phe186
IV
Asp
193
Arg
199
HPP
Lys
165
O2
M2
O3
O1
C1'
Asp
137
III
Lys68
I
O1
M1
O2
O2'
PRPP

Why knowing the 3D structure of proteins?

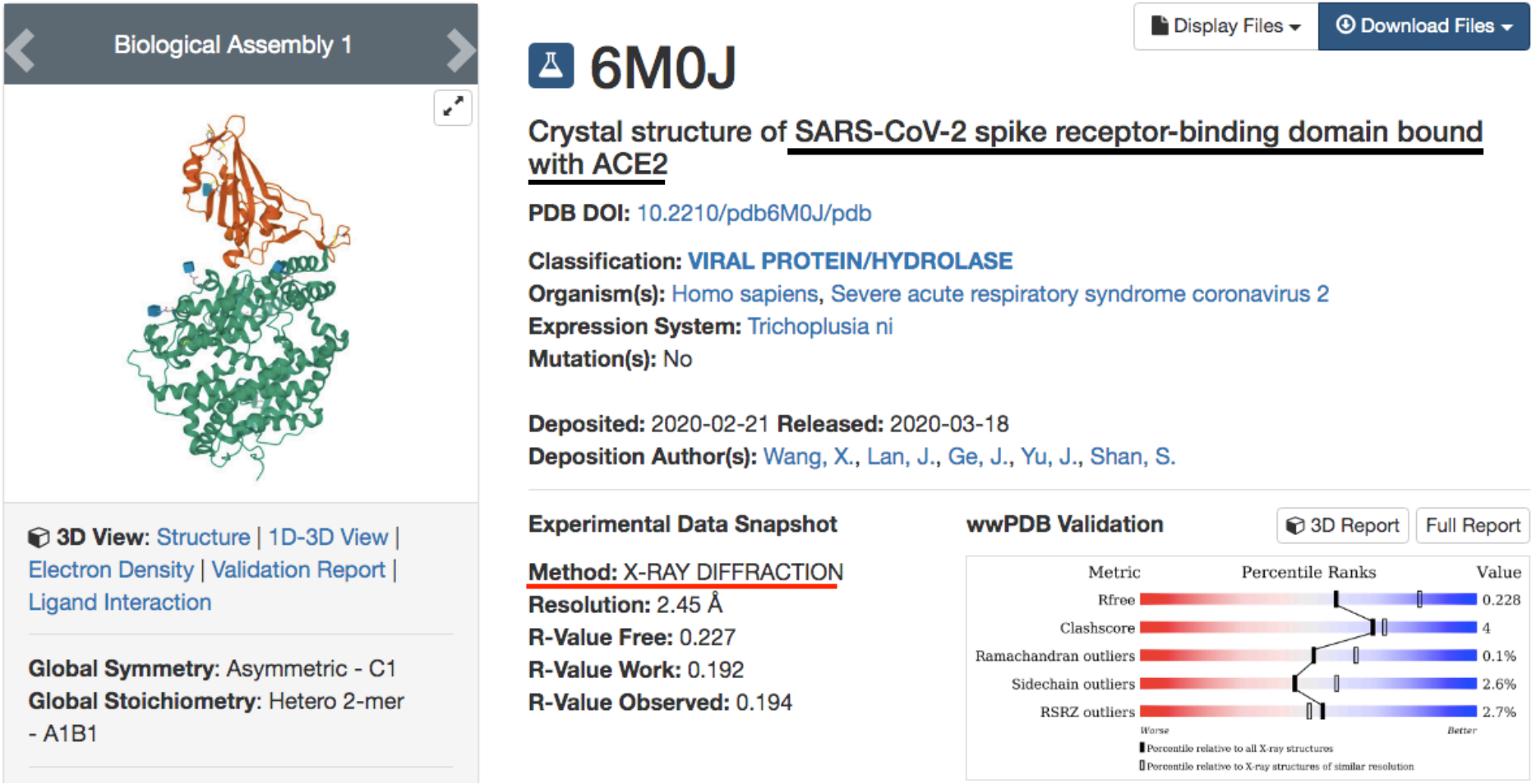
Experimental Models (X-Ray, NMR, EM)

Theoretical Models *(protein modelling)*



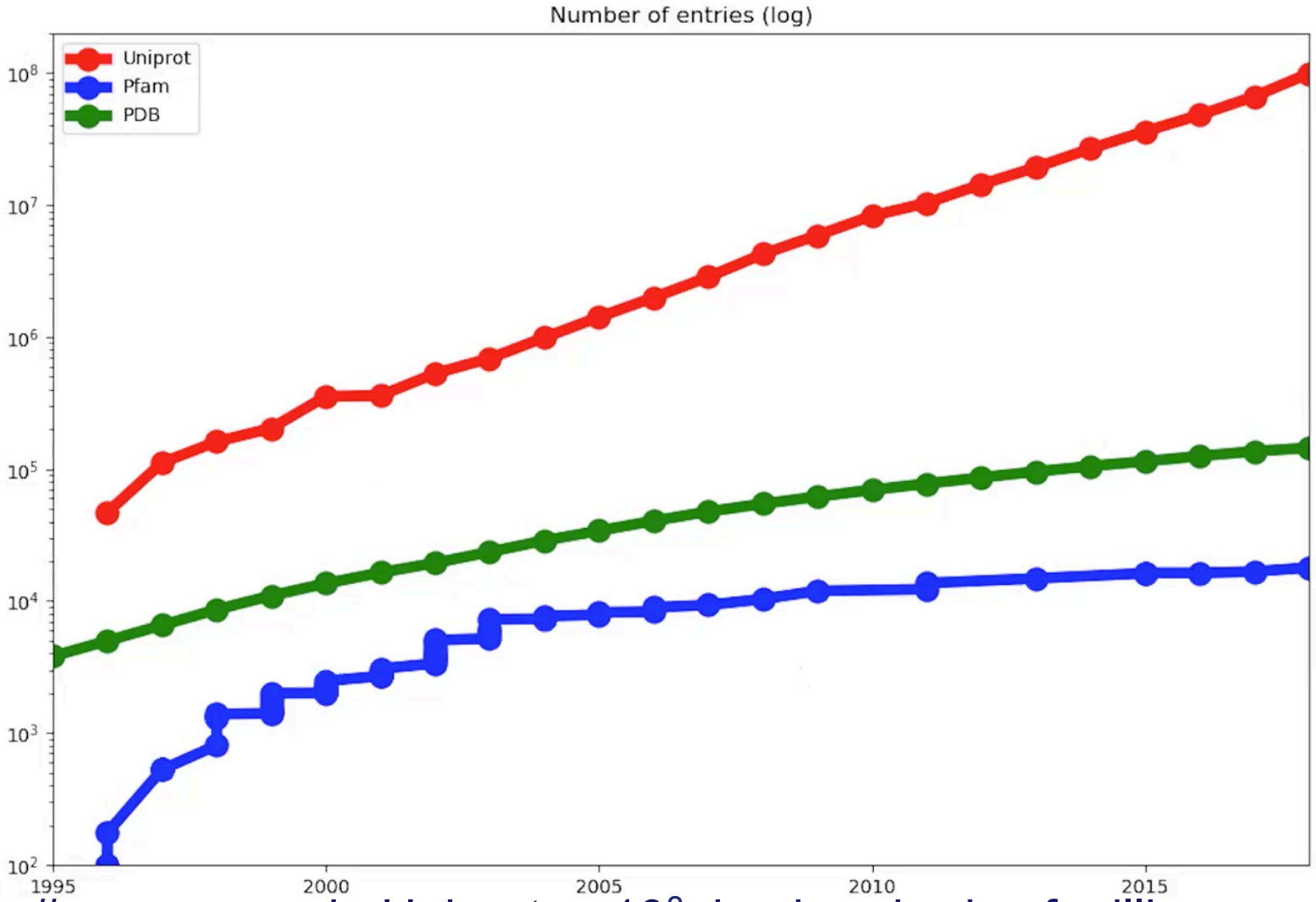
1. Unraveling the function/s if not known
2. Study of the molecular mechanisms of interaction
3. Study of biochemical mechanisms
4. Rational drug design

Example of experimental 3D structure in the PDB



Besides the coordinates of the macromolecule(s) structure, additional information is reported, including especially the experimenthal method (here X-ray diffraction*) and relative quality parameters (*Resolution *and* R-values*, the lower the better)*

Why predicting the 3D structure of proteins?

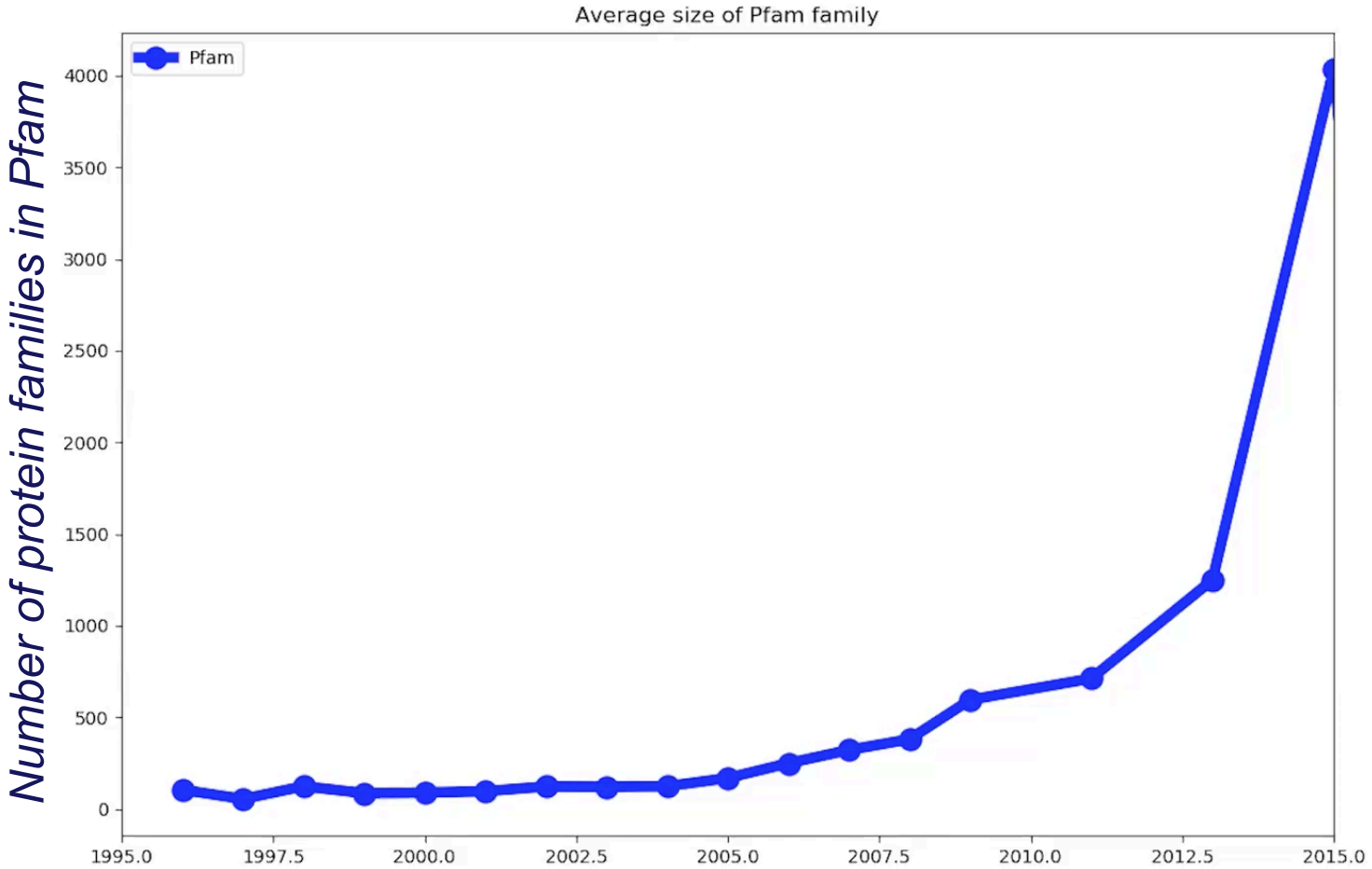


sequences in Uniprot: ≈ 10^8, i.e. hundreds of millions

#3D structures in the PDB: ≈ 10^4, i.e. tens of thousands

Be aware of the logarithmic scale!

Why predicting the 3D structure of proteins?

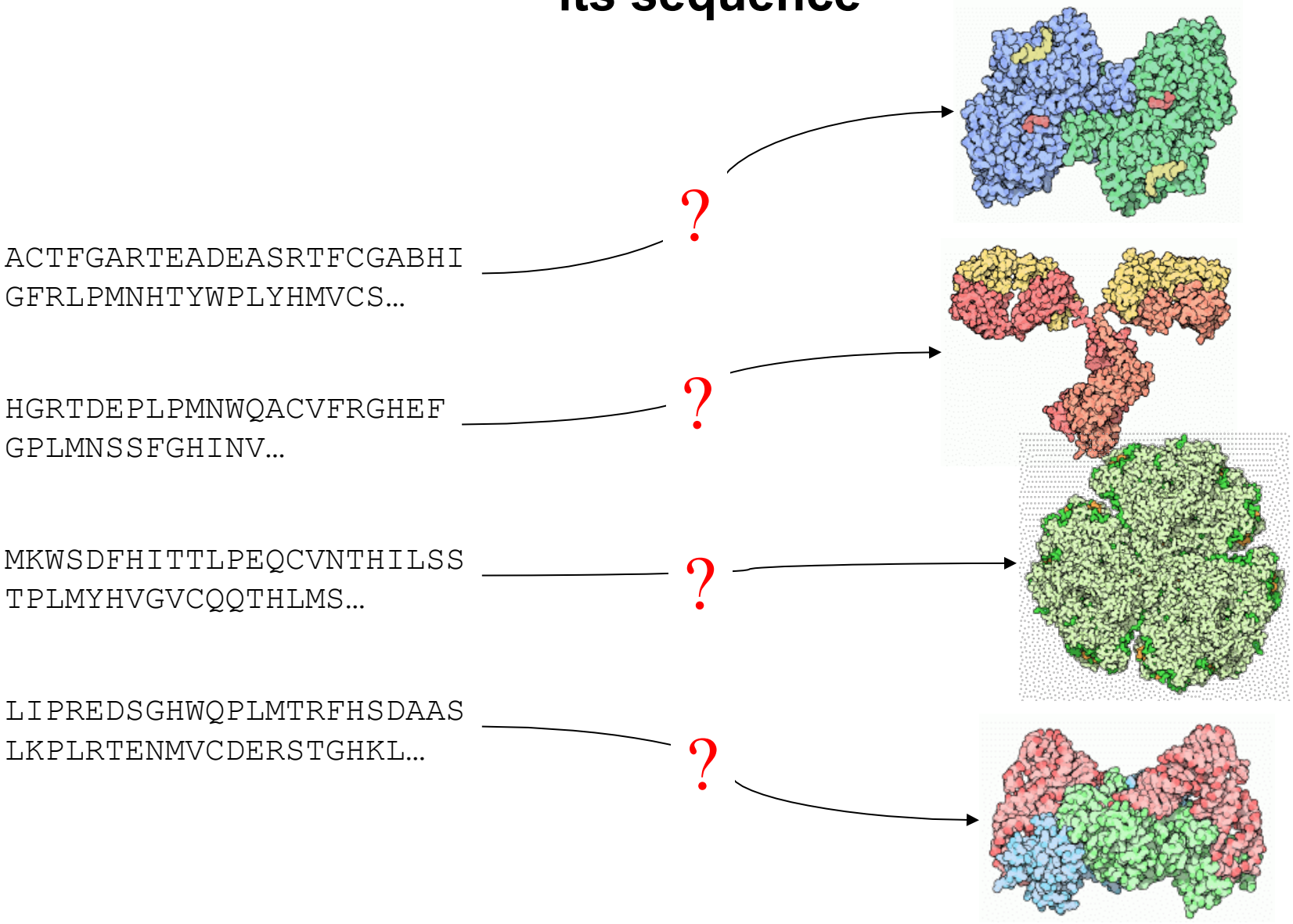


Between 2012 and 2015, the number of protein families has more than doubled, then it remained pretty stable

The hundreds of millions of known protein sequences are grouped in a finite number (thousands) of protein families

The folding problem:

Predicting the fold (3D structure) of a protein starting from its sequence



The folding problem:

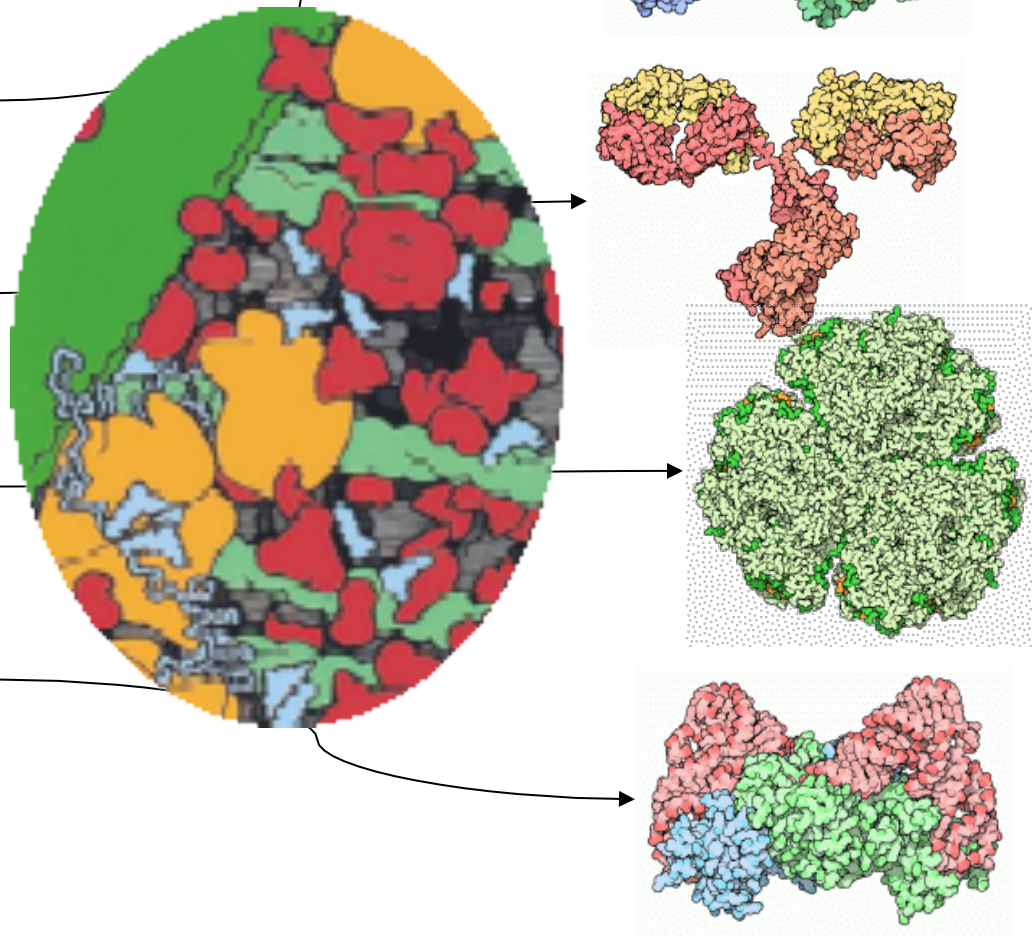
Predicting the fold (3D structure) of a protein starting from its sequence

ACTFGARTEADEASRTFCGABHI
GFRLPMNHTYWPLYHMVCS…

HGRTDEPLPMNWQACVFRGHEF
GPLMNSSFGHINV…

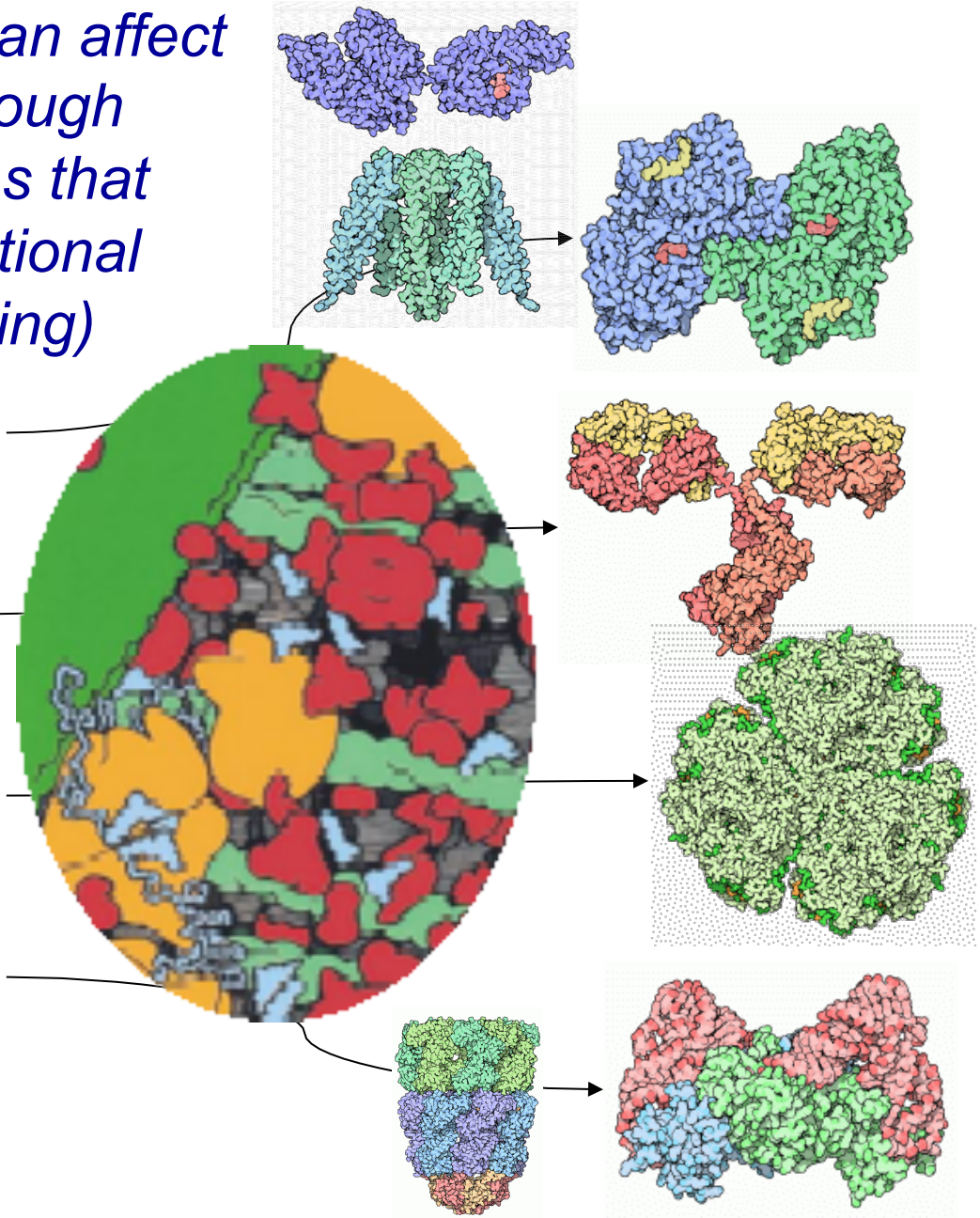
MKWSDFHITTLPEQCVNTHILSS
TPLMYHVGVCQQTHLMS…

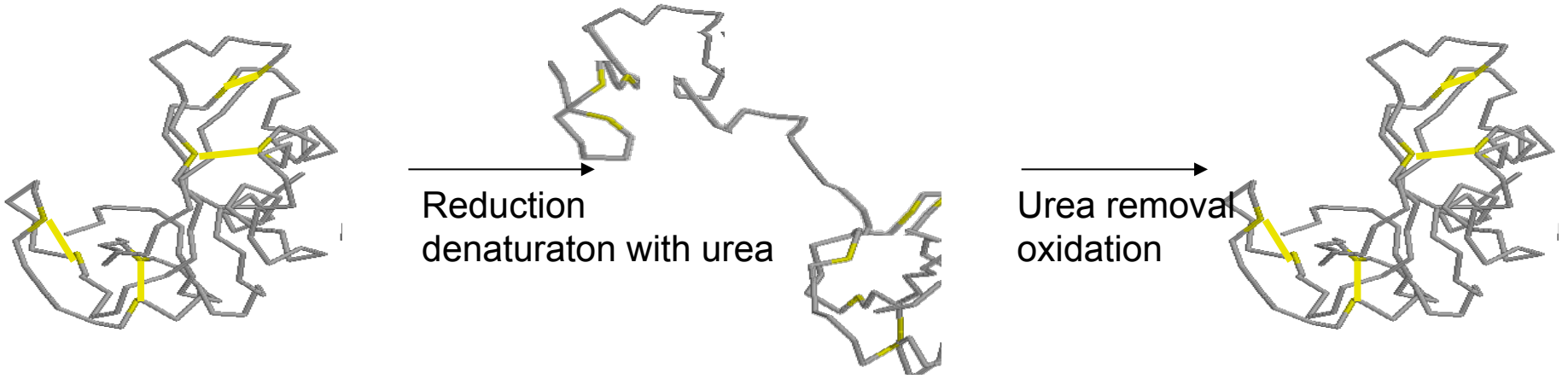
LIPREDSGHWQPLMTRFHSDAAS
LKPLRTENMVCDERSTGHKL…



The folding problem:

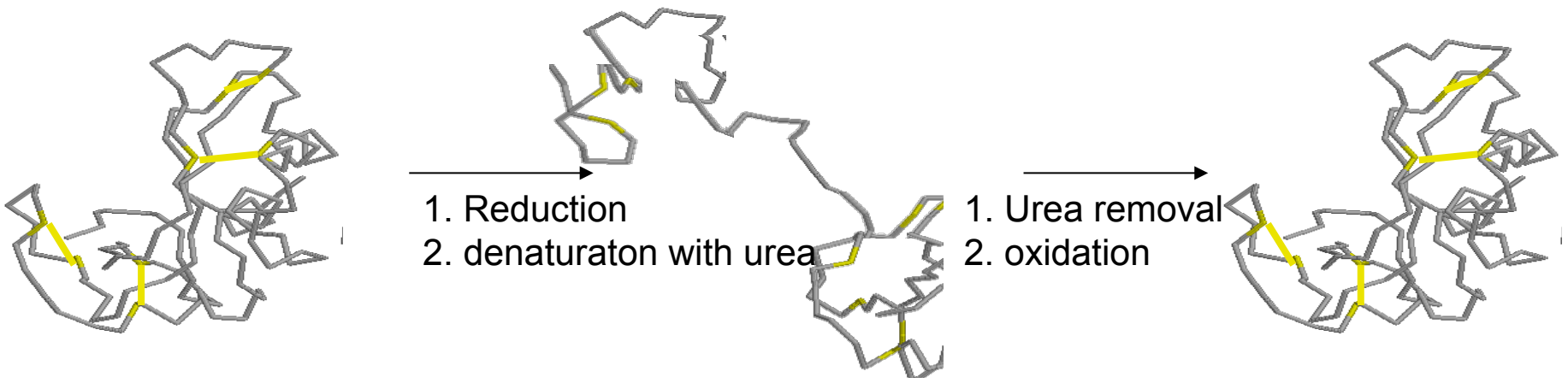
The cell environment can affect the folding, e.g. through chaperones (proteins that assist the conformational folding and unfolding)



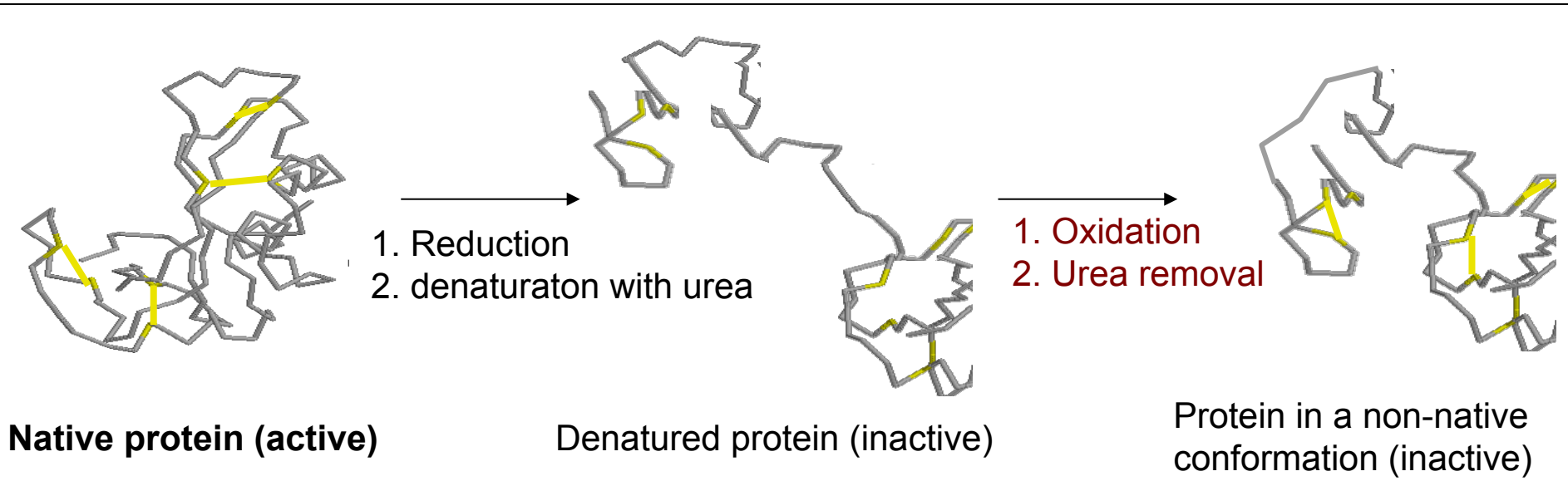


Native protein (active) Denatured protein (inactive) **Native protein (active)**

The Anfinsen's experiment



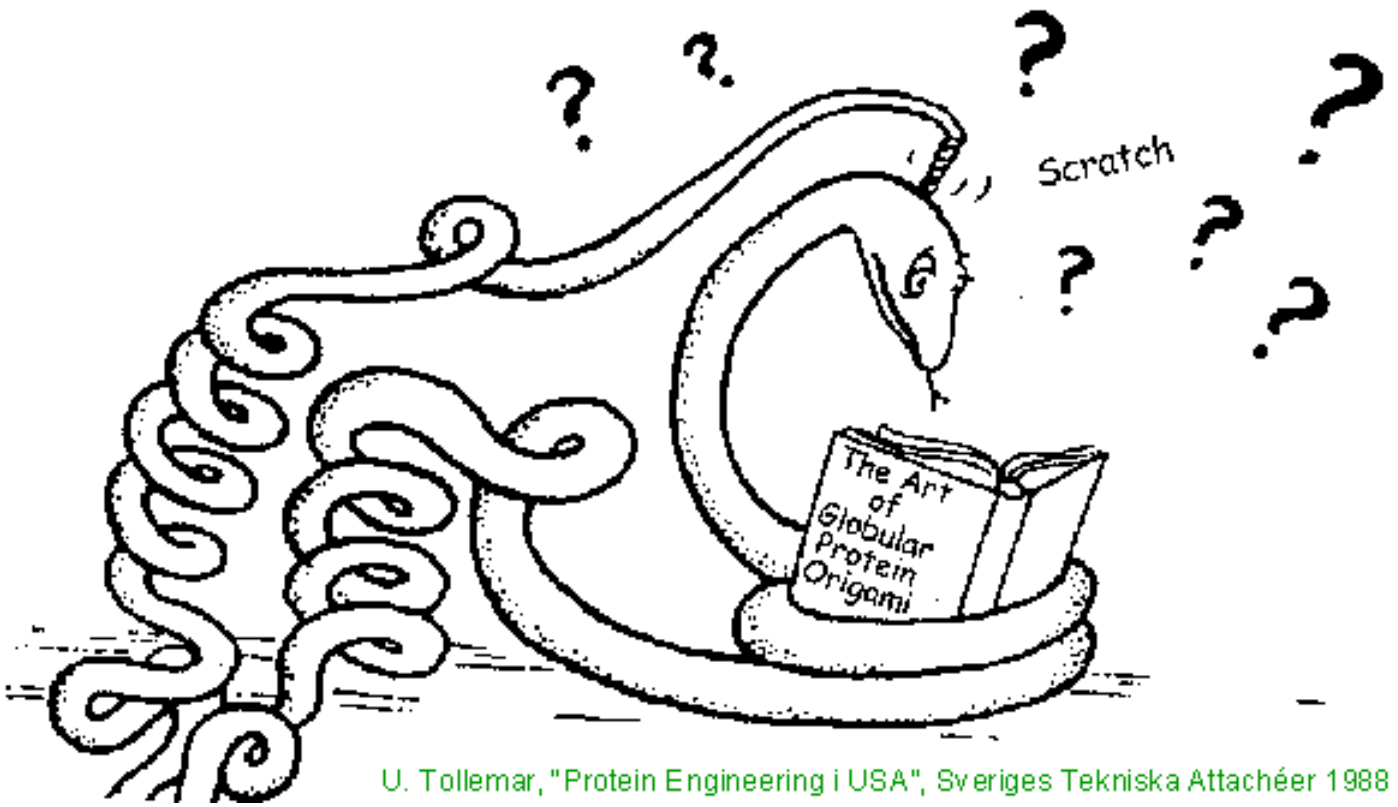
Experiment



Control

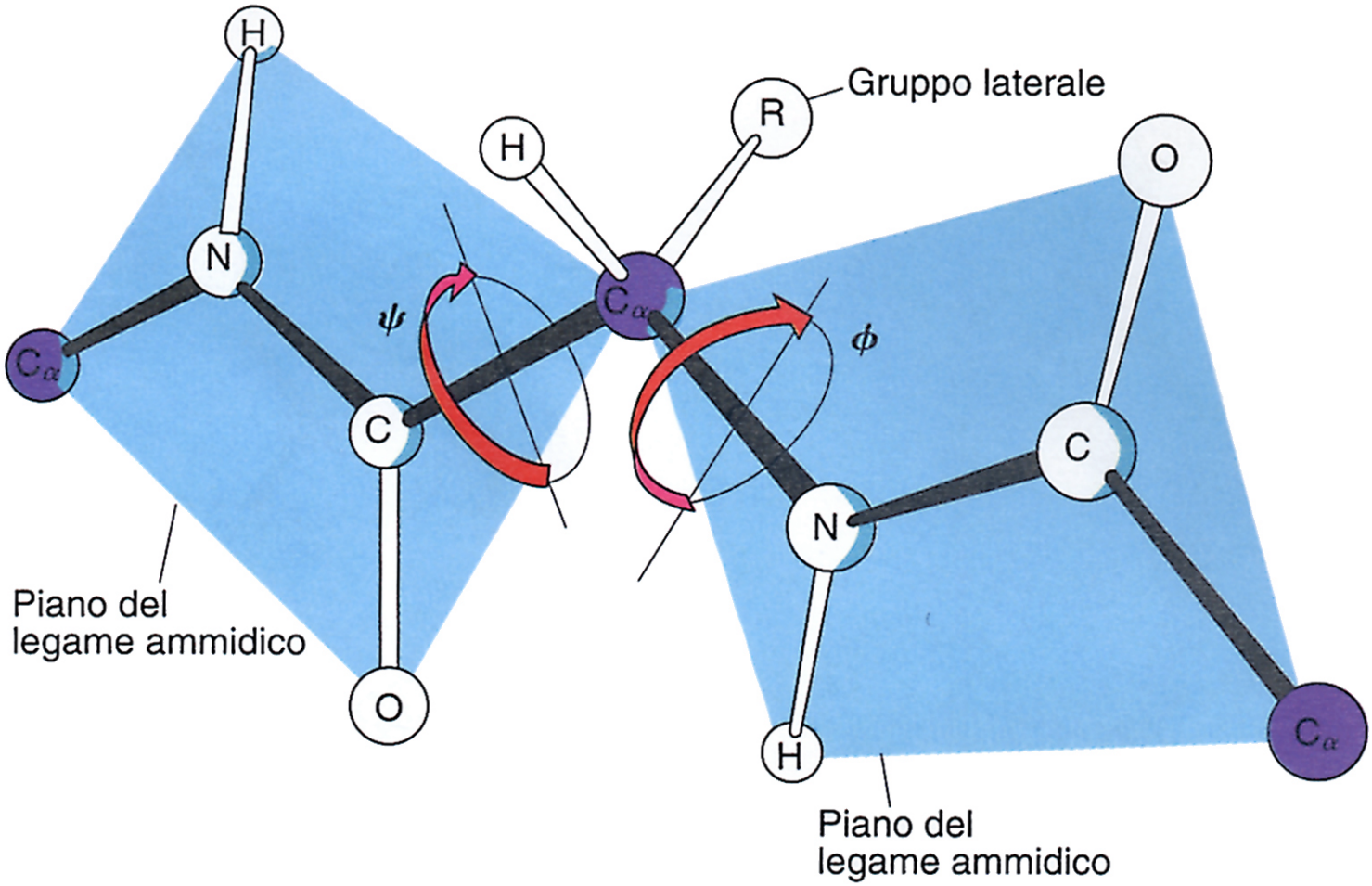
The Anfinsen's experiment

"Now, pack the hydrophobic core,
fold helix A along the dotted line,
taking charged residues of A
close to the ionic groups on the
surface of helix B…"



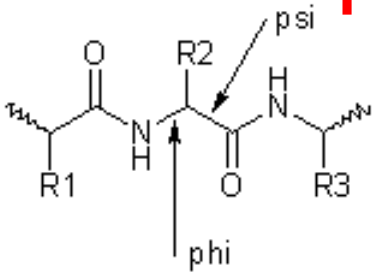
U. Tollemar, "Protein Engineering i USA", Sveriges Tekniska Attachéer 1988

The Levinthal's paradox

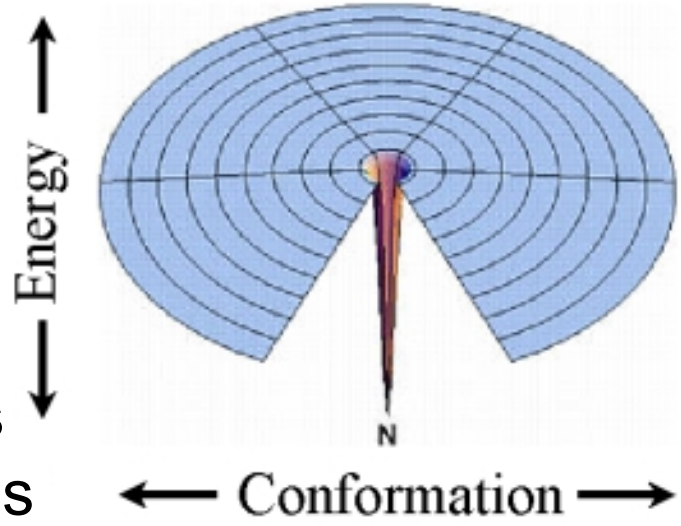


The protein backbone can rotate around the dihedral angles ϕ & ψ

The Levinthal's paradox



Even if we changed ϕ e ψ only by Intervals of 120 degrees…



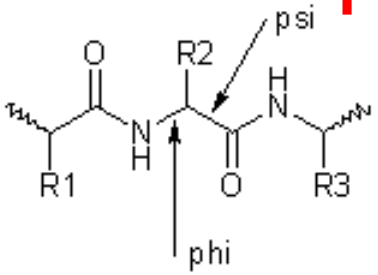
N amino acids: $3^{2(n-1)}$ alternatives

100 amino acids: 10^{47} alternatives

If each attempt would take 100 femtoseconds (10^{-13} sec) trying all the possible combinations would take:

$10^{47} \times 10^{-13}$ seconds = 10^{34} seconds = 2×10^{27} years !!

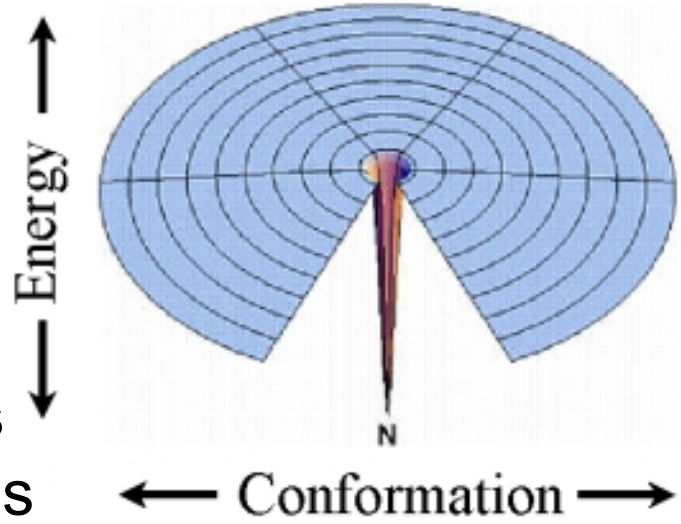
The Levinthal's paradox



Even if we changed ϕ e ψ only by Intervals of 120 degrees…

N amino acids: $3^{2(n-1)}$ alternatives

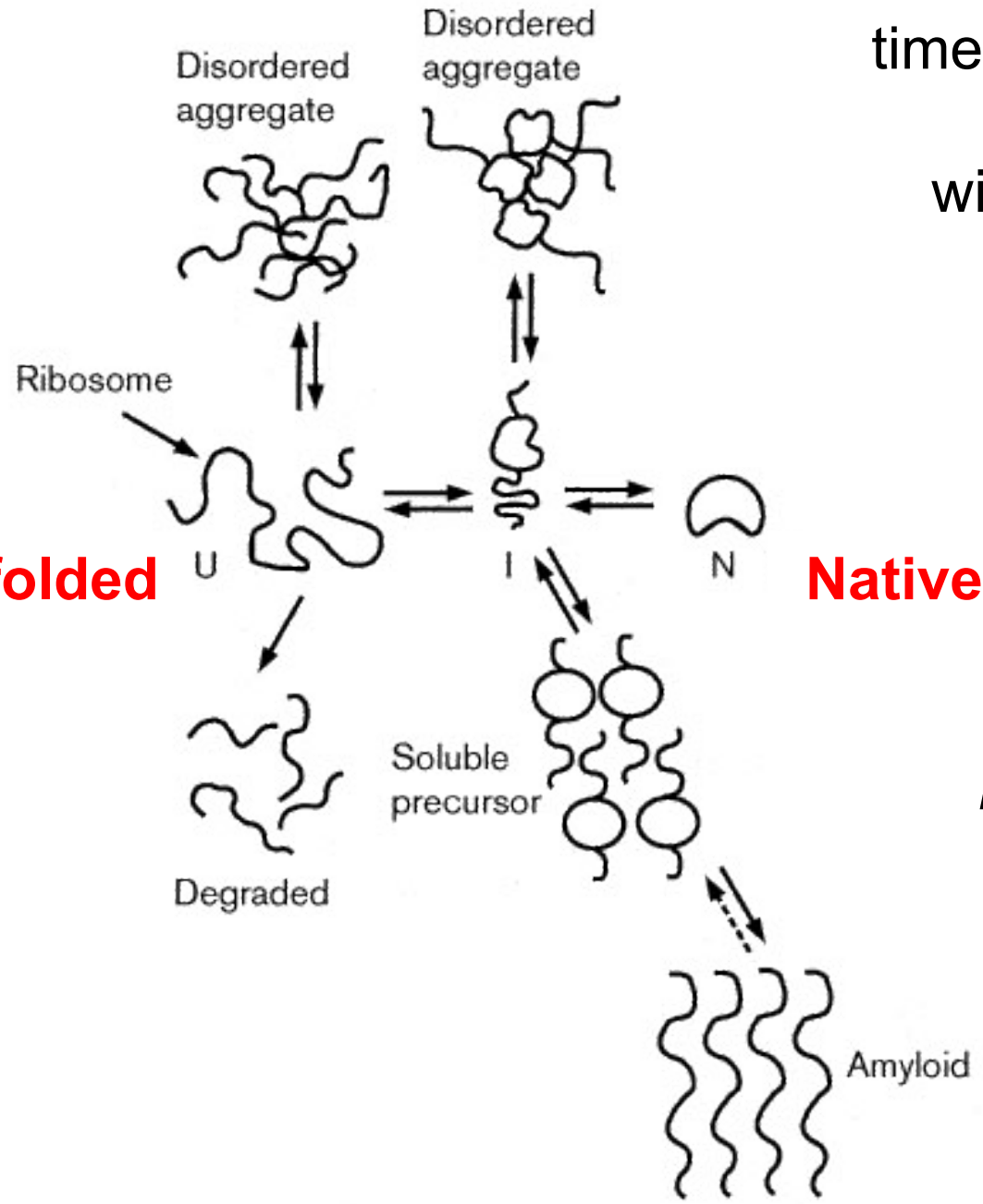
100 amino acids: 10^{47} alternatives



If each attempt would take 100 femtoseconds (10^{-13} sec) trying all the possible combinations would take:

$10^{47} \times 10^{-13}$ seconds = 10^{34} seconds = 2×10^{27} years !!

And we are considering only (a subset) of the backbone conformations

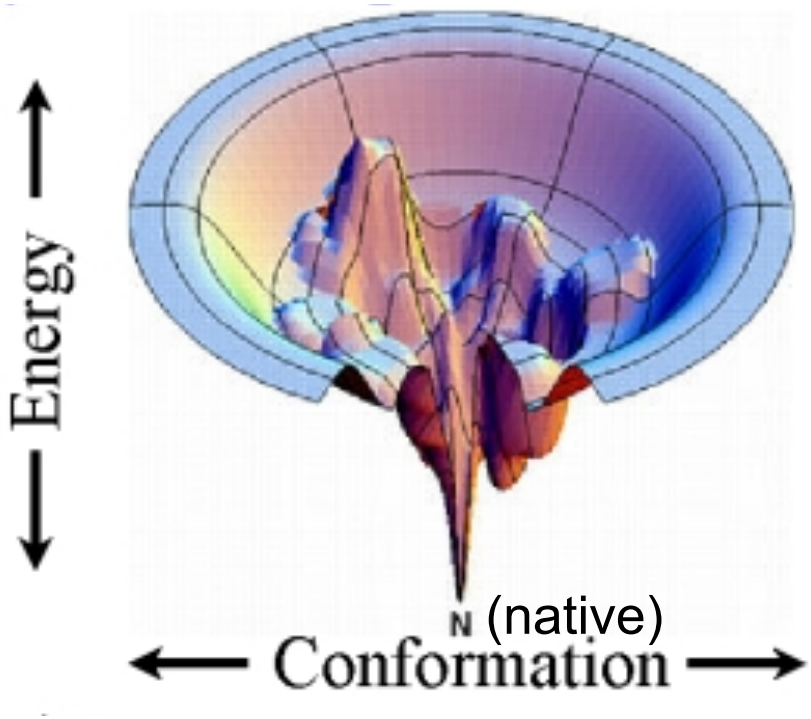


Most proteins fold on timescales on the order of a millisecond (ms), with a median of ~5 ms

Unfolded **Native**

The "speed limit" of protein folding has been proposed to be set by N/100 μs, where N is the length of the protein (note: this only works with single-domain proteins)

Potential energy surface



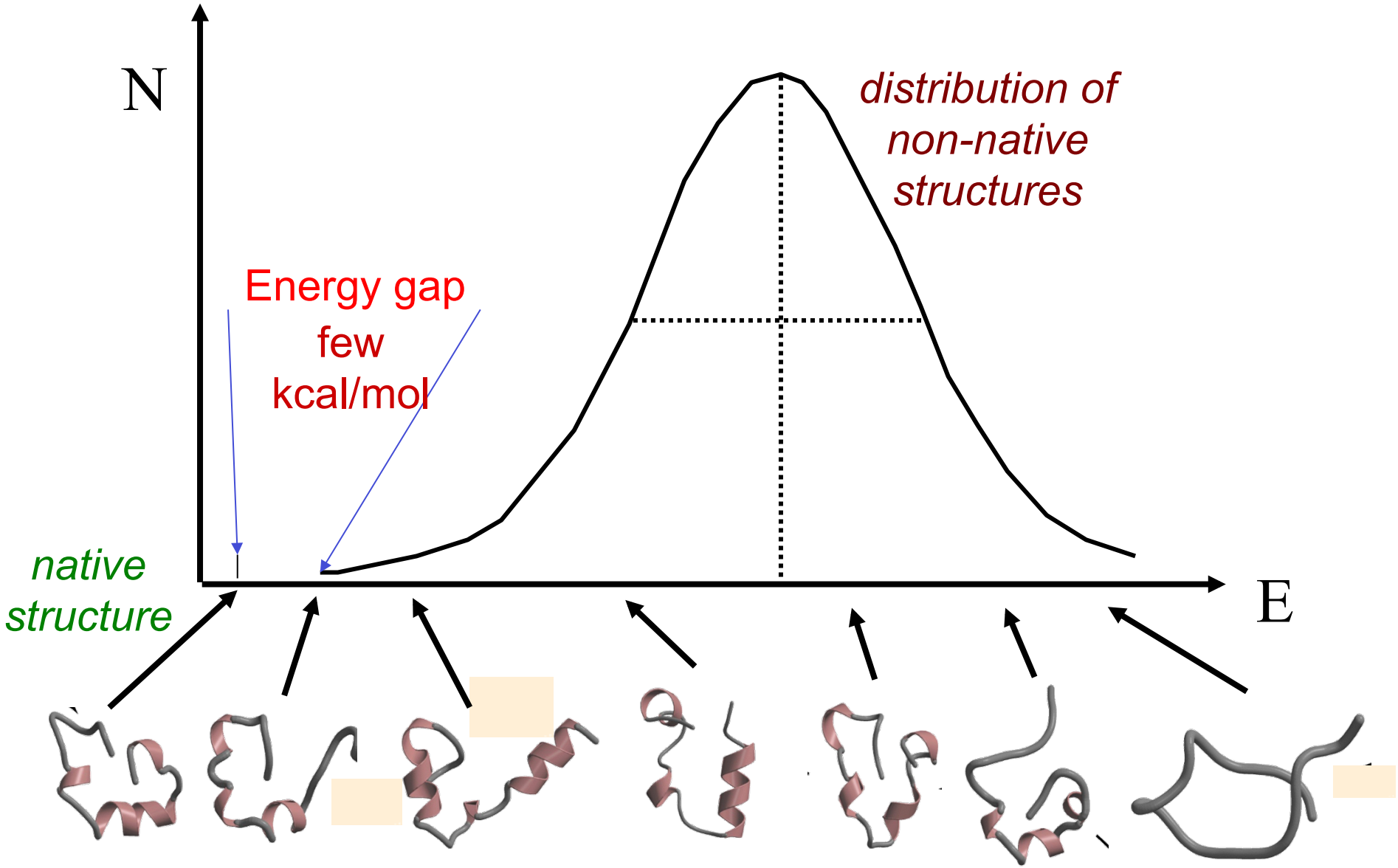
a cross-section of the energy landscape

A surface can be drawn that represents the variation of the potential energy as the conformation varies

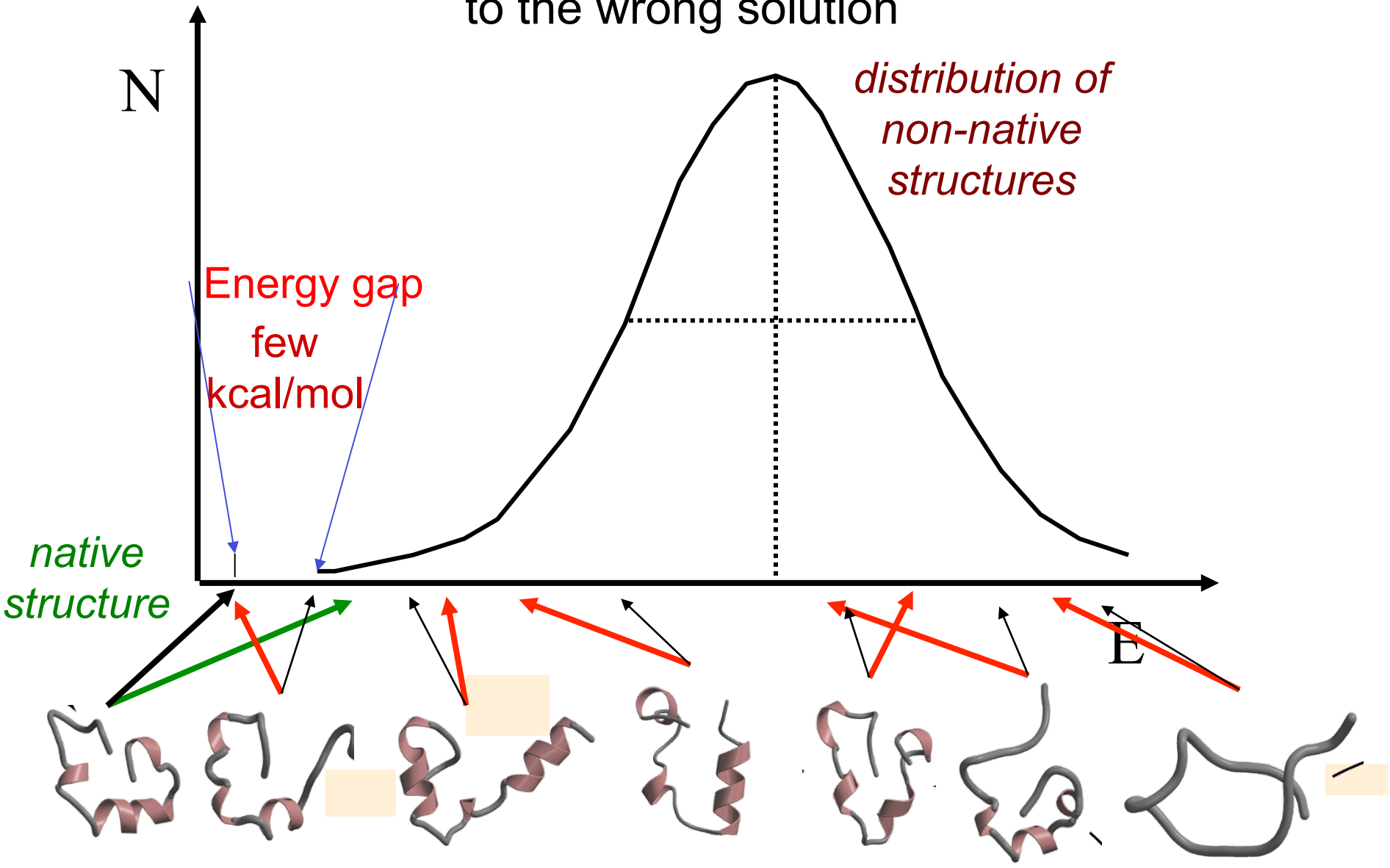
It resembles a landscape where valleys and peaks represent energy minima and maxima, respectively (the most frequently occuring conformations lie at the bottom of the deapest valleys)

Peaks are energy barriers, therefore reaching the global minimum is not trivial

The difference in energy between the native and non-native conformation can be very small



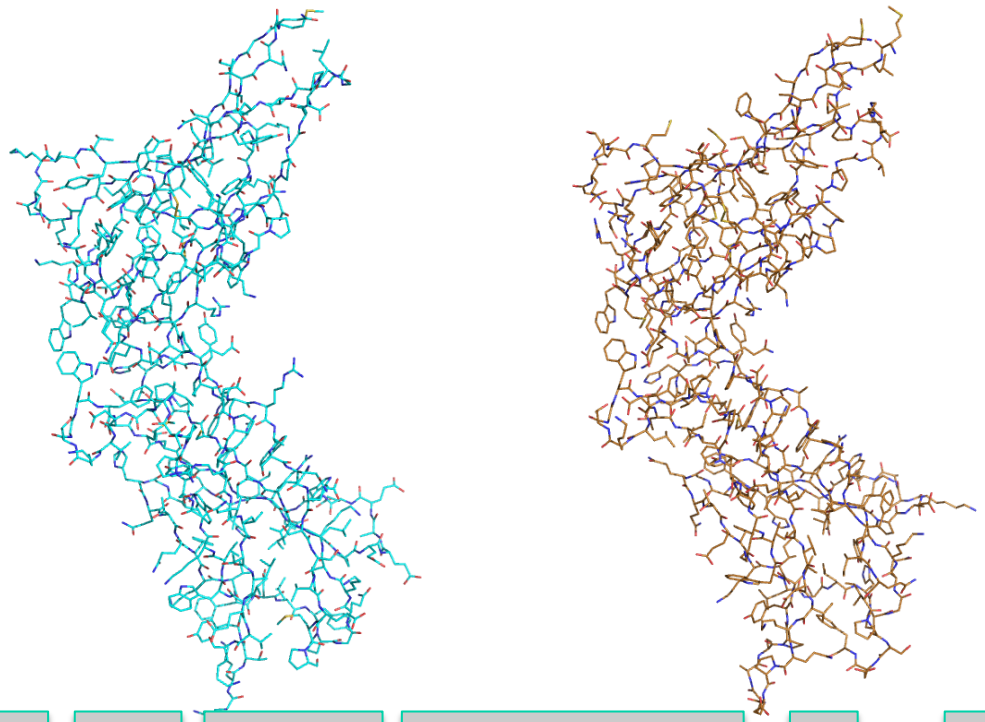
Even if we were able to generate all the possible conformations (and we aren't) a small error in the energy estimate would lead to the wrong solution



The problem of folding: a hand from evolution

SPOT THE DIFFERENCE

Example of homologous proteins



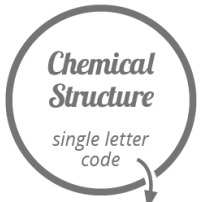
prolactin receptor

```
human   3    PPGKPEIFKCRSPNKETFTCWWRPGTDGGLPTNYSLTYHREGETLMHECPDYITGGPNSC   62
Rat     3    PPGKP                                          PDYKTSGPNSC   62

human   63   HFGKQ                                          EVKQPEDRKPY   122
Rat     63   FFSKQ                                          EVKQLKDKKTY   122

Human   123  LWIKWSPPTLIDLKTGWFTLLYEIRLKPEKAAEWEIHFAGQQTEFKILSLHPGQKYLVQV   182
rat     123  LWVKWSPPTITDVKTGWFTMEYEIRLKPEEAEEWEIHFTGHQTQFKVFDLYPGQKYLVQT   182

human   183  RCKPDHGYWSAWSPATFIQIPSDFTMND  210
rat     183  RCKPDHGYWSRWSQESSVEMPNDFTLKD  210

Identities = 151/208 (73%),
 Positives = 176/208 (85%)
```

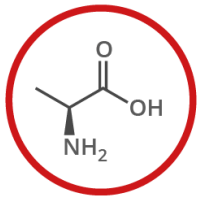
SPOT THE DIFFERENCE

AMINO ACIDS ARE THE BUILDING BLOCKS OF PROTEINS IN LIVING ORGANISMS. THERE ARE OVER 500 AMINO ACIDS FOUND IN NATURE - HOWEVER, THE HUMAN GENETIC COD ONLY DIRECTLY ENCODES 20. 'ESSENTIAL' AMINO ACIDS MUST BE OBTAINED FROM THE DIET, WHILST NON-ESSENTIAL AMINO ACIDS CAN BE SYNTHESISED IN THE BODY.

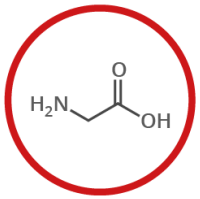
Chart Key: ALIPHATIC · AROMATIC · ACIDIC · BASIC · HYDROXYLIC · SULFUR-CONTAINING · AMIDIC · NON-ESSENTIAL · ESSENTIA



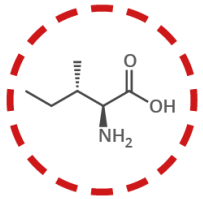
NAME (A)
three letter code
DNA codons



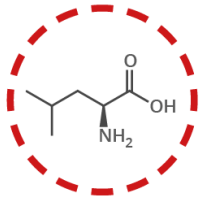
ALANINE (A)
Ala
GCT, GCC, GCA, GCG



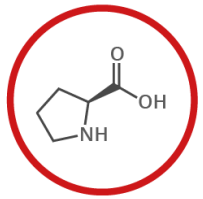
GLYCINE (G)
Gly
GGT, GGC, GGA, GGG



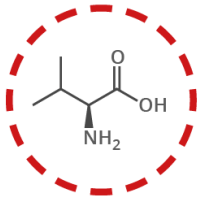
ISOLEUCINE (I)
Ile
ATT, ATC, ATA



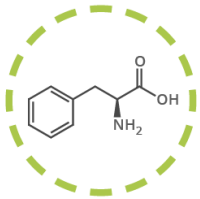
LEUCINE (L)
Leu
CTT, CTC, CTA, CTG, TTA, TTG



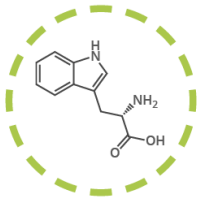
PROLINE (P)
Pro
CCT, CCC, CCA, CCG



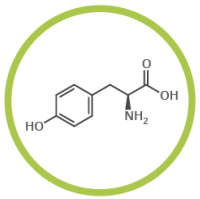
VALINE (V)
Val
GTT, GTC, GTA, GTG



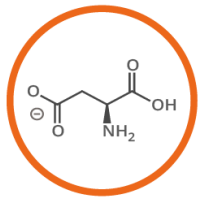
PHENYLALANINE (F)
Phe
TTT, TTC



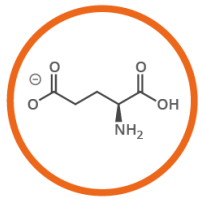
TRYPTOPHAN (W)
Trp
TGG



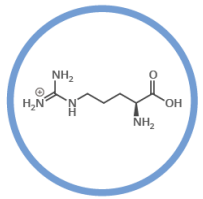
TYROSINE (Y)
Tyr
TAT, TAC



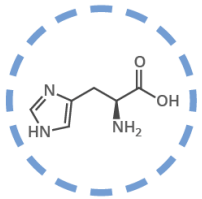
ASPARTIC ACID (D)
Asp
GAT, GAC



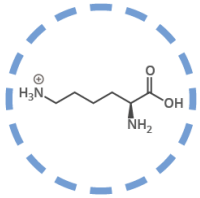
GLUTAMIC ACID (E)
Glu
GAA, GAG



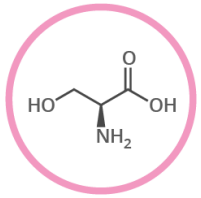
ARGININE (R)
Arg
CGT, CGC, CGA, CGG, AGA, AGG



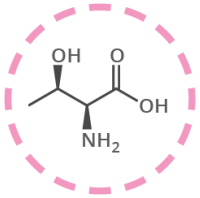
HISTIDINE (H)
His
CAT, CAC



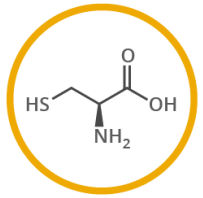
LYSINE (K)
Lys
AAA, AAG



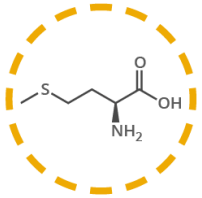
SERINE (S)
Ser
TCT, TCC, TCA, TCG, AGT, AGC



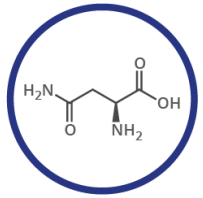
THREONINE (T)
Thr
ACT, ACC, ACA, ACG



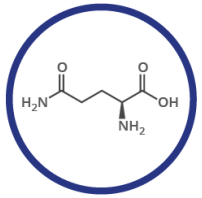
CYSTEINE (C)
Cys
TGT, TGC



METHIONINE (M)
Met
ATG



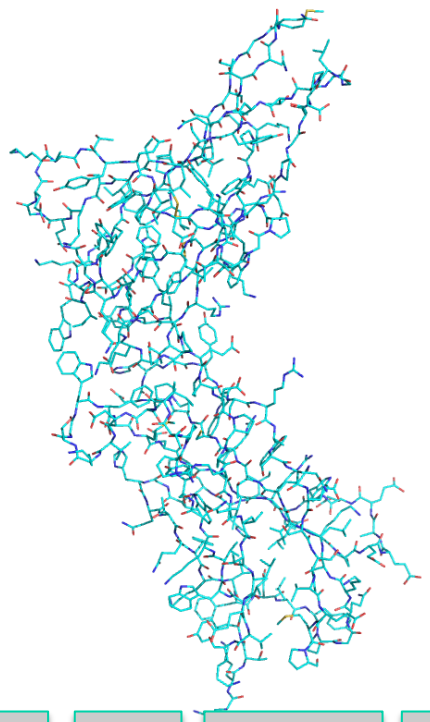
ASPARAGINE (N)
Asn
AAT, AAC

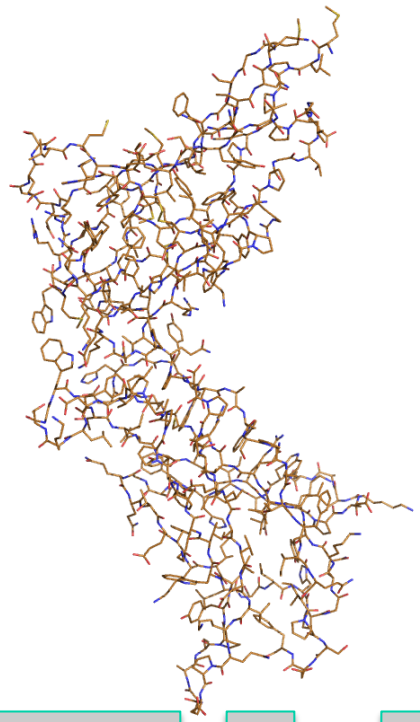


GLUTAMINE (Q)
Gln
CAA, CAG

Note: This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between asparagine/aspartic acid and glutamine/glutamic acid is difficult. In these cases, the codes asx (B) and glx (Z) are respectively used.

Example of homologous proteins

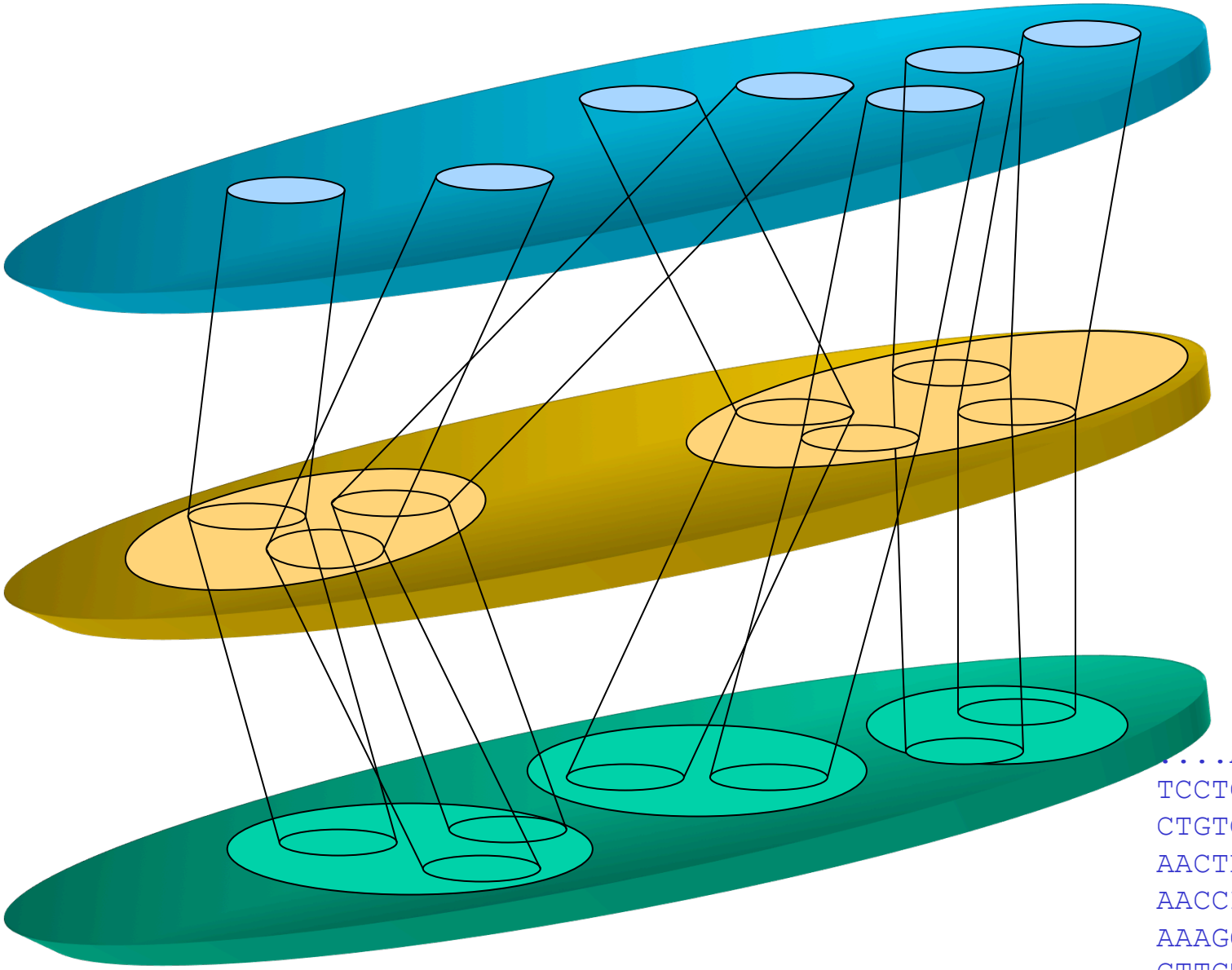




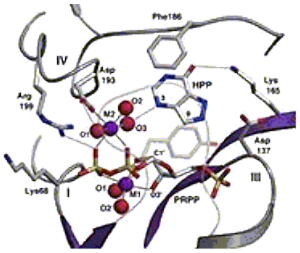
prolactin receptor

```
human   3    PPGKPEIFKCRSPNKETFTCWWRPGTDGGLPTNYSLTYHREGETLMHECPDYITGGPNSC   62
Rat     1                                                                  2

human                                                                     22
Rat                                                                       22

Human   123  LWIKWSPPTLIDLKTGWFTLLYEIRLKPEKAAEWEIHFAGQQTEFKILSLHPGQKYLVQV  182
rat     123  LWVKWSPPTITDVKTGWFTMEYEIRLKPEEAEEWEIHFTGHQTQFKVFDLYPGQKYLVQT  182

human   183  RCKPDHGYWSAWSPATFIQIPSDFTMND  210
rat     183  RCKPDHGYWSRWSQESSVEMPNDFTLKD  210

Identities = 151/208 (73%),
Positives = 176/208 (85%)
```

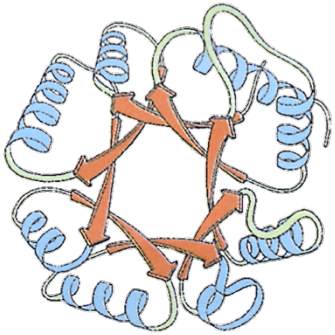
This is the basis of the "comparative modeling" technique



function



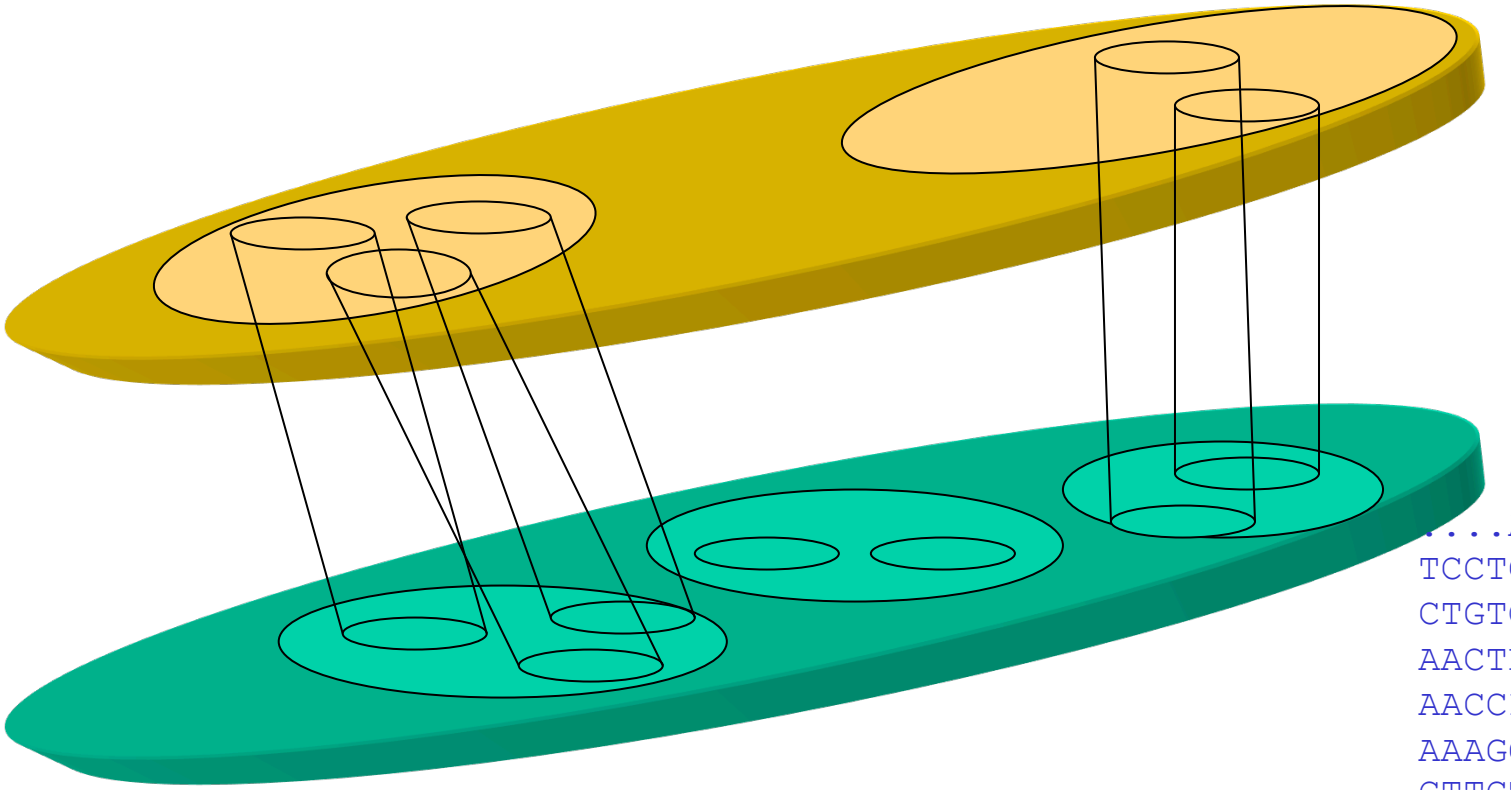
structure



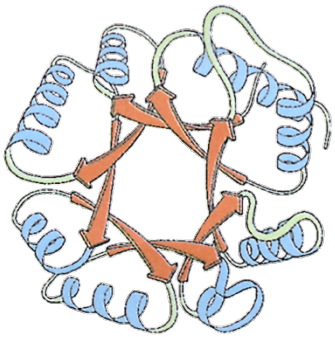
sequence

```
....ACCEAA GAAGTCAGLV
TCCTCGADGC TGAAGQRNDF
CTGTQQGCTC LIAVRCCTAC
AACTRGACAA GTDFHAATGC
AACCFHTTTG CTILMGAGGA
AAAGGAGTLI GCNDEGAGGR
CTTCTGAGFR CGGCAAFHRL
CCATTQFCTG ACAGDYGHTW
TTTACYWACT TGCFTRNQKI
CCTCTGAWHY ...
```

Similarity in sequence implies similarity in structure!



structure



sequence

```
....ACCEAA GAAGTCAGLV
TCCTCGADGC TGAAGQRNDF
CTGTQQGCTC LIAVRCCTAC
AACTRGACAA GTDFHAATGC
AACCFHTTTG CTILMGAGGA
AAAGGAGTLI GCNDEGAGGR
CTTCTGAGFR CGGCAAFHRL
CCATTQFCTG ACAGDYGHTW
TTTACYWACT TGCFTRNQKI
CCTCTGAWHY ...
```

Homology (or Comparative) Modelling

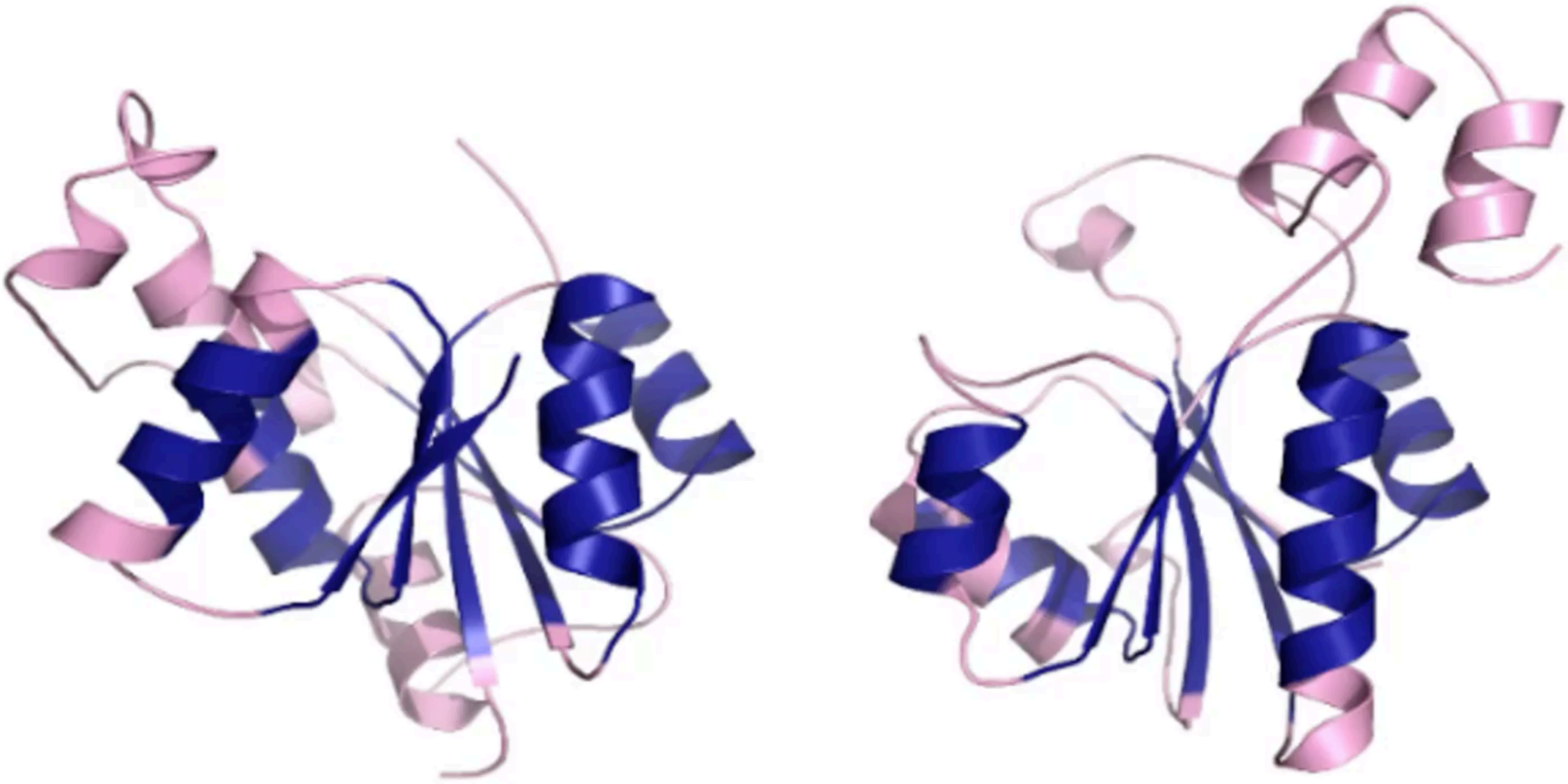
that is

How to build a molecular model by homology

The **tertiary structure** of ***homologous*** proteins is better preserved than their **primary structure (sequence)**

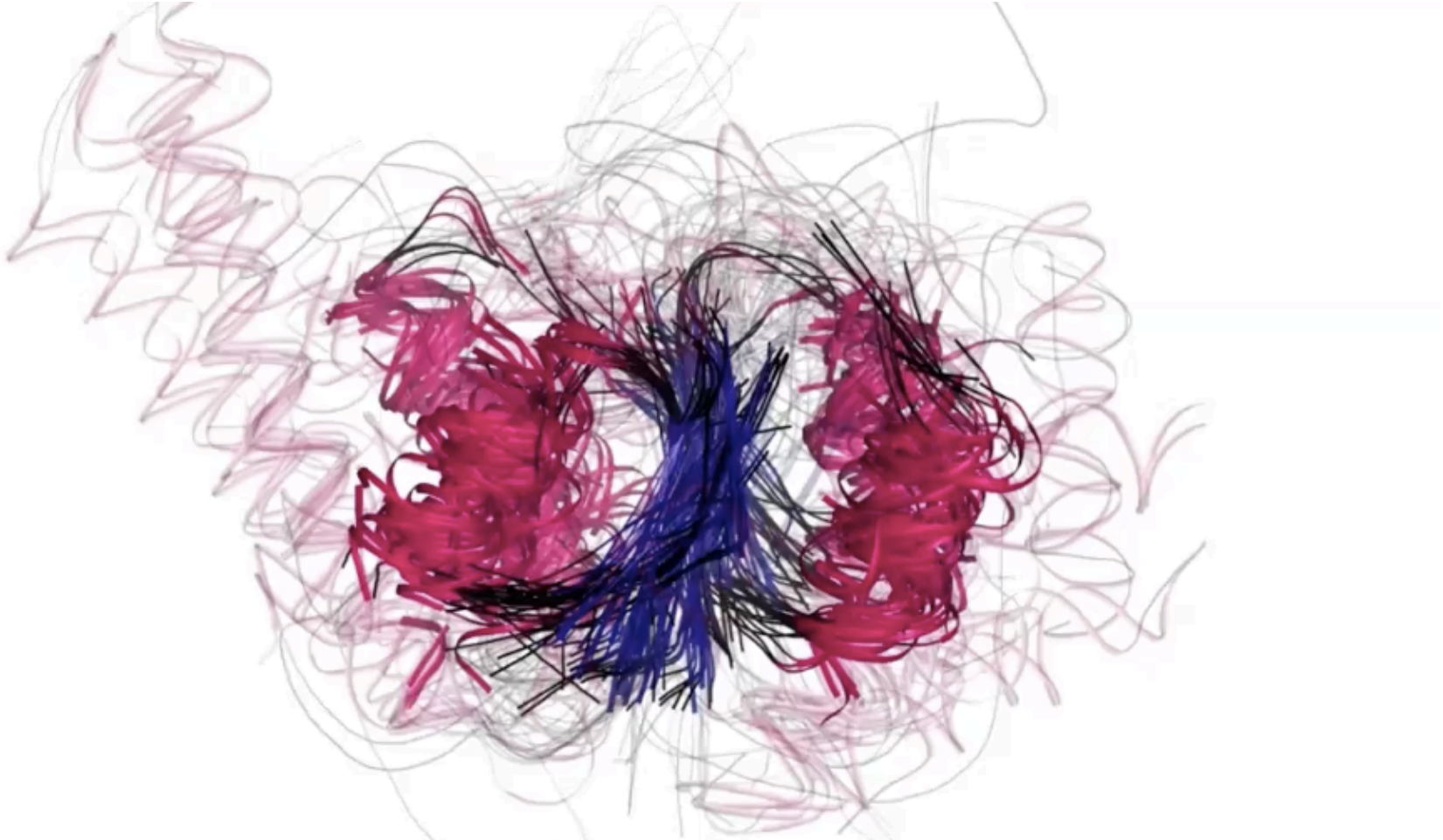
Homology Modelling

Let’s define the protein core



“Core”: the structural conserved region

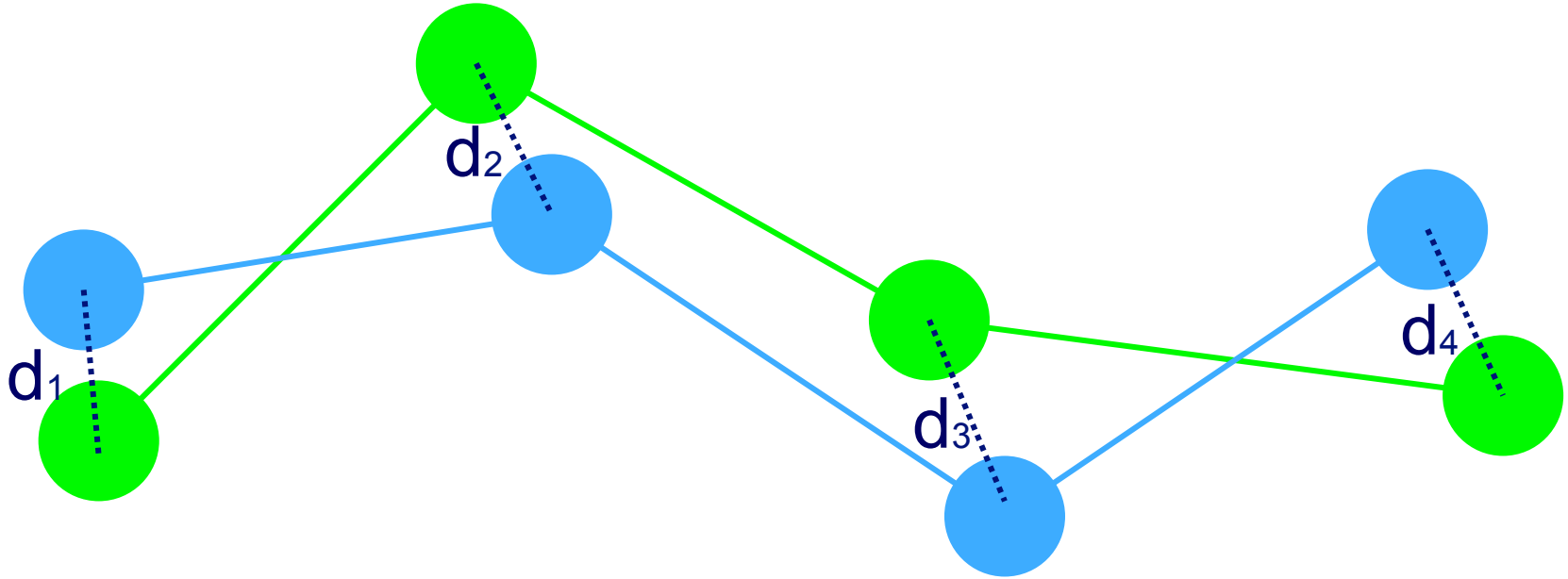
Peripheral regions (indels) or “structural divergent regions” (SDR)



(Even extremely divergent) relatives in superfamilies contain well-conserved structural cores

CATHEDRAL: PLoS Comp. Biol. (2007), *FLORA: PLoS Comp Biol. (2007), GRATH: J. Mol. Biol (2005), SSAP: J. Mol Biol (1989)*

Root Mean Square Deviation (RMSD)



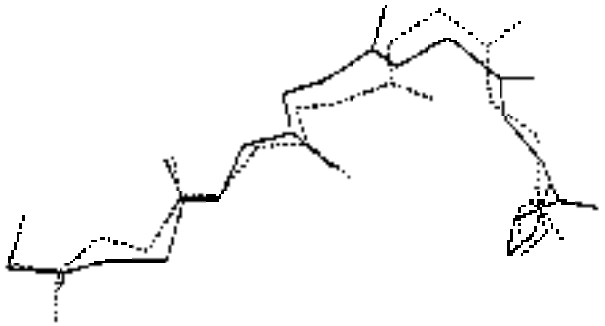
N = number of corresponding atoms

$$RMSD_{N-1} = \sqrt{\frac{\sum_{i=1}^{N}(d_i)^2}{N-1}}$$

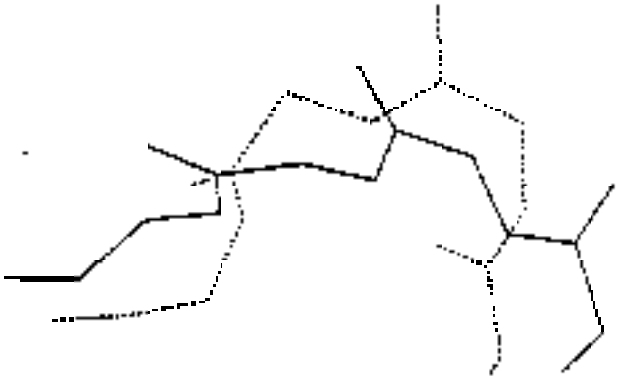
It is a similarity measure for protein 3D structures

RMSD = 0.5 Å

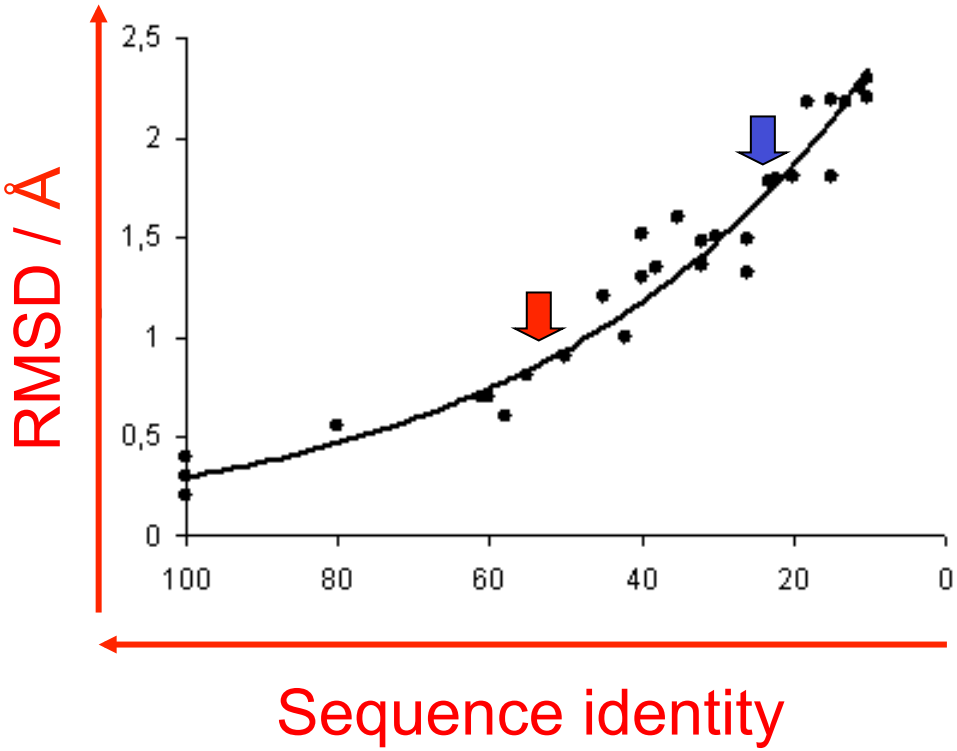
RMSD = 1.2 Å



RMSD = 1.8 Å



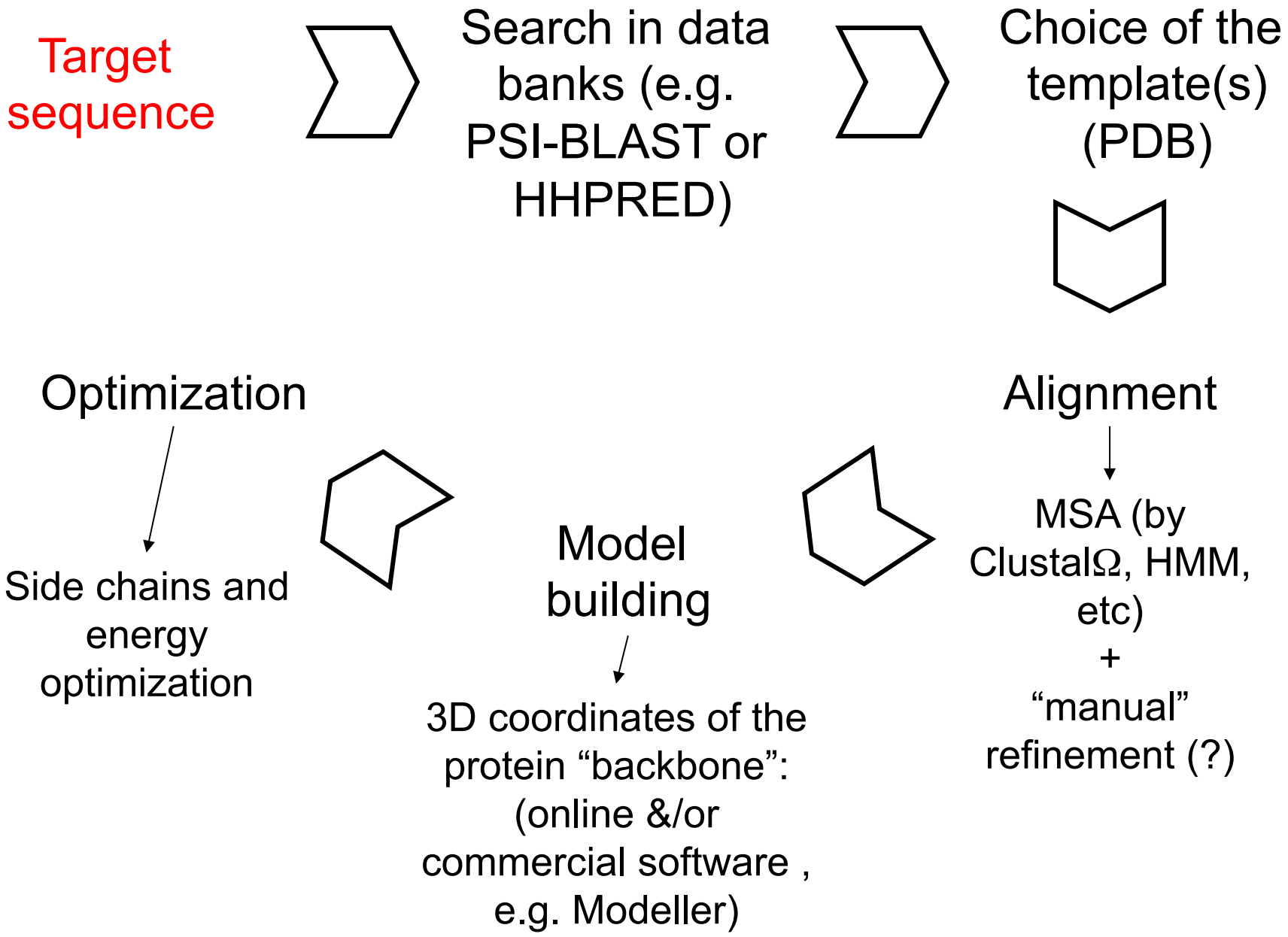
A (non biunivocal) relationship exists between the sequence identity and the structural similarity



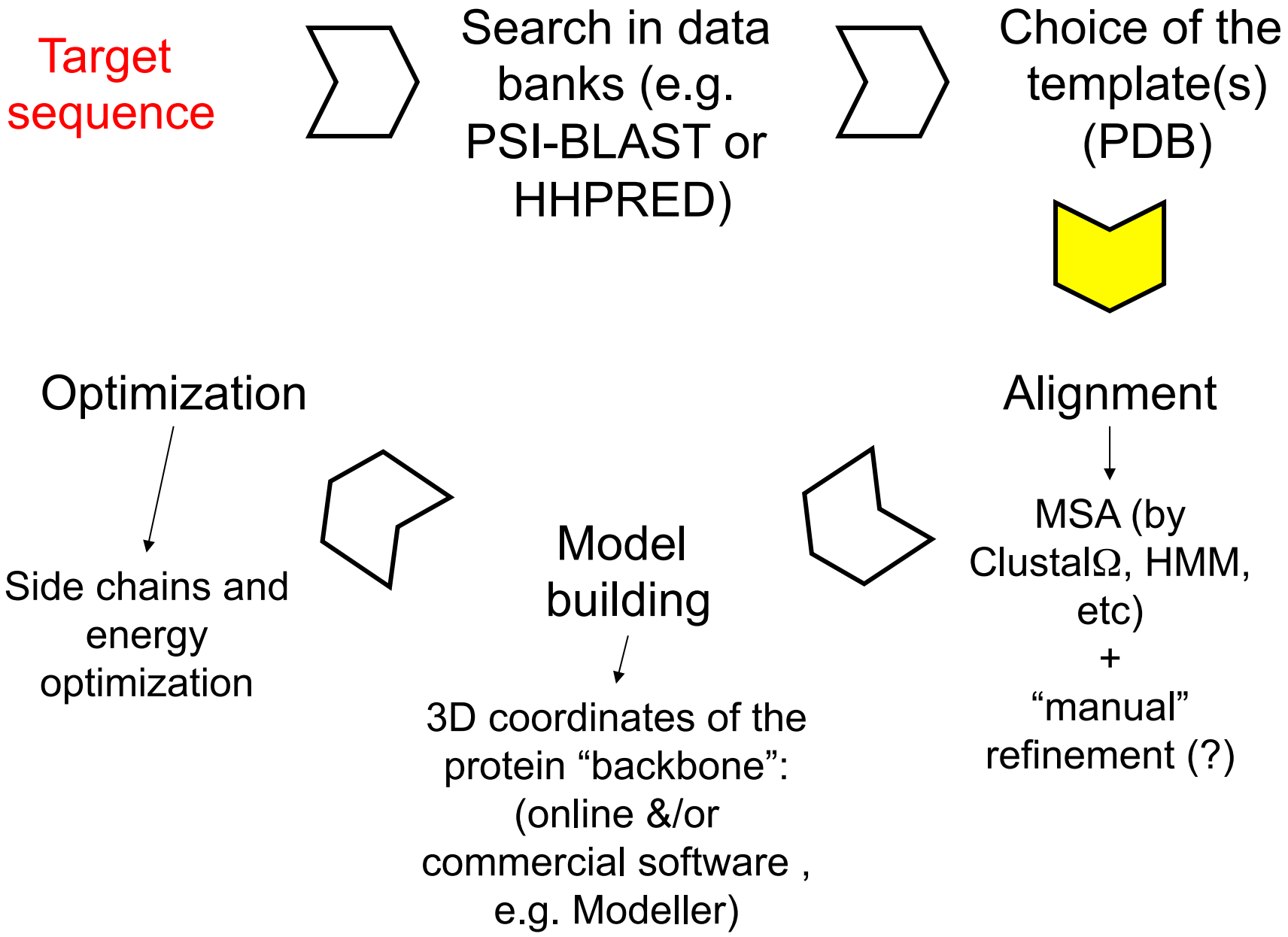
- Seq. Id. < 20%: "core" takes **~50%** of the structure & r.m.s.d. of the backbone around 1.8 Å
- Seq. Id. > 50%: "core" takes **~ 90%** of the structure & r.m.s.d. of the backbone around 1.0 Å

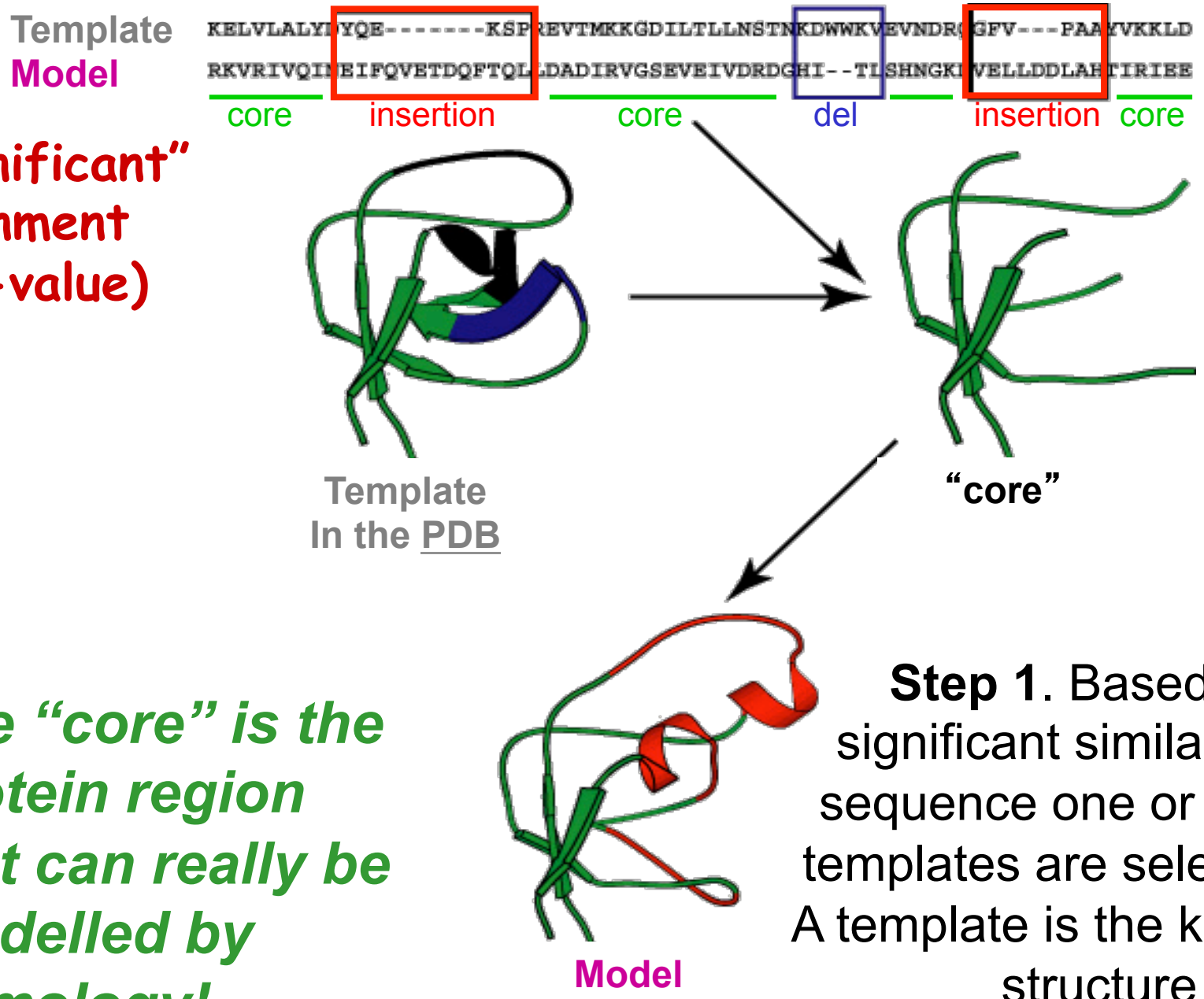
Chothia & Lesk, *EMBO Journal* (1986) vol 5, pag. 823-826
derived from 32 pairs of homologous proteins

Comparative modelling



Comparative modelling

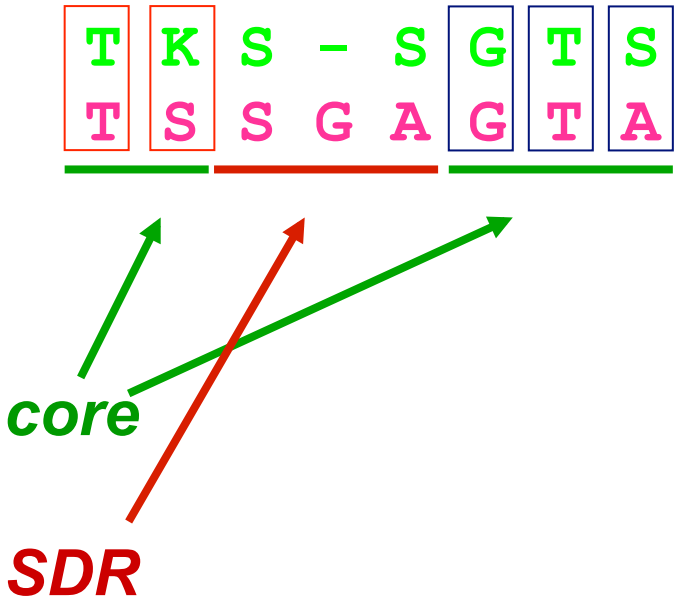


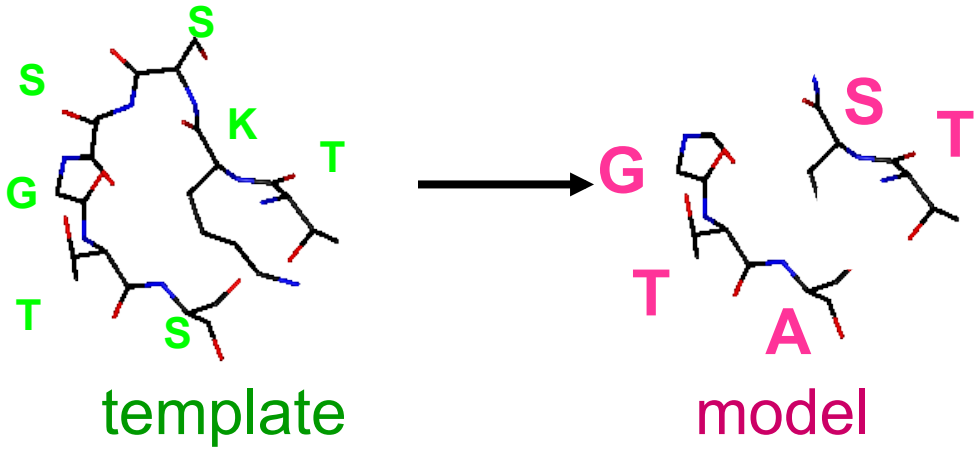

Template
Model
core
insertion
core
del
insertion
core
"Significant" alignment (E-value)
Template In the PDB
"core"
Model
The "core" is the protein region that can really be modelled by homology!
Step 1. Based on a significant similarity in sequence one or more templates are selected. A template is the known structure of an homologous protein

Comparative modelling

Step 2. Transfer of the ***core*** backbone coordinates from the template to the model

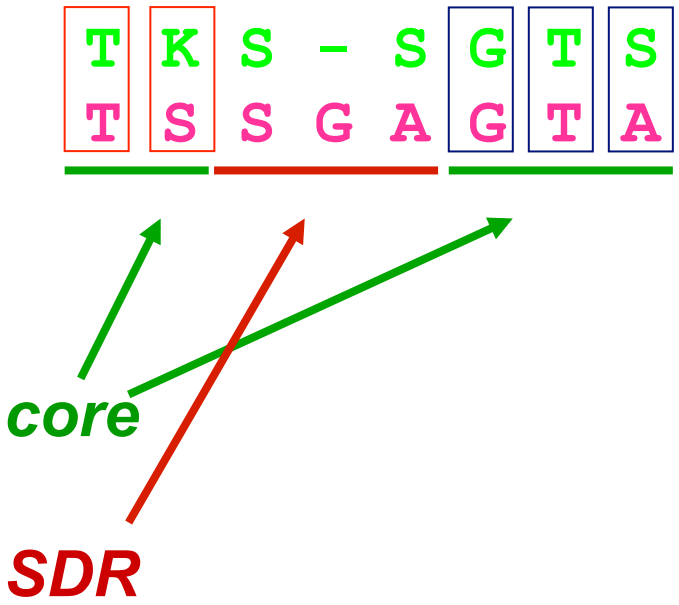
- Substitution of the **side chains**

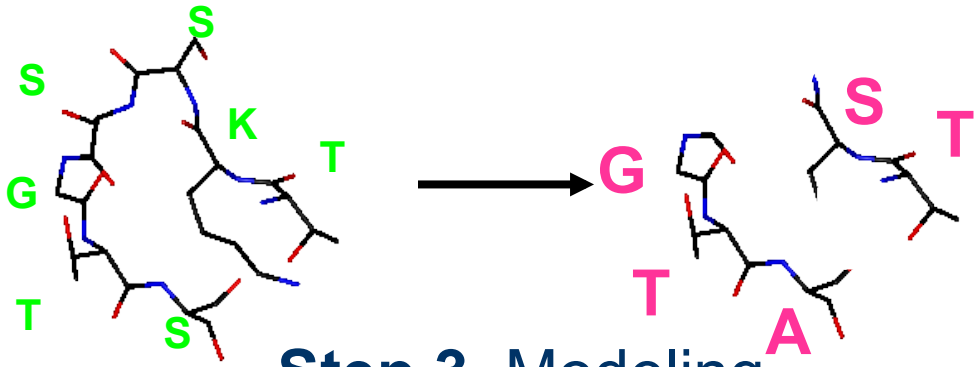




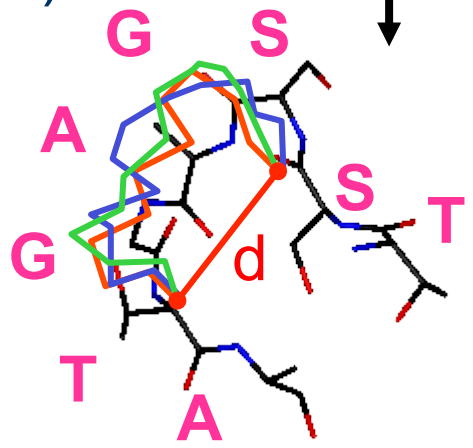
Remember that the sequence alignment is our hypothesis of evolutionary correspondence between the target and template proteins!

Comparative modelling

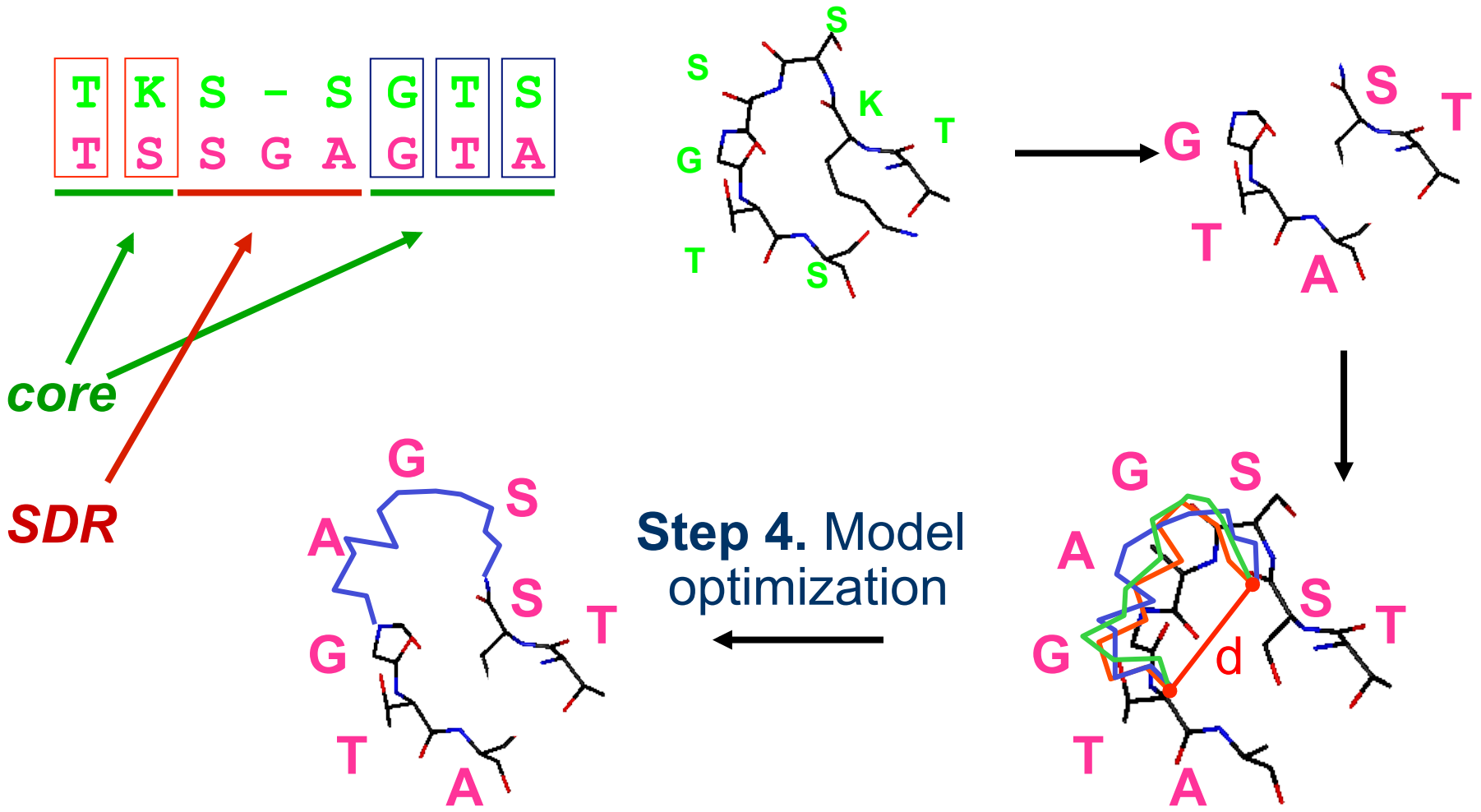




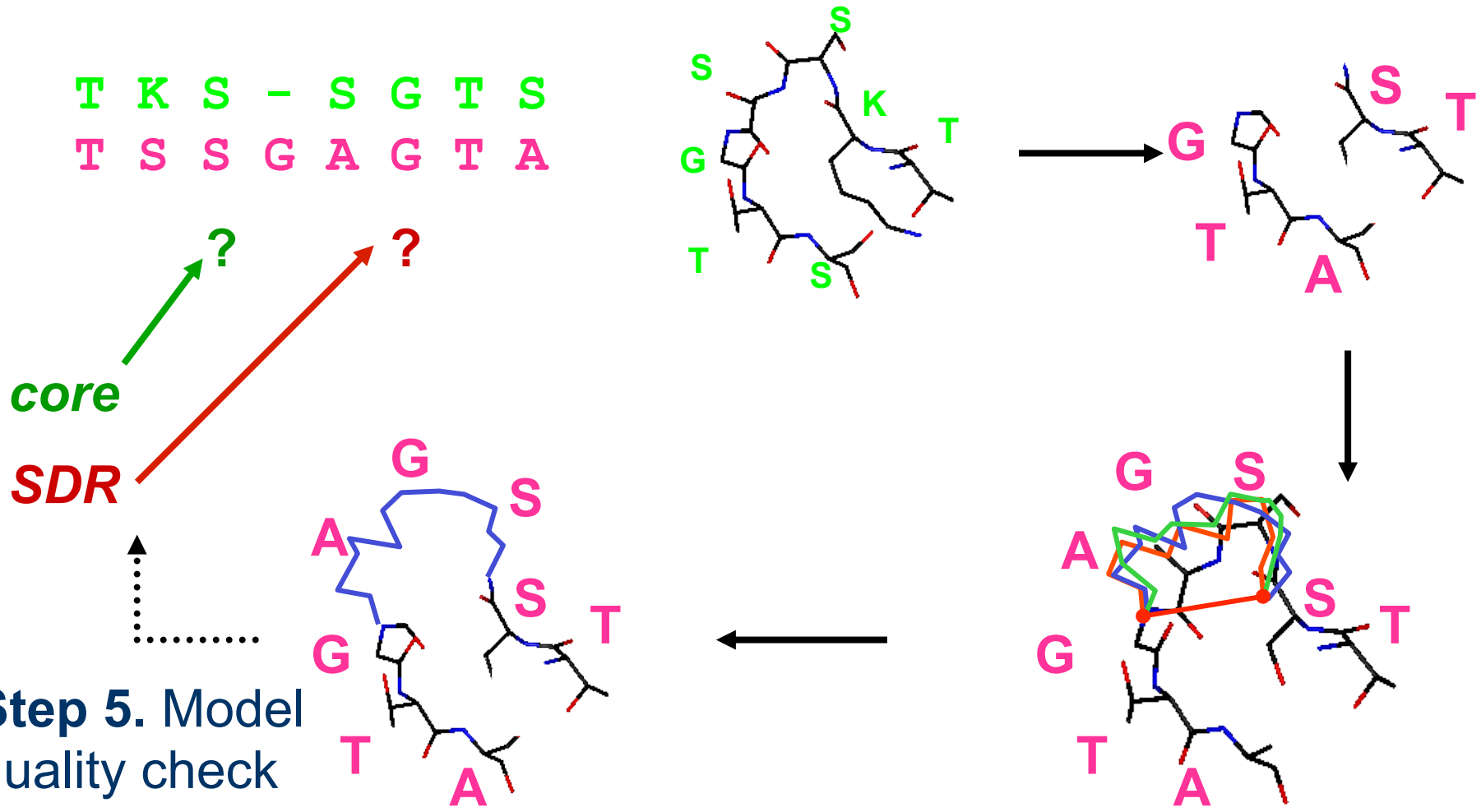
Step 3. Modeling divergent regions (**loops**)



Comparative modelling



Comparative modelling



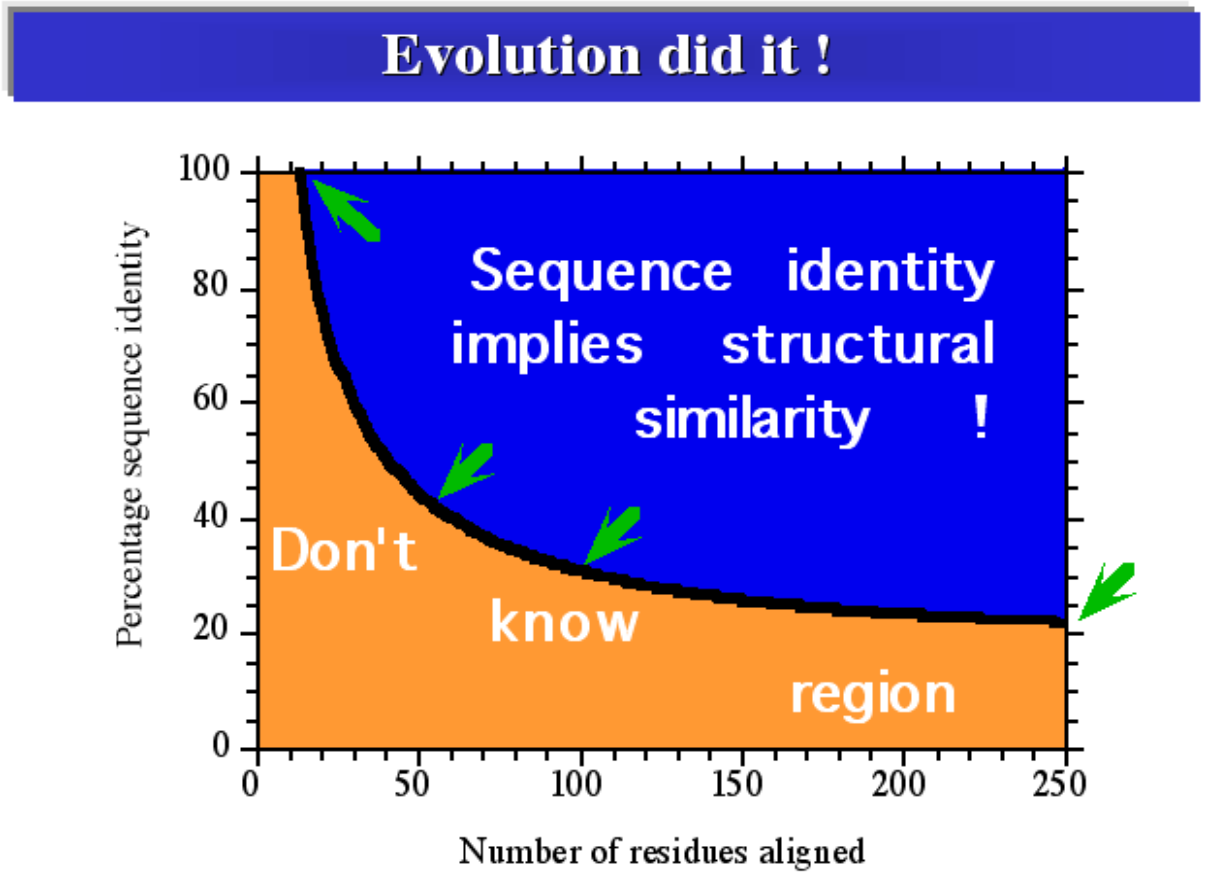
Step 5. Model quality check

The sequence alignment between template & target represents our hypothesis on the structural/functional corrispondence between amino acids of the two sequences

The key step in building a model by homology consists in obtaining a good sequence alignment!

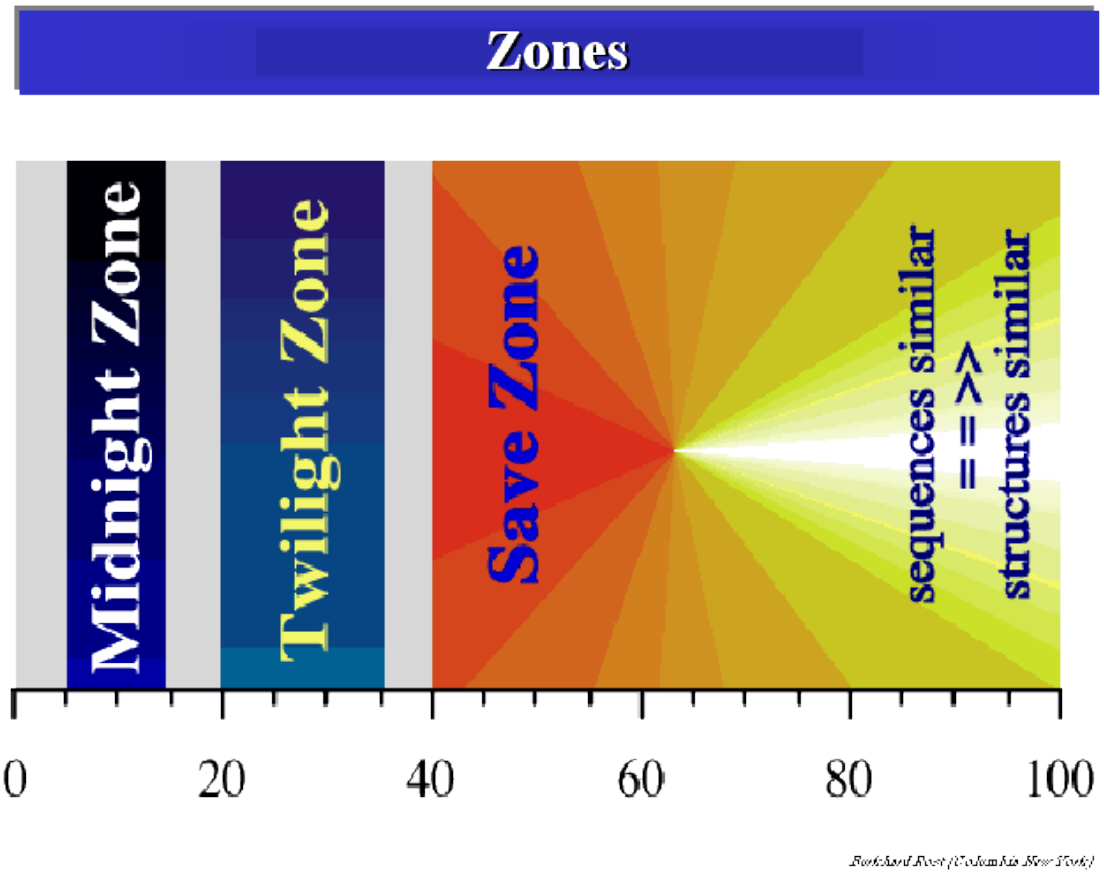
The sequence identity with the template correlates with the extension of the protein core and affects the alignment quality (further improvable through MSAs and the structure analysis)

When Homology Modelling can trusthworthly be applied ?



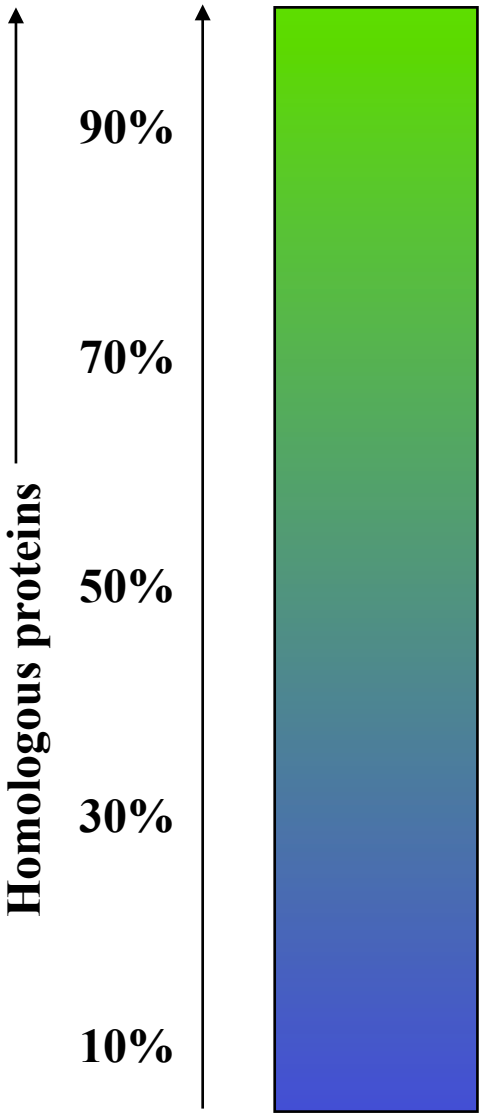
Although… alignments of HMM profiles have challenged this “rule of thumb”

When Homology Modelling can trusthworthly be applied ?



Although… alignments of HMM profiles have challenged this “rule of thumb”

Rate of model confidence



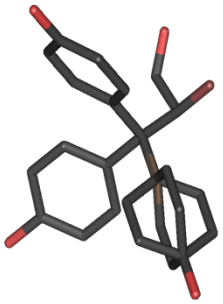
Main chain of the "core" very similar, extended common "core"

Main chain of the "core" somewhat similar, limited common "core"

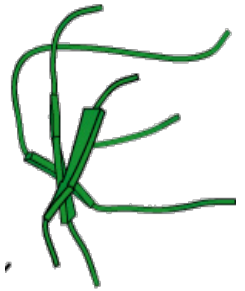
Same topology

Sequence identity

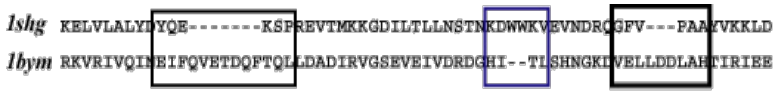
Drug design?



Biochemistry?



Molecular biology?

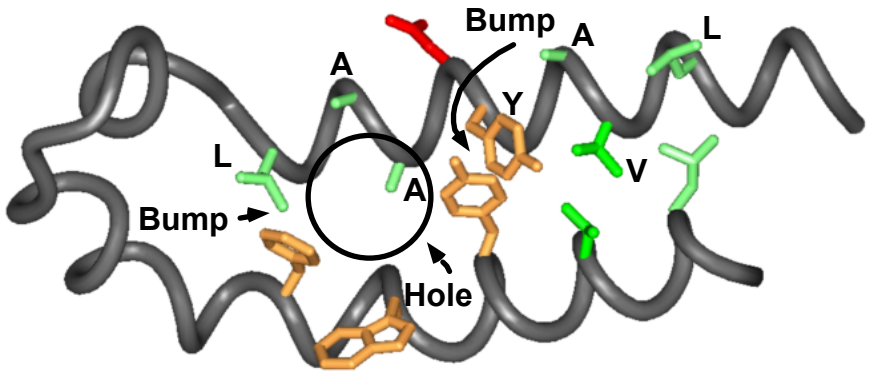


Comparative (homology) modelling

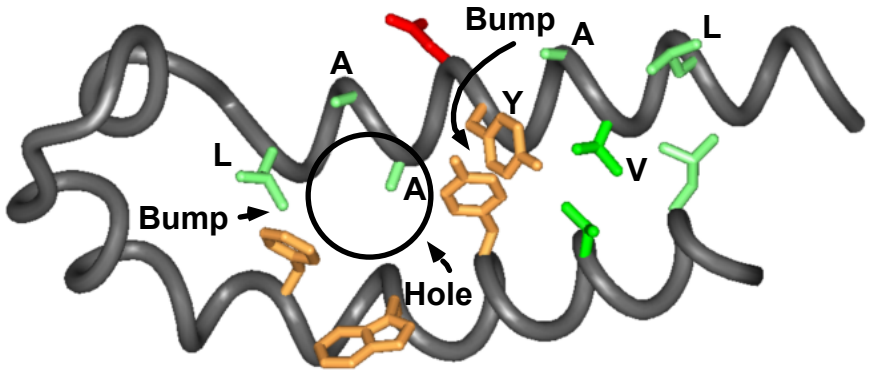
Model optimization:

- Serious errors cannot be corrected
  - We can eliminate sterically unfavorable interactions
  - We can optimize the backbone geometry in particular regions (e.g., around loops)

Comparative (homology) modelling

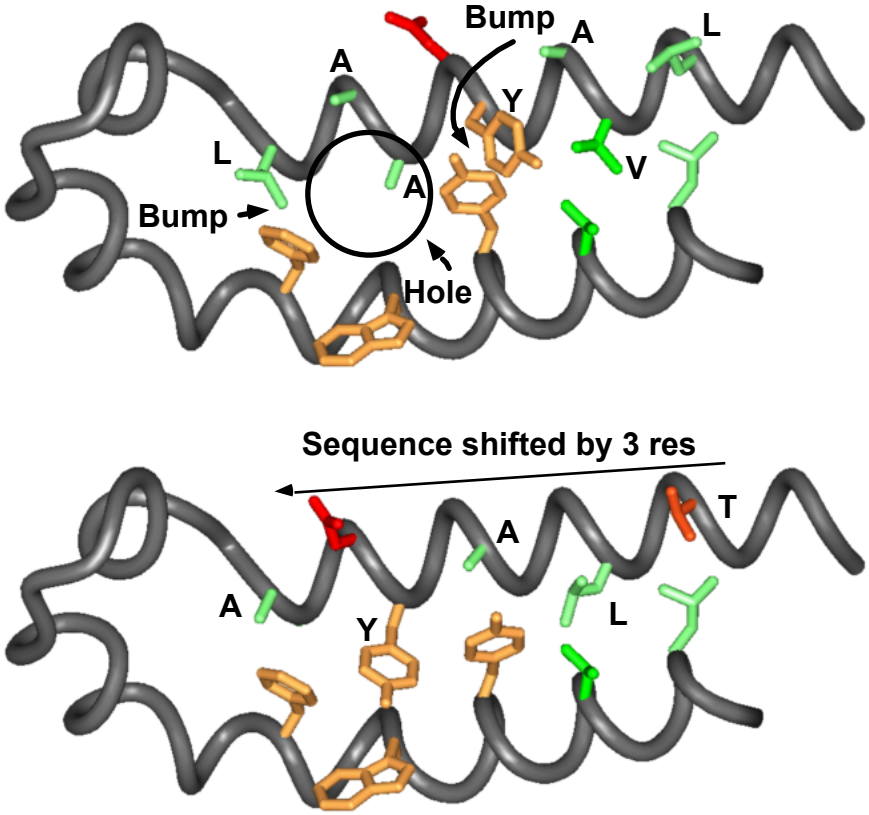


Comparative (homology) modelling



serious error in the protein packing

Comparative (homology) modelling

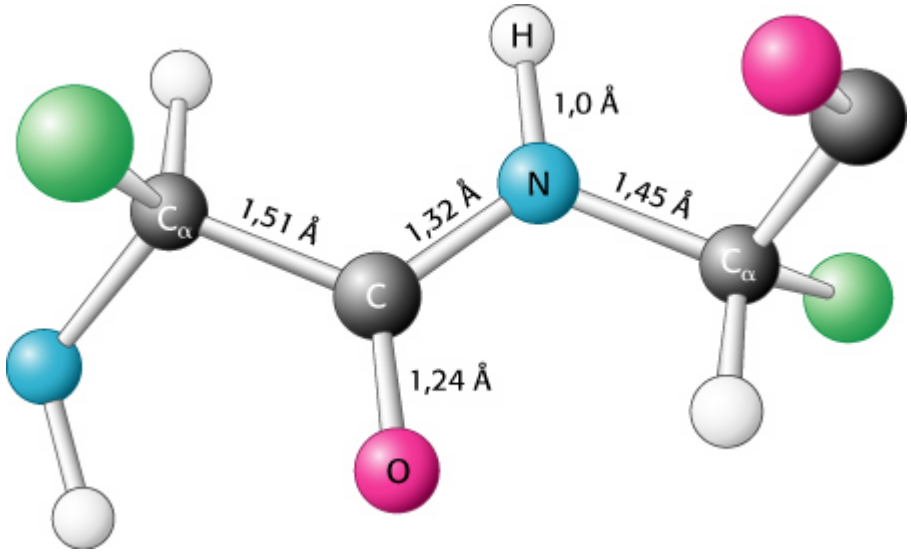


serious error in the protein packing

new alignment!

=> better packing of side chains

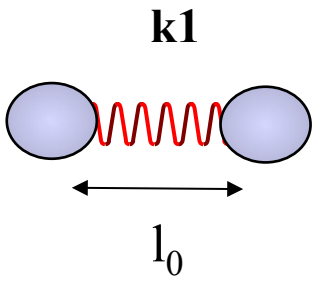
Energy optimization



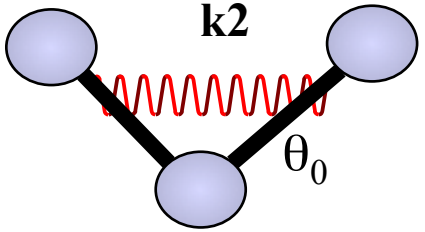
Molecular mechanics (MM) approximates atoms as rigid spheres subjected to forces

MM force fields

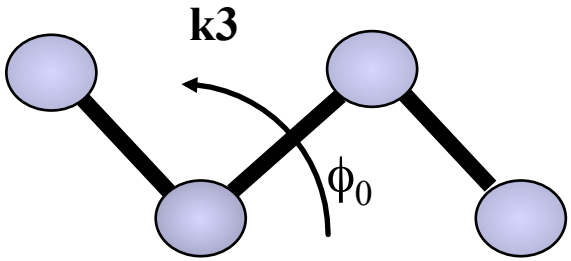
$$E_{total} =$$



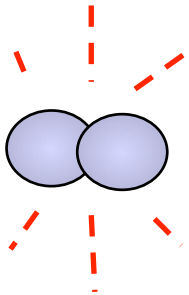
$$+$$



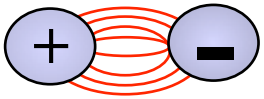
$$+$$



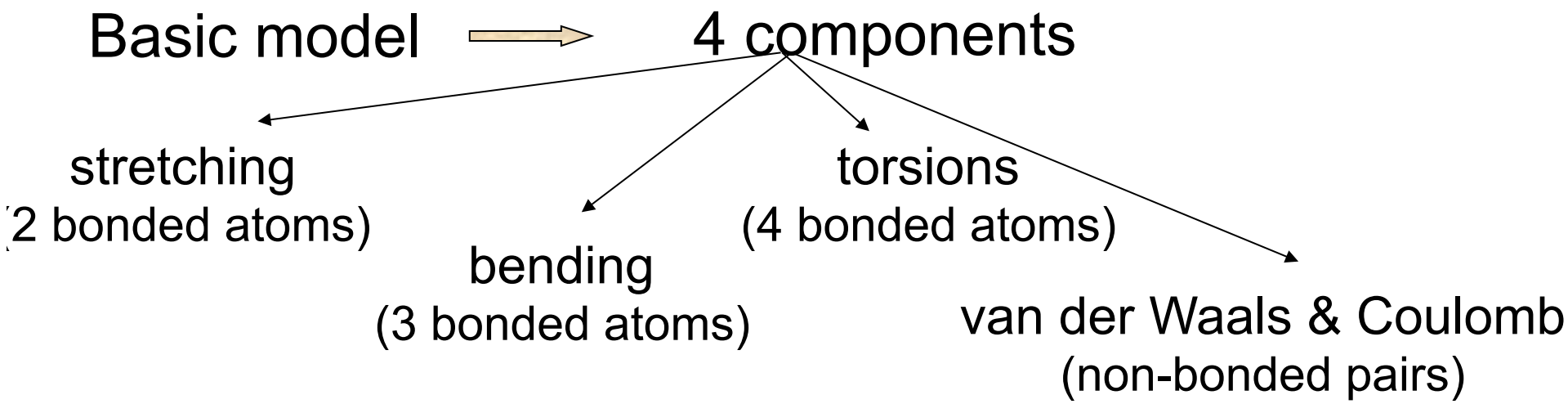
$$+$$



$$\alpha \frac{1}{r^9} - \frac{1}{r^6} \quad +$$



$$\frac{q_i q_j}{\varepsilon r}$$



Typical form:

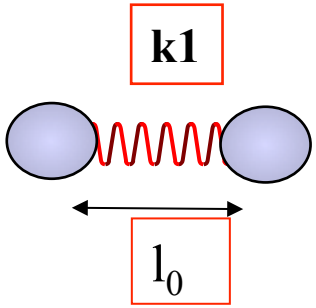
$$V(\mathbf{r}^N) = \sum_{bonds} \frac{k_i}{2}\left(l_i - l_{i,0}\right)^2 + \sum_{angles} \frac{k_i'}{2}\left(\theta_i - \theta_{i,0}\right)^2$$

$$+ \sum_{torsions} \frac{V_i}{2}\left[1 + \cos\left(i\,\omega - \gamma\right)\right] +$$

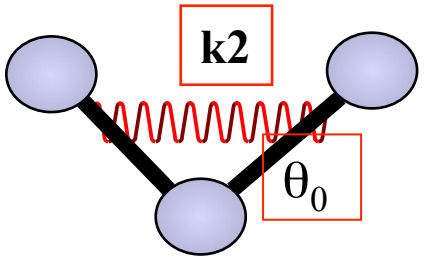
$$+ \sum_{i,j>i} \left\{ 4\varepsilon_i \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^6 - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \right] + \frac{q_i q_j}{4\pi\varepsilon\, r_{ij}} \right\}$$

MM force fields

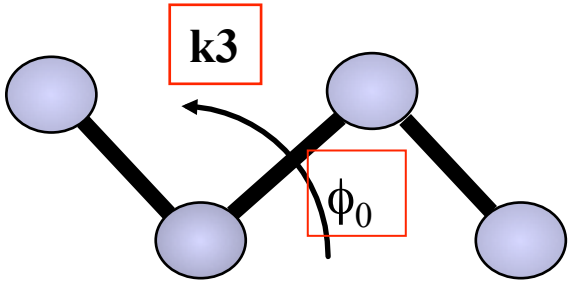
$$E_{total} =$$



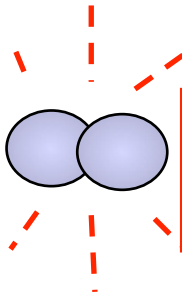
$+$



$+$

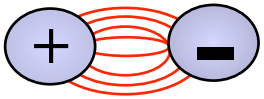


$+$

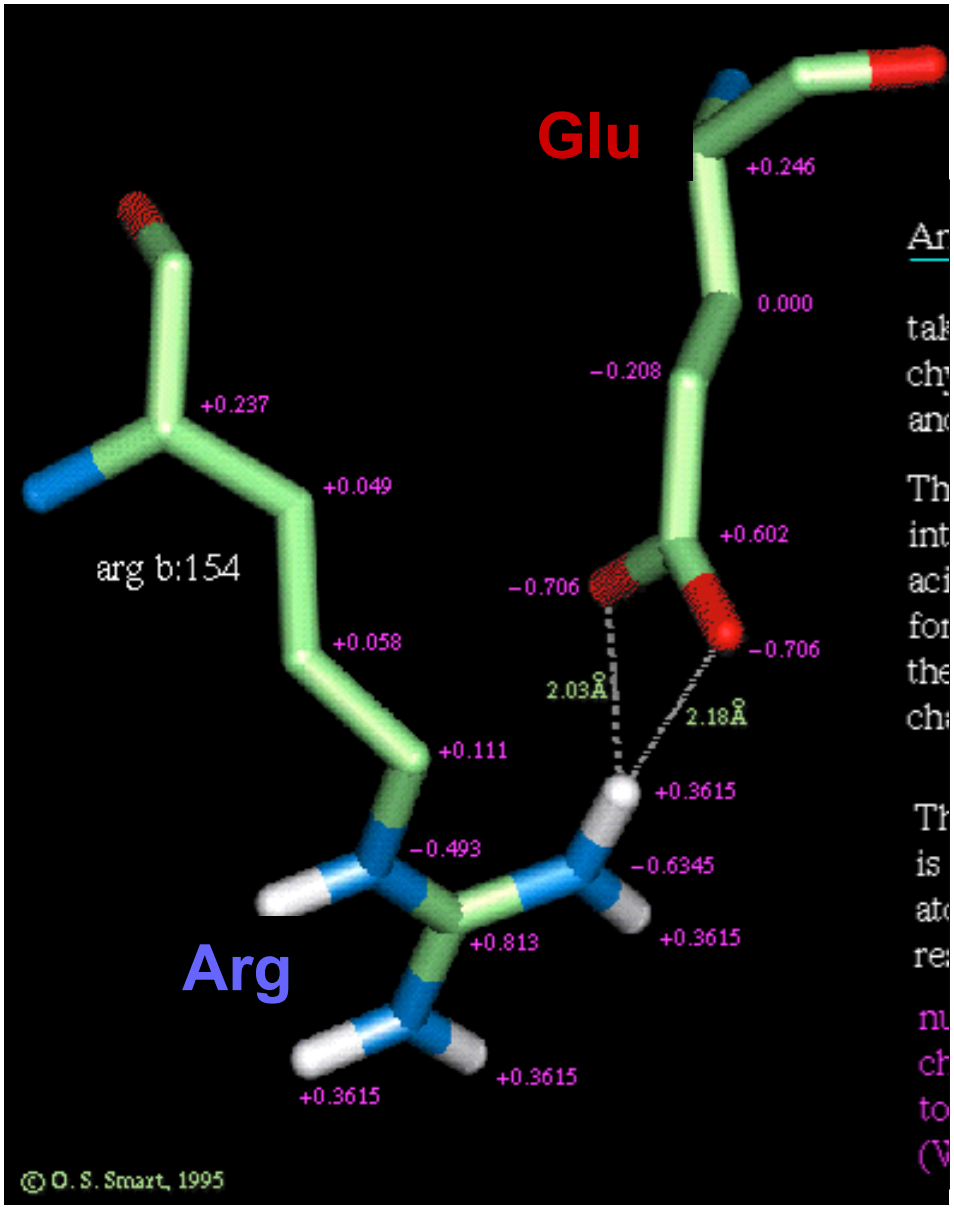


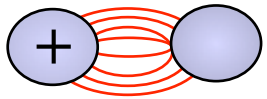
$$\alpha \frac{1}{r^9} - \frac{1}{r^6}$$

$+$



$$\frac{q_i q_j}{\varepsilon r}$$





$$\frac{q_i q_j}{\varepsilon r}$$

Example of Coulomb's interaction

Evaluating the model quality

”How much I can trust?"

In recognizing 'misfolded’ models:

EM (energy minimization) & **MD** (molecular dynamics) are **inefficient**:

Normality indices can be more efficient:

e.g. regularity on the geometry, atomic contacts, distribution of polar and apolar residues, solvation potentials, etc., derived from the observation od deposited (experimental) 3D structures

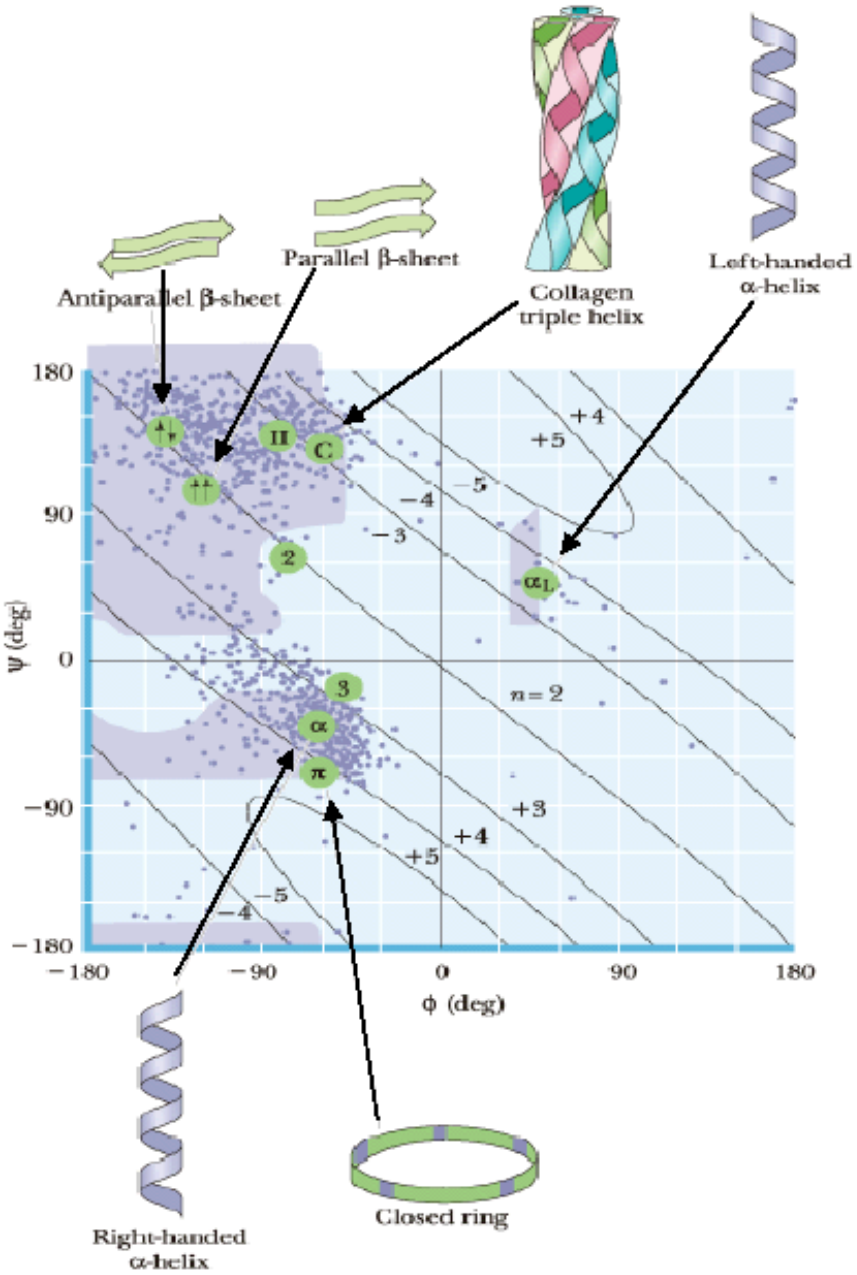
Model quality assessment

Estimating the quality of protein structure models is a vital step

We may have a set of alternative models (e.g. from different modeling servers or based on alternative template structures and alignments) from which the best candidate shall be selected, or a singe model has been built from which we need to predict the quality in order to have an idea about its suitability for subsequent experiments

Traditionally, geometrical checks were performed on the models (see Ramachandran plots), however "**statistical potentials**" (also named "**knowledge-based potentials**") such as the packing quality or the inside/outside profile, based on the comparison with experimental structures, have been shown to be much more effective

Ramachandran plot



A traditional (and obsolete) model quality check

Model quality assessment

"**statistical (or** "knowledge-based) **potentials**"

Statistical potentials or knowledge-based potentials are scoring functions derived from the analysis of known protein structures in the Protein Data Bank (PDB)
The most frequent something is, the most stable it is

$$u_{ij}(r) = -k_B T \ln[g_{ij}(r)]$$

$$\text{energy score} = \sum_{\text{atom pair}} u_{ij}(r)$$

K_B is the Boltzmann constant
T is the temperature
i and ***j*** are two atoms
g_{ij}(r) is the pair distribution function (derived from a set of structure-known proteins)

Model quality assessment: QMEAN

Nowadays, one of the the most used/cited protein model quality check estimators is QMEAN (Qualitative Model Energy ANalysis, https://swissmodel.expasy.org/qmean/)

The QMEAN server provides access to three scoring functions for the quality estimation of protein structure models, which allow to:
i) rank a set of models and ii) identify potentially unreliable regions in them

Both single models and set of models can be analysed.

Model quality assessment: QMEAN

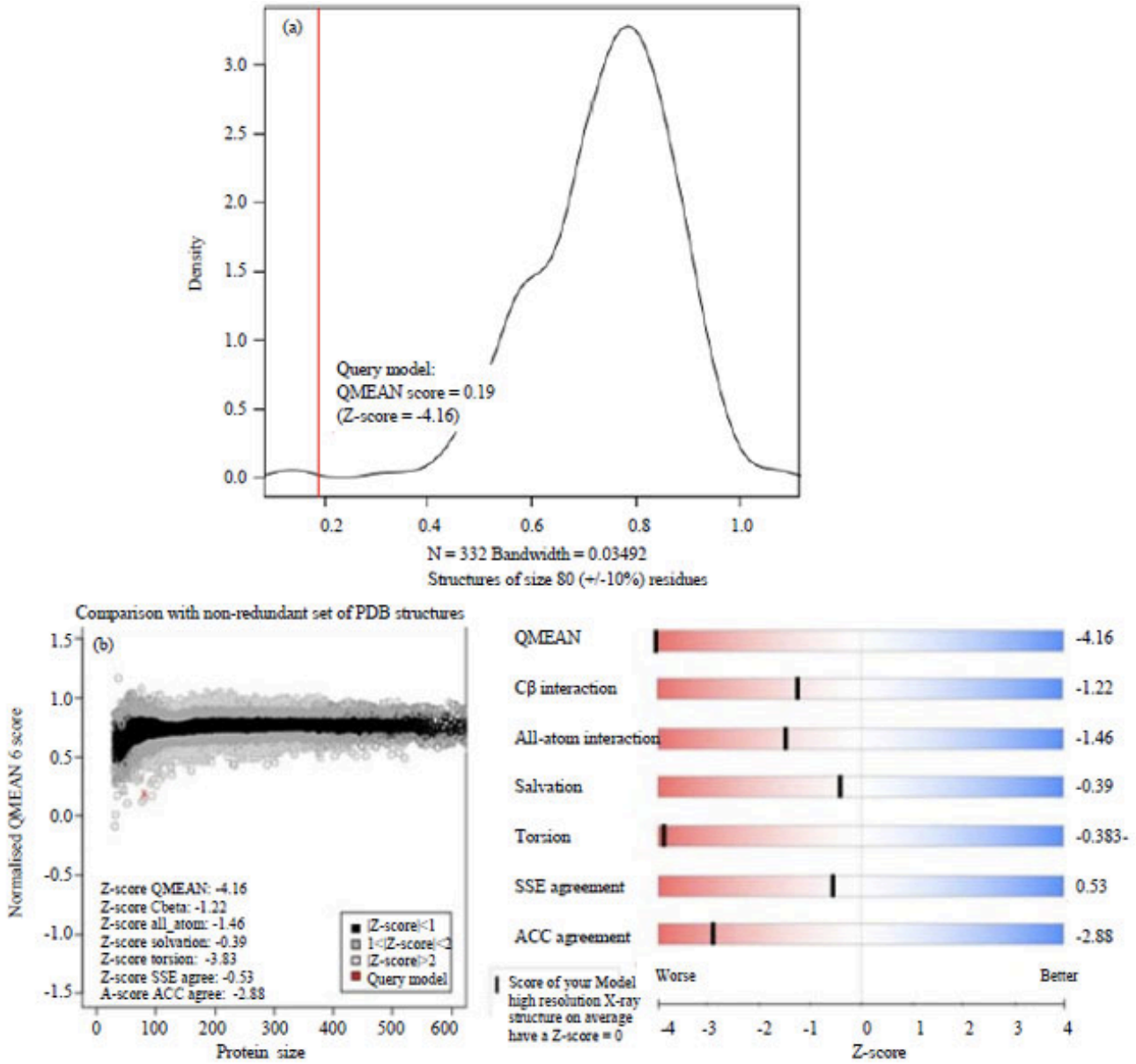
QMEAN provides both global (i.e. for the entire structure) and local (i.e. per residue) quality estimates on the basis of one single model

It is a linear combination of statistical potential terms: a new kind of torsion angle potential over three consecutive amino acids - to assess local geometry; a secondary structure-specific distance-dependent pairwise residue-level potential - to assess long-range interactions; a solvation potential term - to describe burial status of the residues.

QMEAN6 additionally uses two agreement terms evaluating the agreement of predicted (based on sequence) and calculated secondary structure and solvent accessibility

Both global scores are originally in a range [0,1] with one being good.

Model quality assessment: QMEAN



Model quality assessment: QMEAN

The three scoring functions are:

QMEANDisCo uses the single terms of QMEAN as a basis.

In addition, it has a term predicting local per-residue quality estimates by assessing the agreement of pairwise residue-residue distances with ensembles of distance constraints (DisCo) extracted from structures homologous to the assessed model. If no homologues are found, the DisCo scores are not used.

All terms are combined using neural networks trained to predict per-residue scores in range [0,1]. The global score is the average per-residue score and the provided error estimate is based on global QMEANDisCo scores estimated for a large set of models of similar size to the input

Model quality assessment: ProQ2

ProQ2 also uses machine learning, specifically support vector machines (SVMs), to predict local as well as global quality of protein models

Scalar features from each protein model based on properties that can be derived from its sequence, e.g. conservation, predicted secondary structure, and exposure, or 3D coordinates, e.g. atom-atom contacts, residue-residue contacts, and secondary structure, are calculated and used as features to predict model correctness

http://bioinfo.ifm.liu.se/proq2/

The CASP experiment

Critical assessment of techniques for protein structure prediction

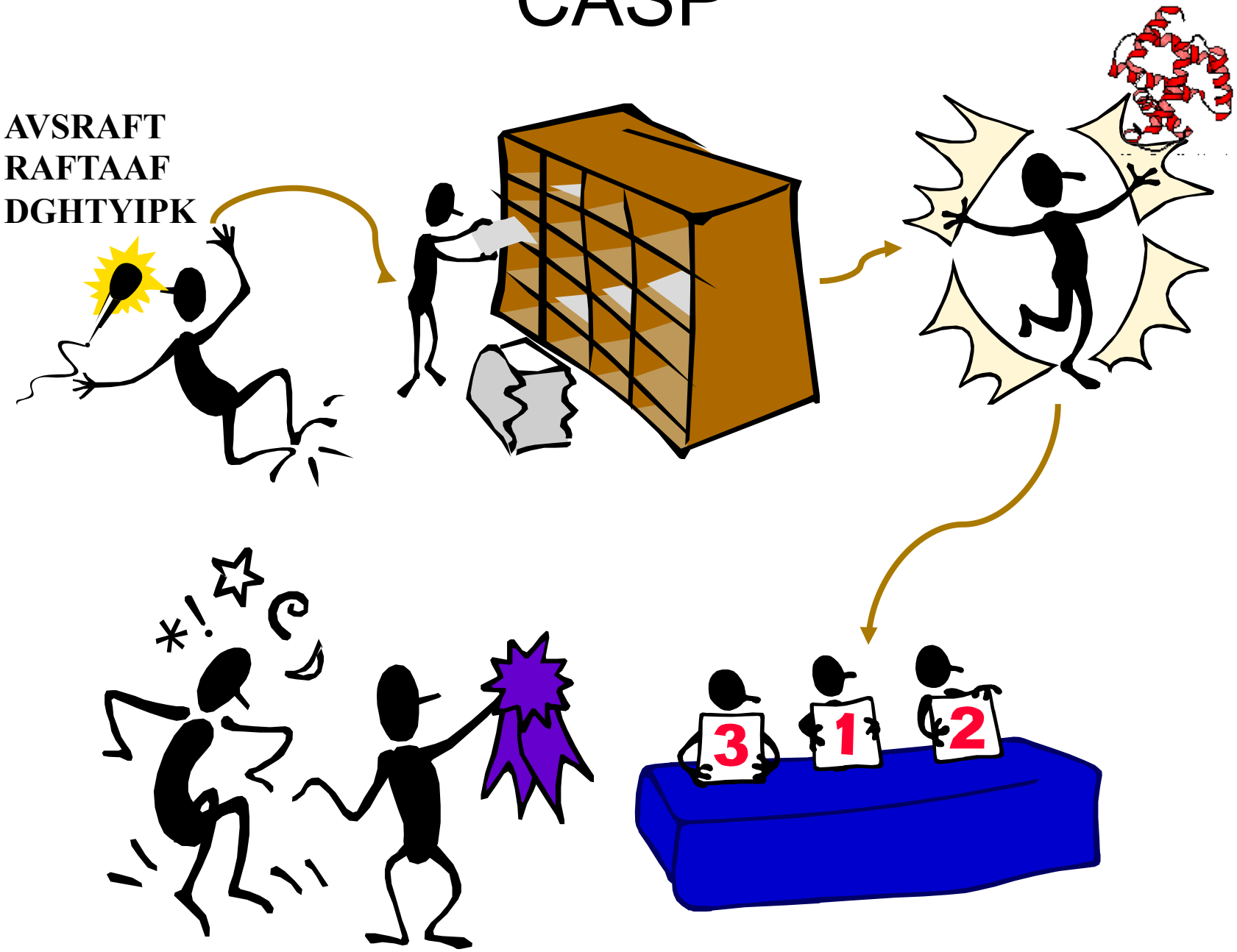
CASP is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994; in it the accuracy of different prediction methods is blindly assessed on common targets

The primary goal of CASP is to help advance the methods for predicting protein three-dimensional structure from its amino acid sequence

Like in a “world championship” in this field of protein structure prediction, over 100 research groups from all over the world participate in CASP on a regular basis

A model quality assessment section is included since CASP-7

CASP



The CASP experiment

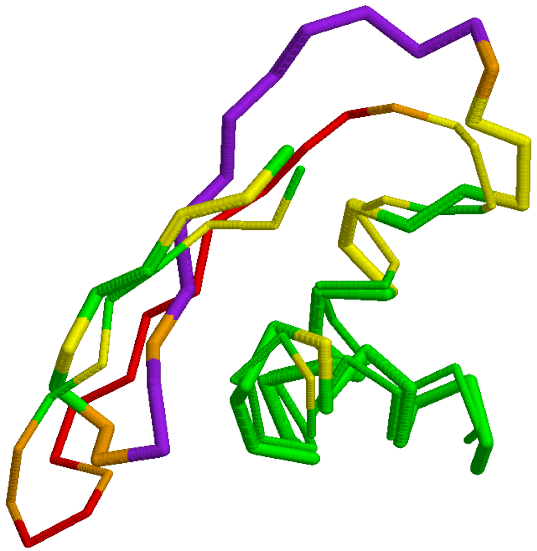
Critical assessment of techniques for protein structure prediction

Traditionally predictions in CASP are divided in:

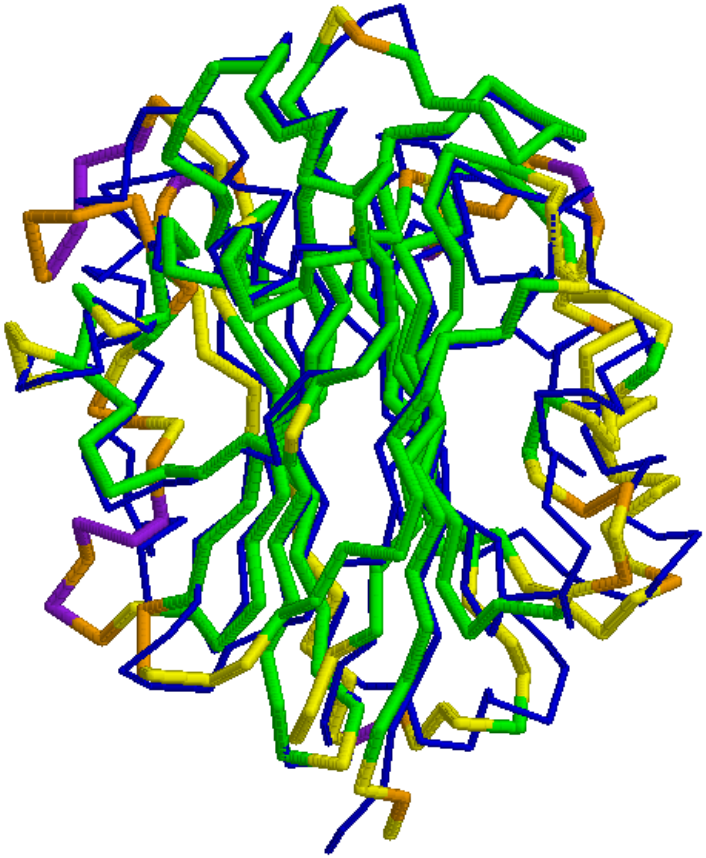
template-based (including comparative modeling)
&
template free modeling

A revolution in CASP happened since 2020…

Target 89_2
Group 126
22 %



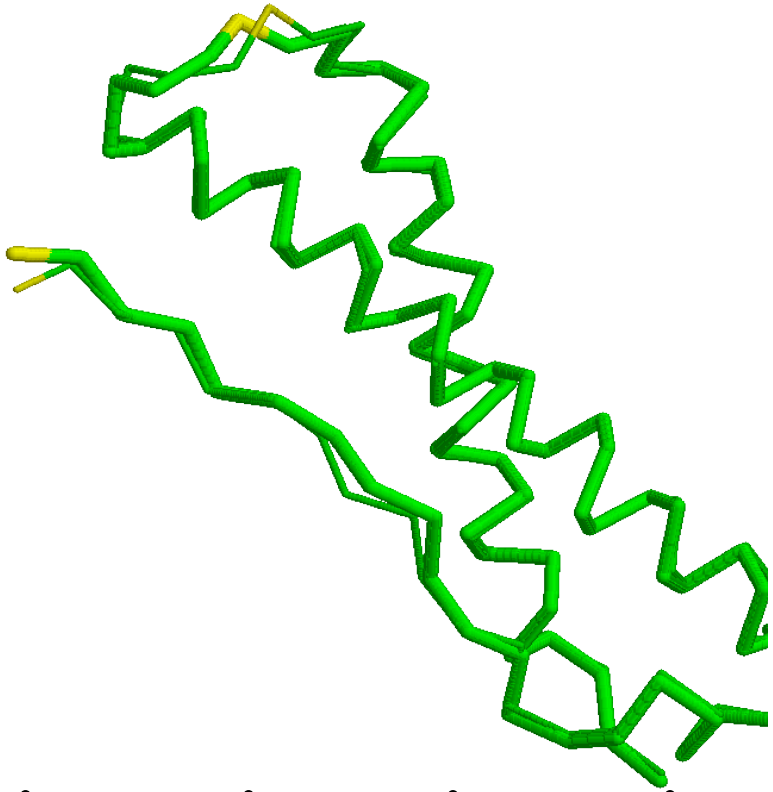
Target 142
Group 020
26 %



< 2.0 Å < 4.0 Å < 8.0 Å > 8.0 Å

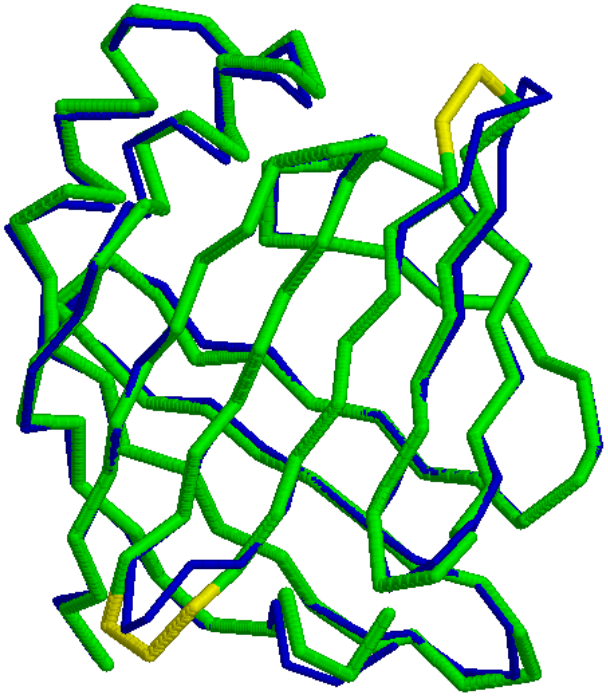
Sample results from CASP 5, held in 2002

Target 128_2
Group 12
60 %



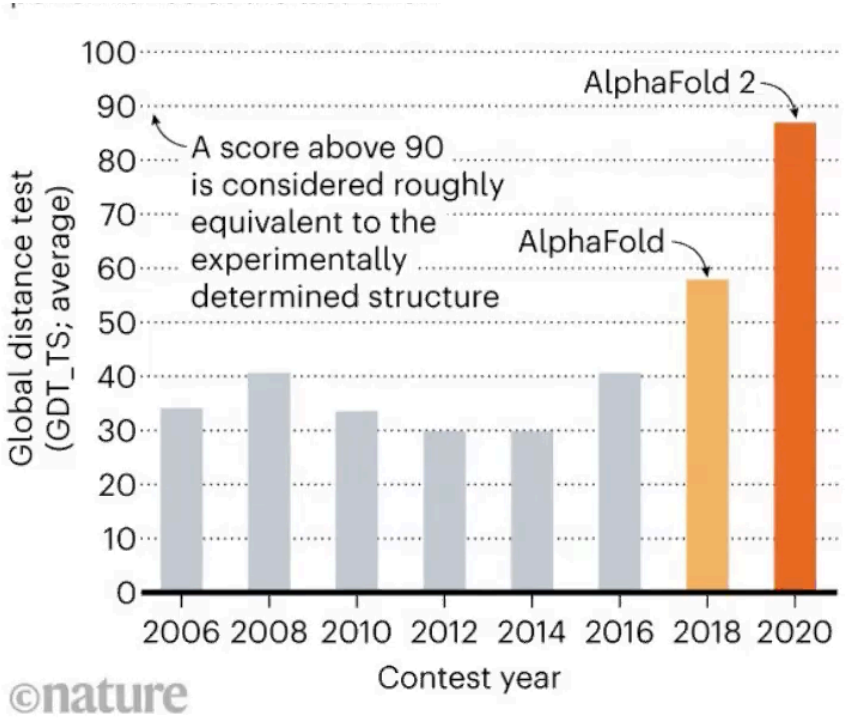
< 2.0 Å < 4.0 Å < 8.0 Å > 8.0 Å

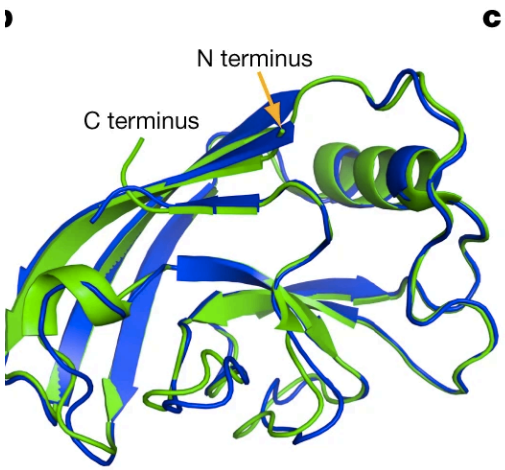
Target 137
Group 427
43 %

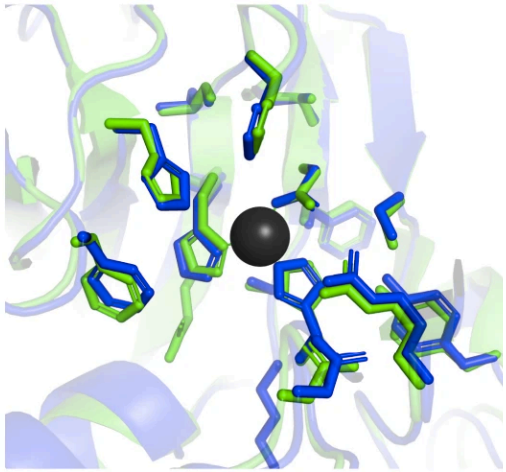


Sample results from CASP 5, held in 2002

Performance of predictors in CASP14 (2020)

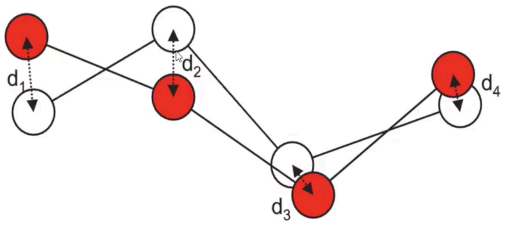






GDT_TS: percentage of corresponding α-carbons within a 4 Å distance

- Root Mean Square Deviation (RMSD)



$$rmsd = \sqrt{\frac{\sum_i d_i^2}{n}}$$

n = Number of aligned residue pairs
d = Distance between each pair of atoms

Lesson 11. Content

1. The protein folding problem. Knowing a protein sequence is not enough to predict its folding by physico-chemical methods.
2. Comparative (Homology) modelling. Choice of template and alignment are the crucial steps of the method. The higher the target-template sequence identity, the better the expected quality of the model.
3. Comparative(Homology) modelling. Modelling of the protein core and loops. Serious errors in the core can be solved by correcting the target-template alignment. The loops (non-core) modelling is the method weakness.
4. Comparative(Homology) modelling. Optimization and quality check. The model quality can be assessed based on statistical potentials and evolutionary information.