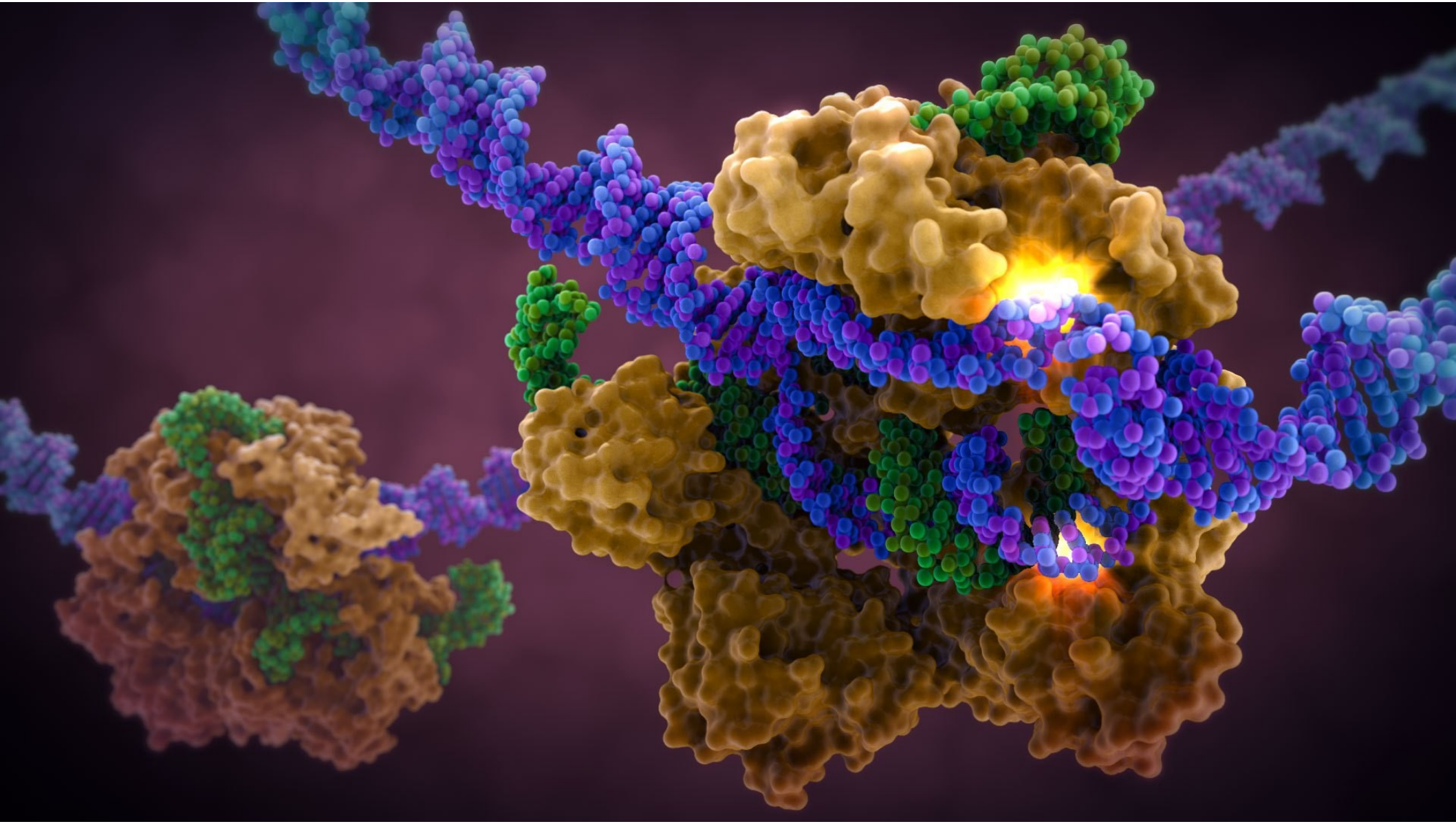
Structural bioinformatics

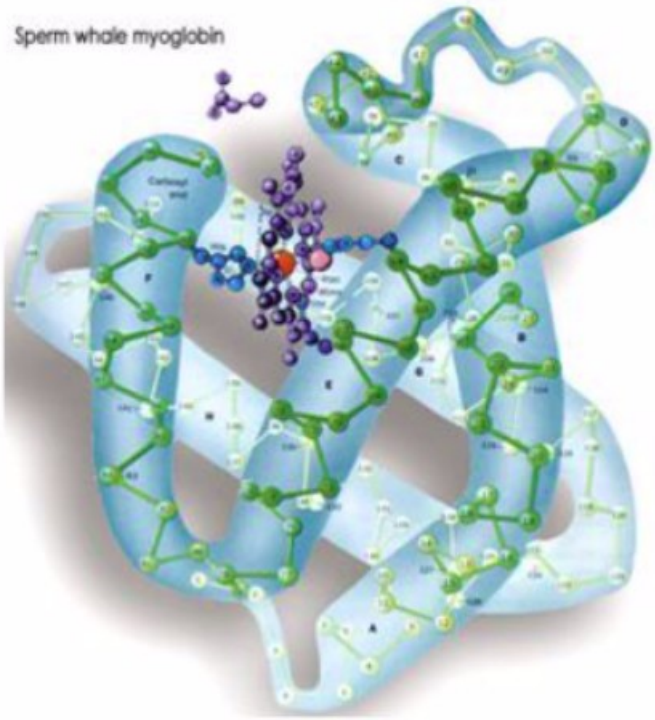


Erdős Gábor (gerdos@caesar.elte.hu)
2018.12.03

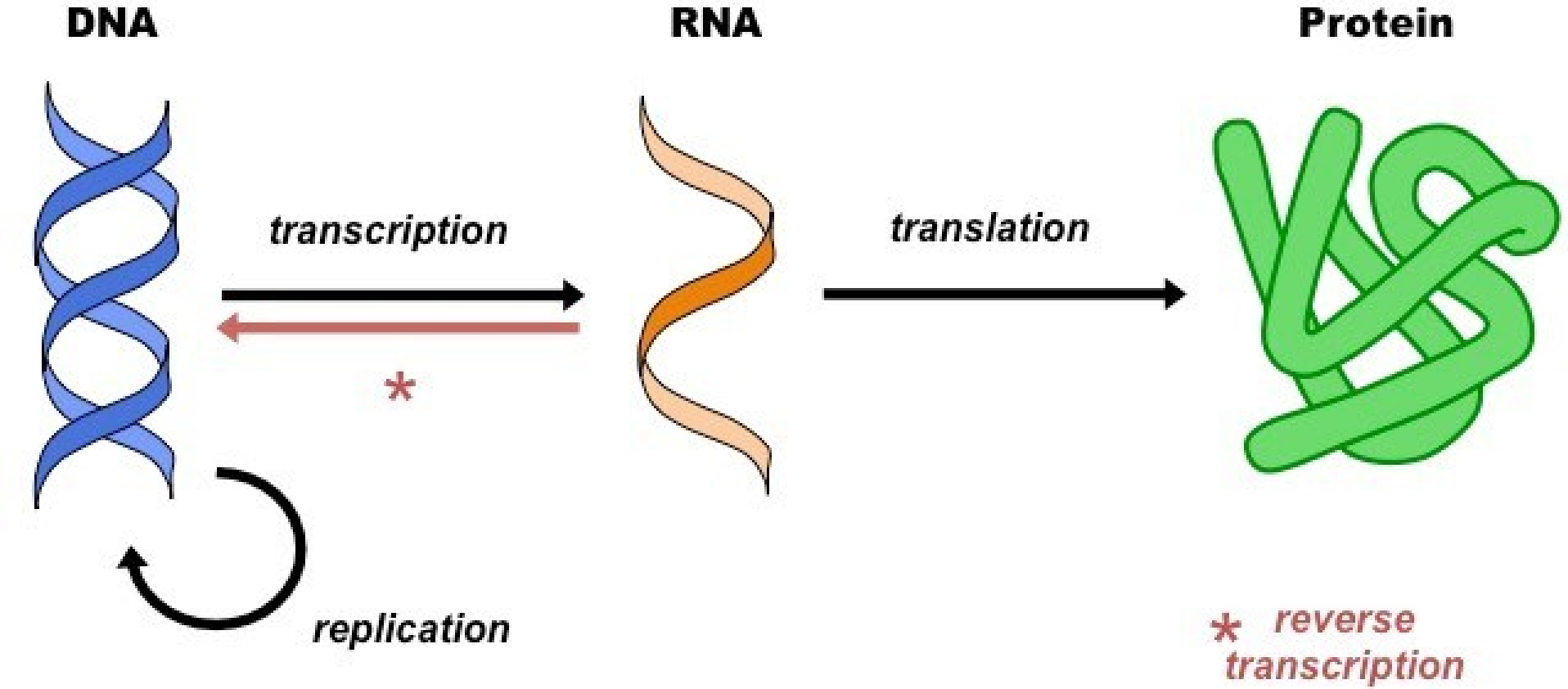
Áttekintés

- Fehérje szerkezetek álltalános tulajdonságai
- Szerkezet meghatározási módszerek
- PDB adatbázis
- Szerkezet vizualizáció
- Szerkezet összehasonlitas
- Szerkezet predikciós módszerek

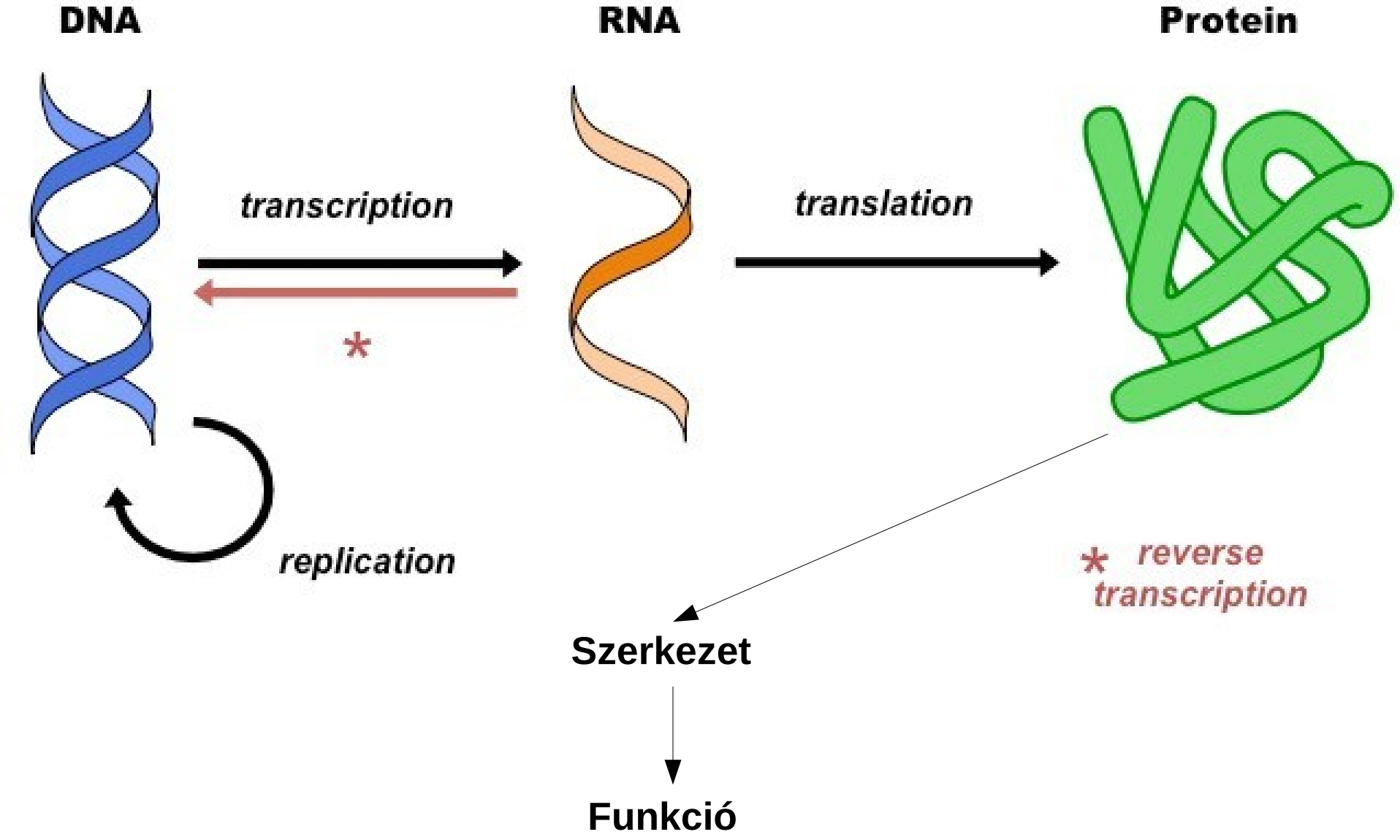
- **1958**
  - John Kendrew et al., published the first structure of a globular protein, myoglobin.
  - " Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry"
- **1962**
  - Nobel prize in Chemistry was awarded to Max Perutz and John Kendrew.
- **Now**
  - ~80,000 structures in protein database (PDB)



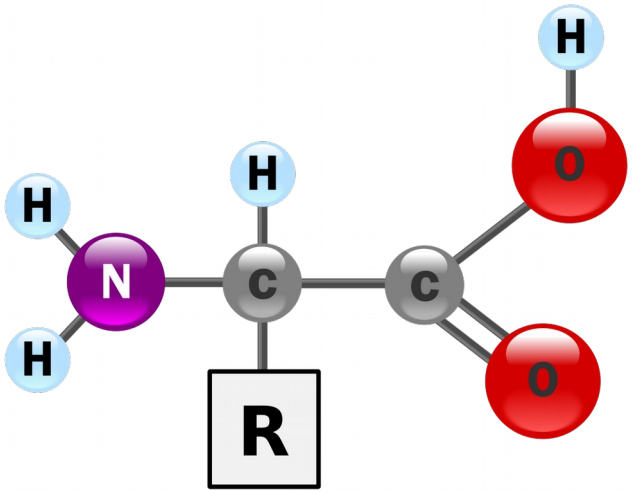
Molekuláris biológia centrális dogmája

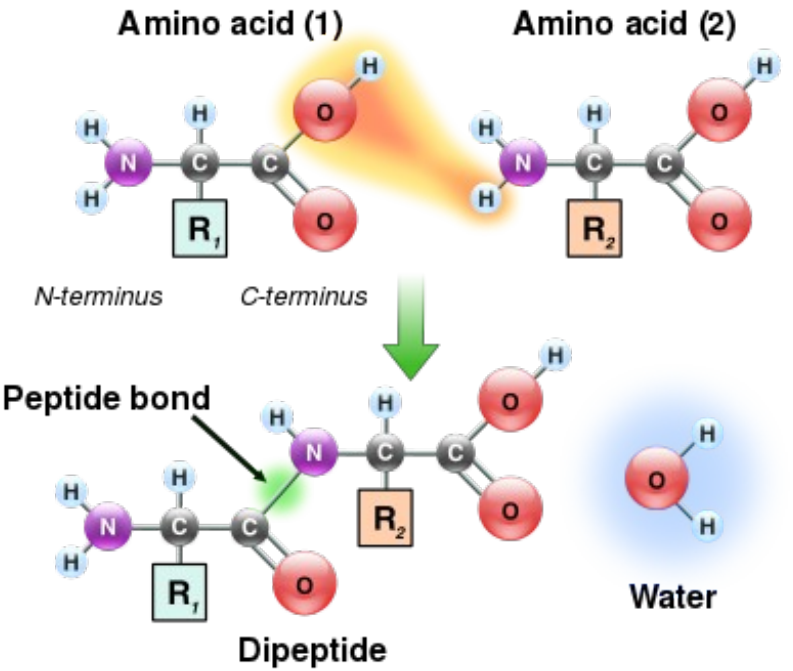


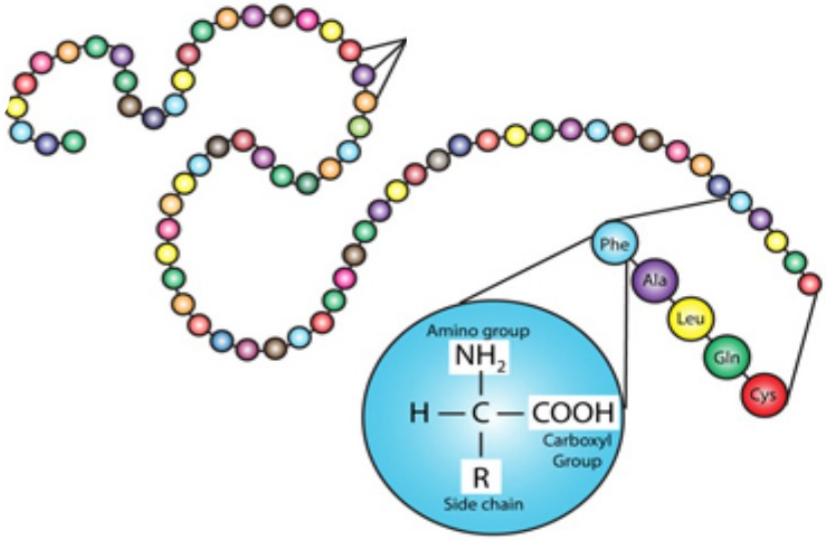
Molekuláris biológia centrális dogmája



Amino acid


H
H
N
H
C
R
C
O
O


Amino acid (1)
Amino acid (2)
N-terminus
C-terminus
Peptide bond
Water
Dipeptide


Phe
Ala
Leu
Gln
Cys
Amino group
NH2
H — C — COOH
Carboxyl Group
R
Side chain

How many different sequences can a 100 amino acid long protein have?

$20^{100} \sim 10^{130}$

Number of protons in the observable universe is around 10^{80}

Proteins are usually longer
The longest one is around 30 000 AA

Why are we interested?

Function is related to the **structure**

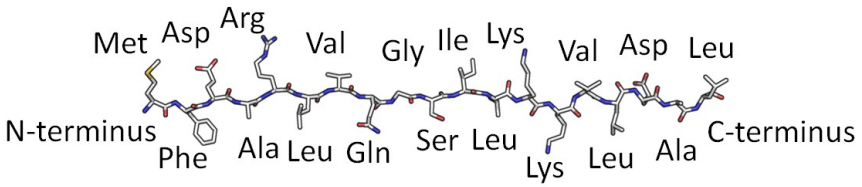
- Understand biological processes (DNA, RNS, enzymes, hormones, receptors)
- Diseases
- Drug design, protein – drug interactions

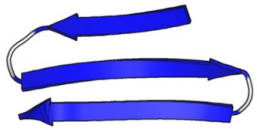
Primary

Secondary

Tertiary

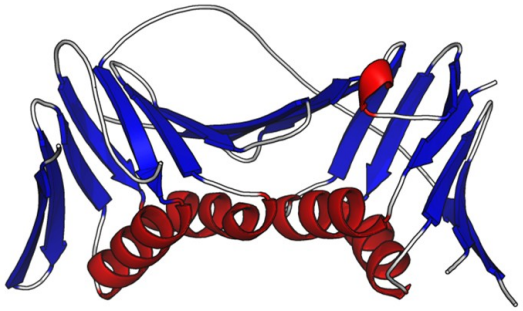
Quaternary

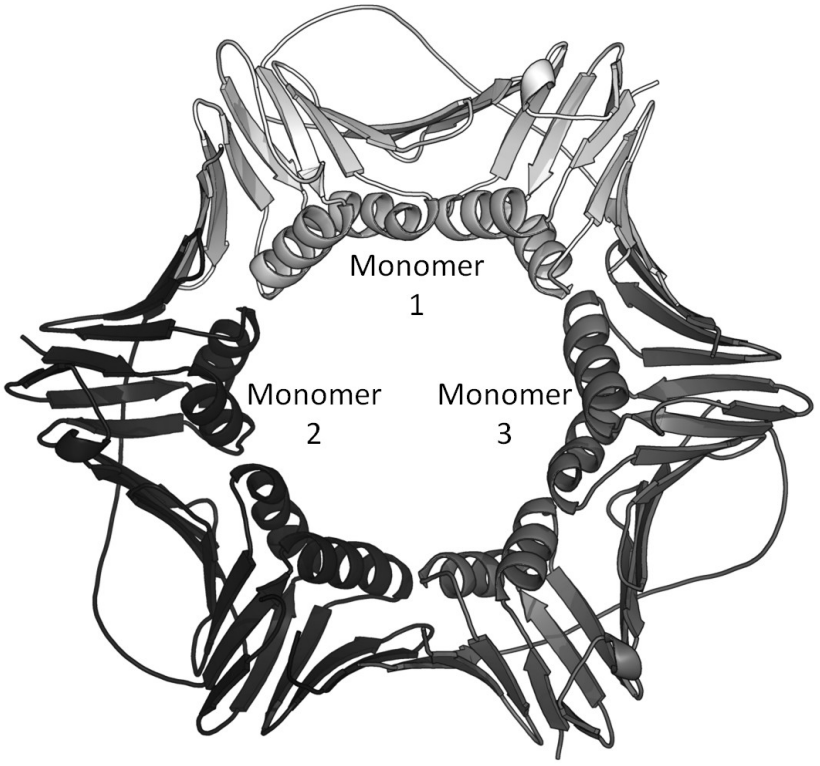




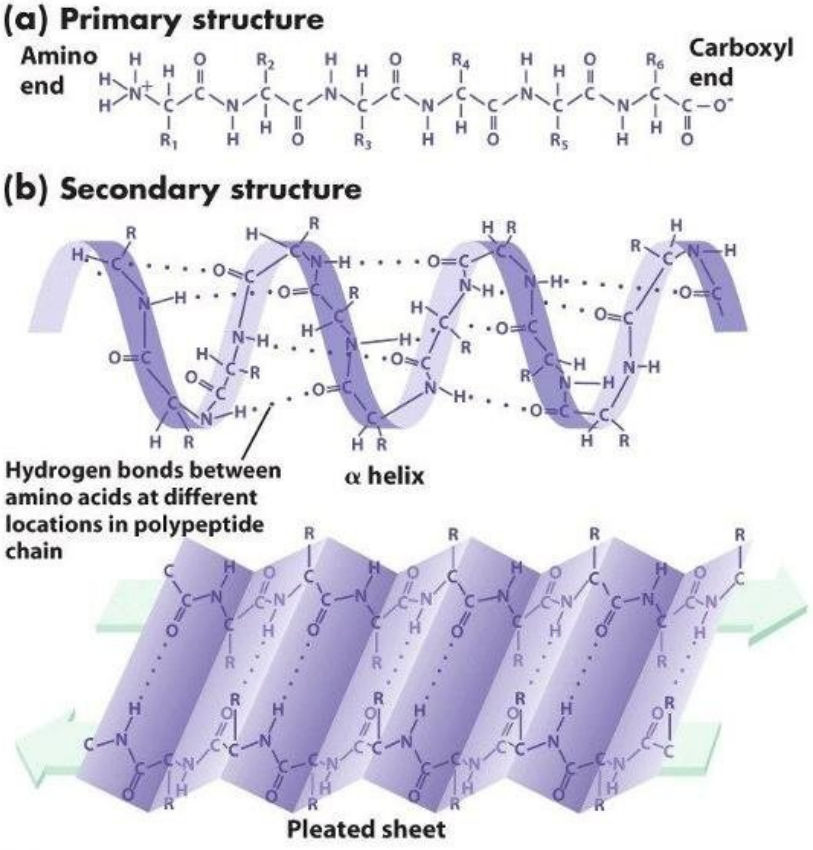
β-Sheet (3 strands)

α-helix

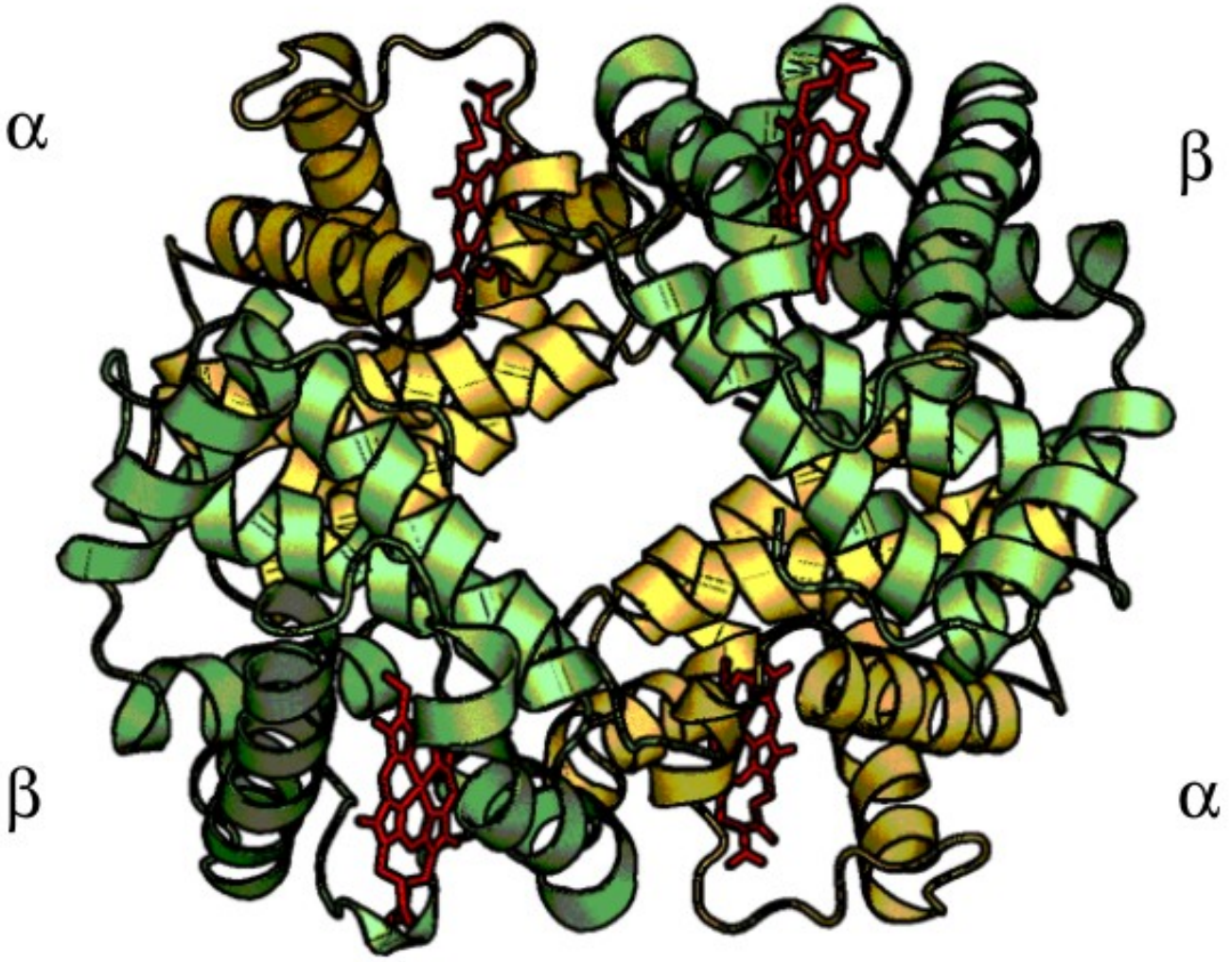




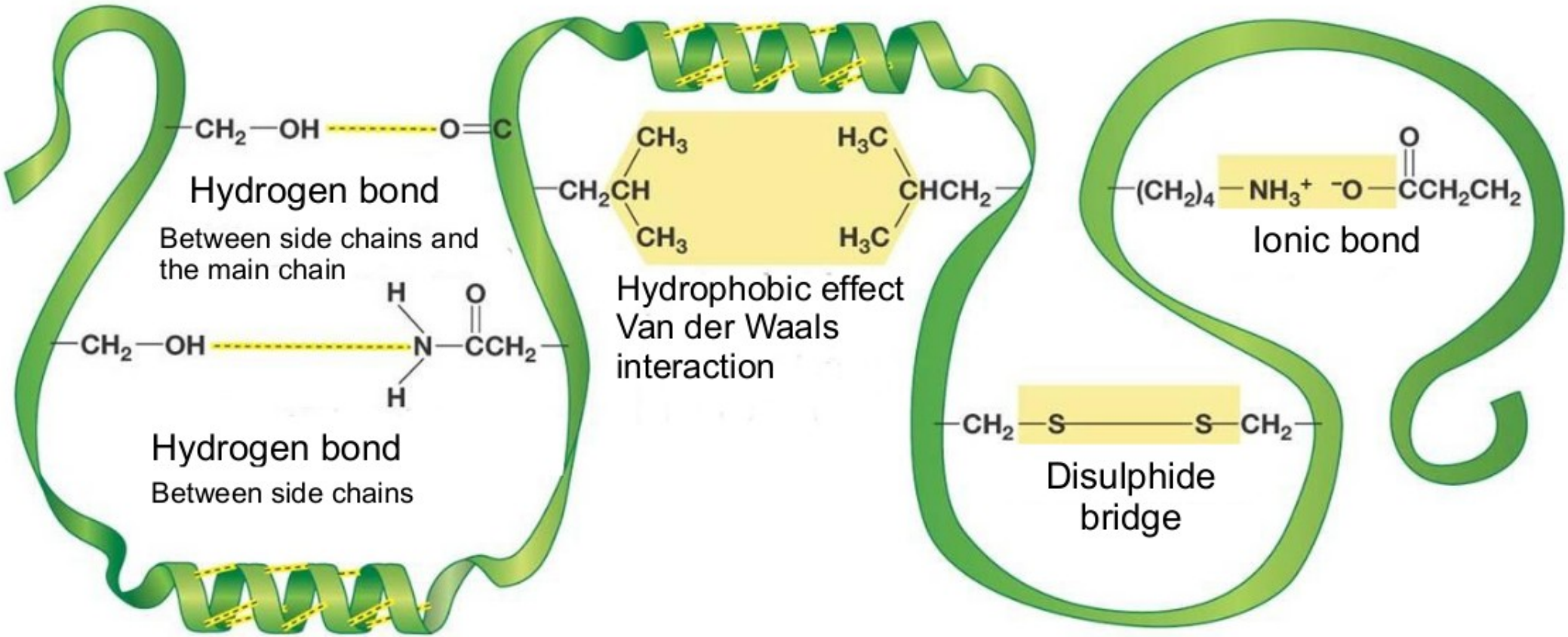
"**Protein structure** is the three-dimensional arrangement of atoms in a protein molecule"



Hemoglobin

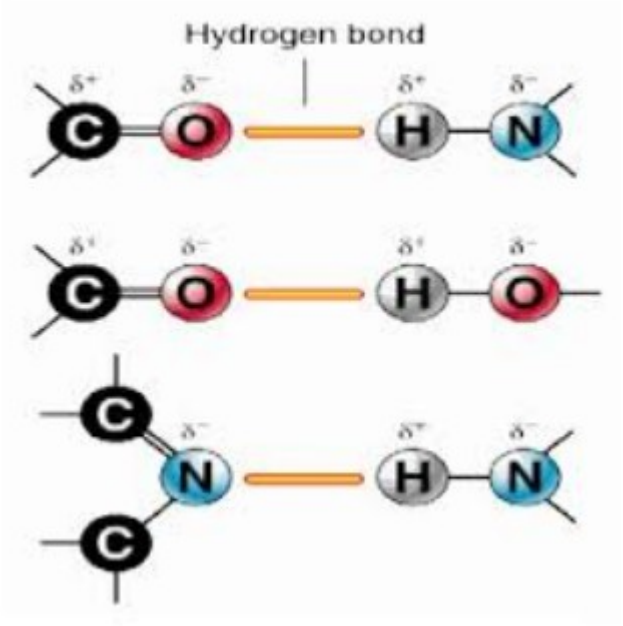


Stabilization of a protein



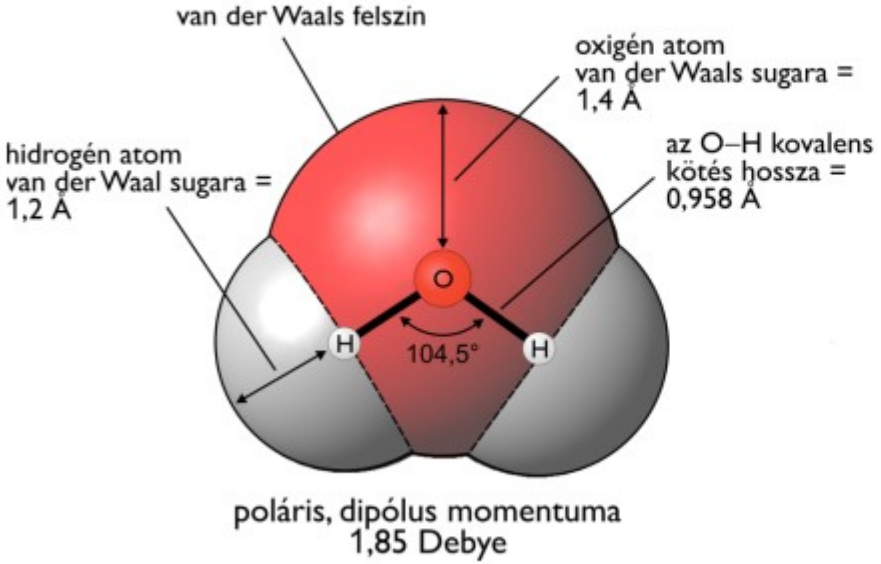
Hidrogén kötés

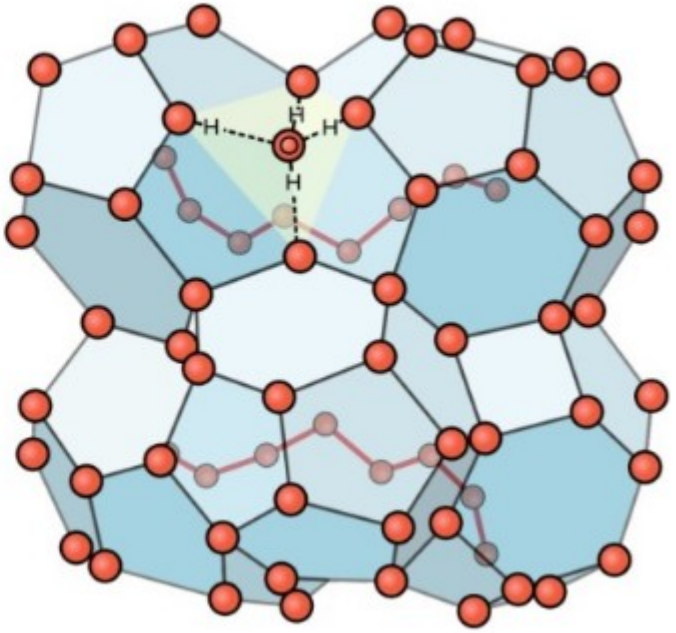
Hidrogén kötés egy hidrogén, és egy nagy elektronegativitású atom között kialakuló kötés



- Gyengébb mint a kovalens kötés (1-7 kcal/mol vs. 100 kcal/mol)
- Domináns stabilizáló erő
- Elsődleges szerkezet stabilabb mint a másodlagos, és így tovább

Víz

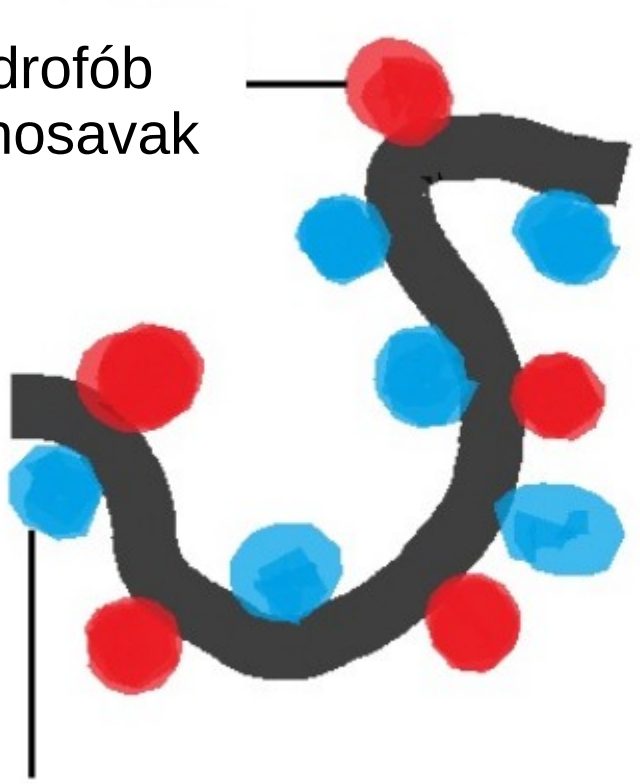




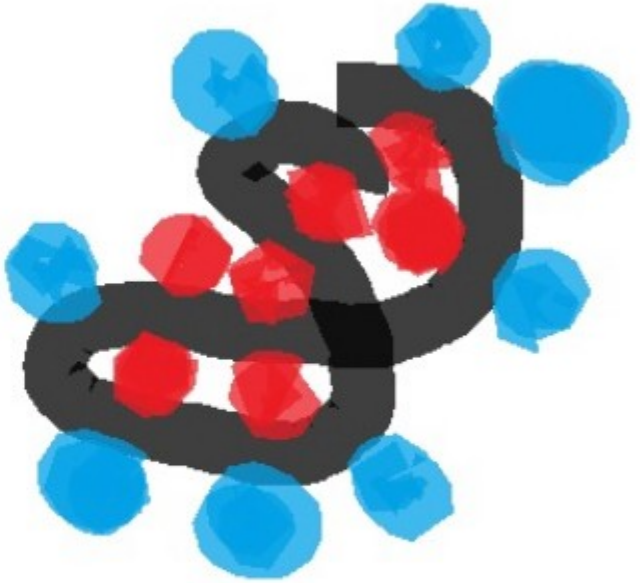
Hidrofób effektus

Entrópikus erők

Hidrofób effektus

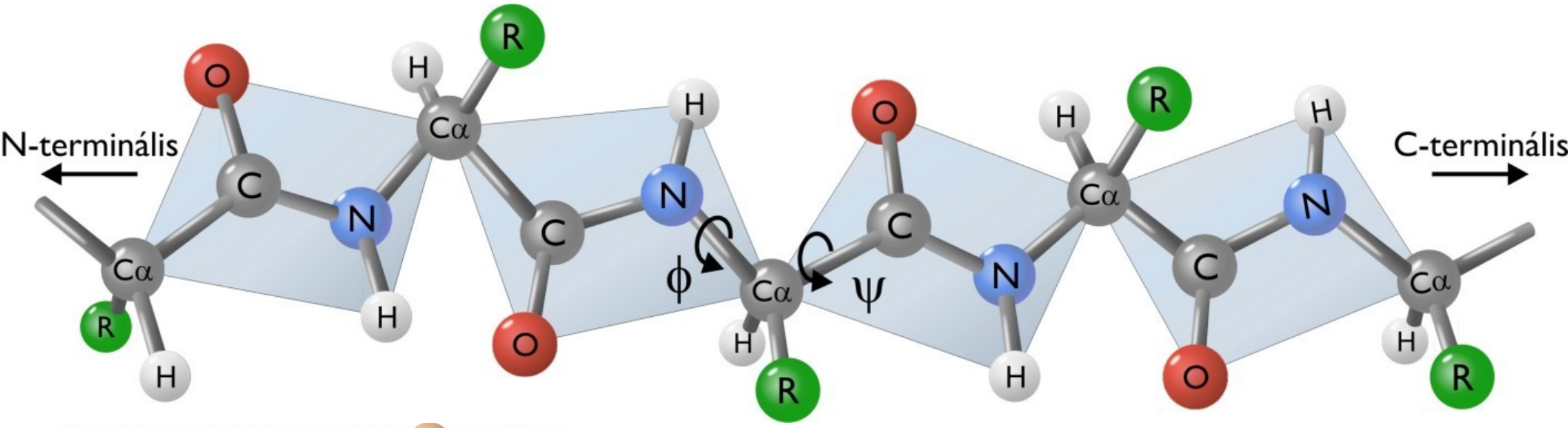


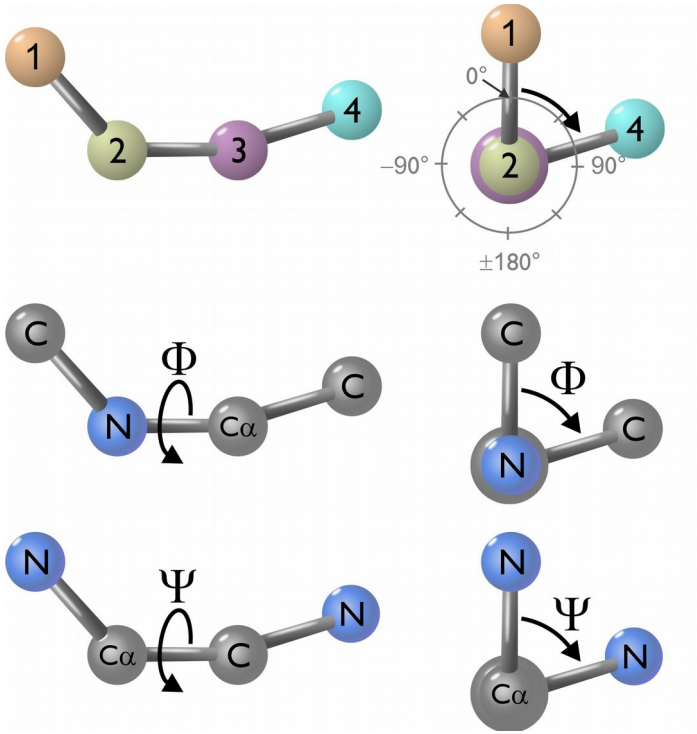
Izolált állapot



Vizes közeg

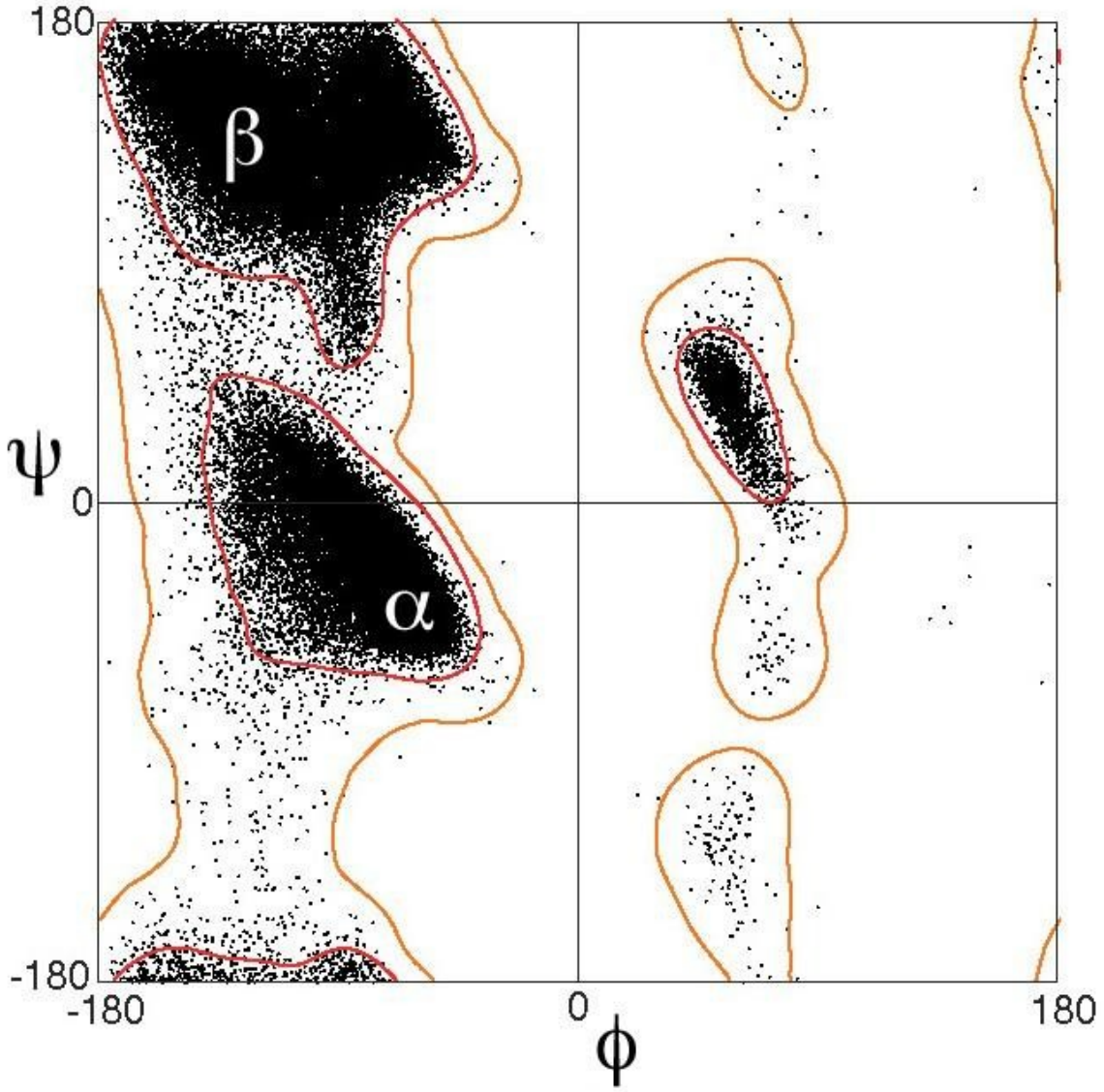
Főlánc konformációja





A főlánc atomjai csak bizonyos (torziós) szögeket vehetnek fel (φ, ψ). Ezek a kényszerek bizonyos preferenciákat alkotnak.

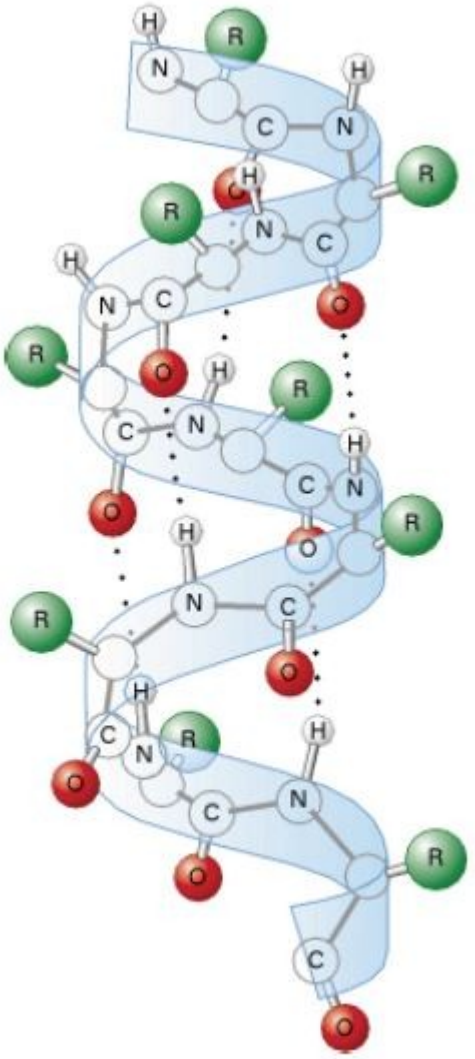
Ramachandran plot



- A φ, ψ szögeket használva elkészíthetünk egy preferencia térképet
- Minden másodlagos szerkezeti elemnek megvan a maga tere
- Bizonyos aminosavak problémásak

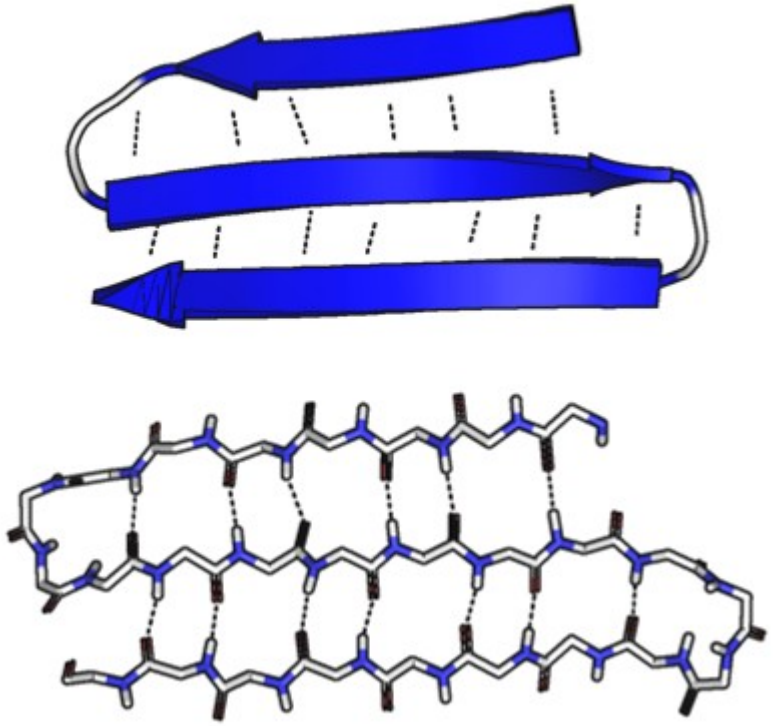
Alfa hélixrek

- Globuláris fehérjék ~30%üa
- 5-40 aminosav hosszú (átlagosan 10)
- Hidrogén híd kötések önmagukban gyengék
- A hélix klialakulás az aminosavak összetételétől függ



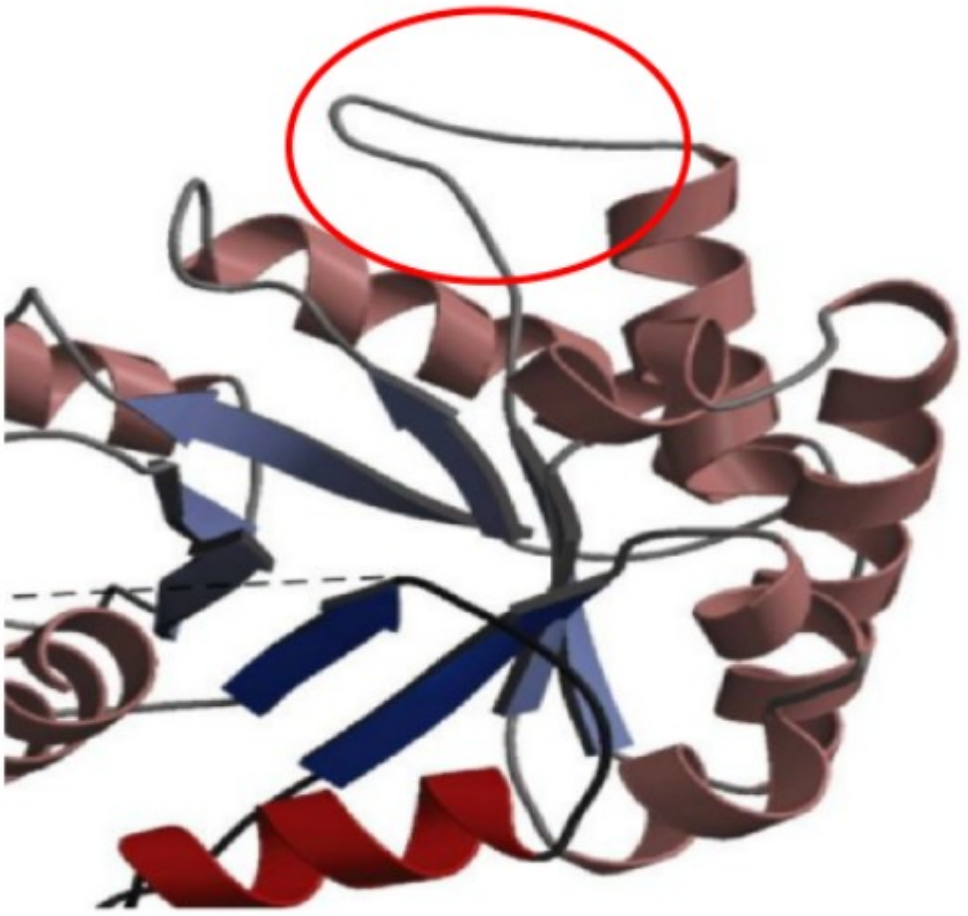
β sheet conformation

- Globuláris fehérjék ~30%-a
- 5-10 aminosav parallel vagy antiparallel redőbre rendeződik
- Hidrogén hidak stabilizálják



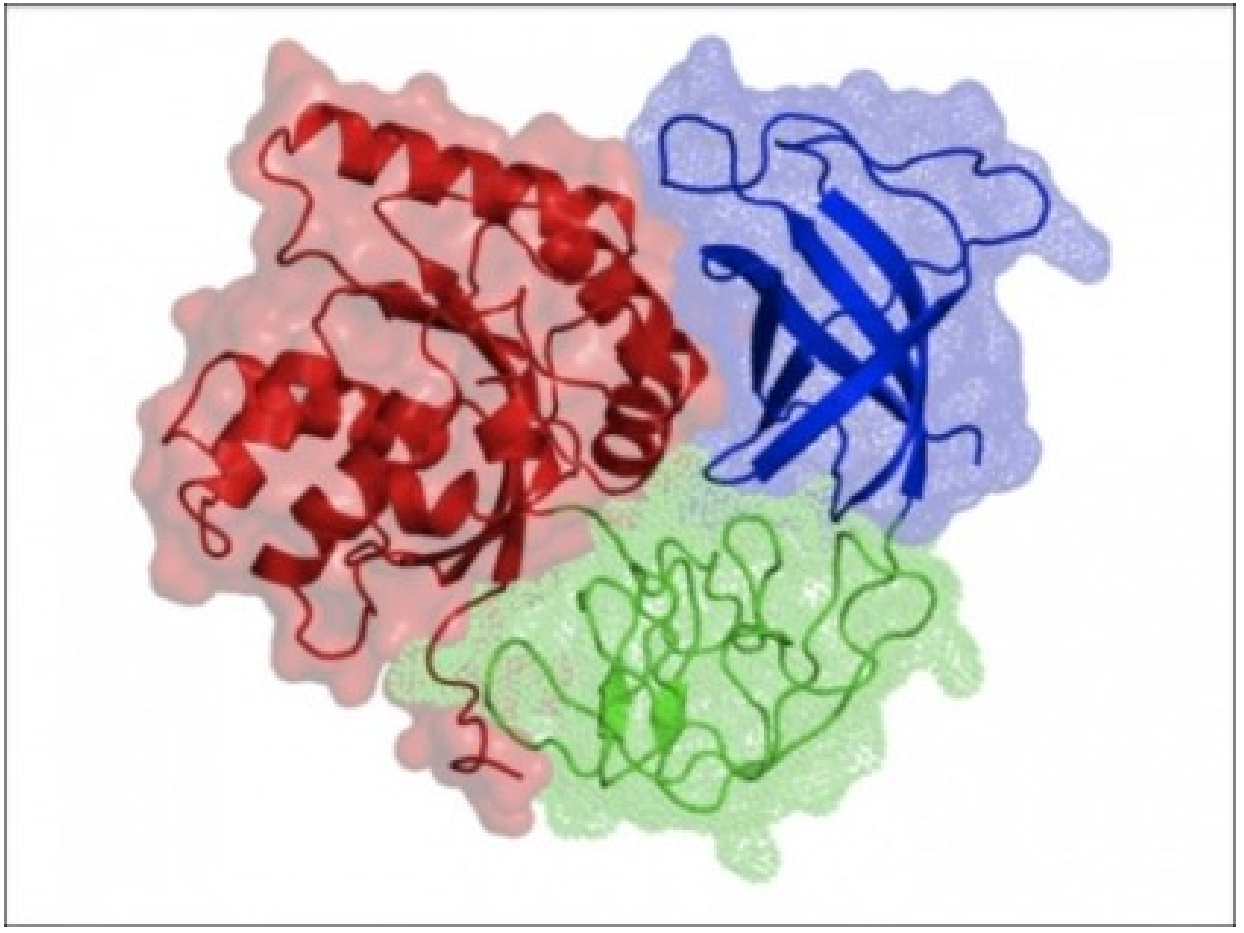
Hurkok és kanyarok

- Tipikusan hidrofil régiók
- A molekula külső részén, oldószernek kitett részen
- Gyakran található bennük kötőhely, aktív centrum
- Hidrogén híd kötések az oldószerrel



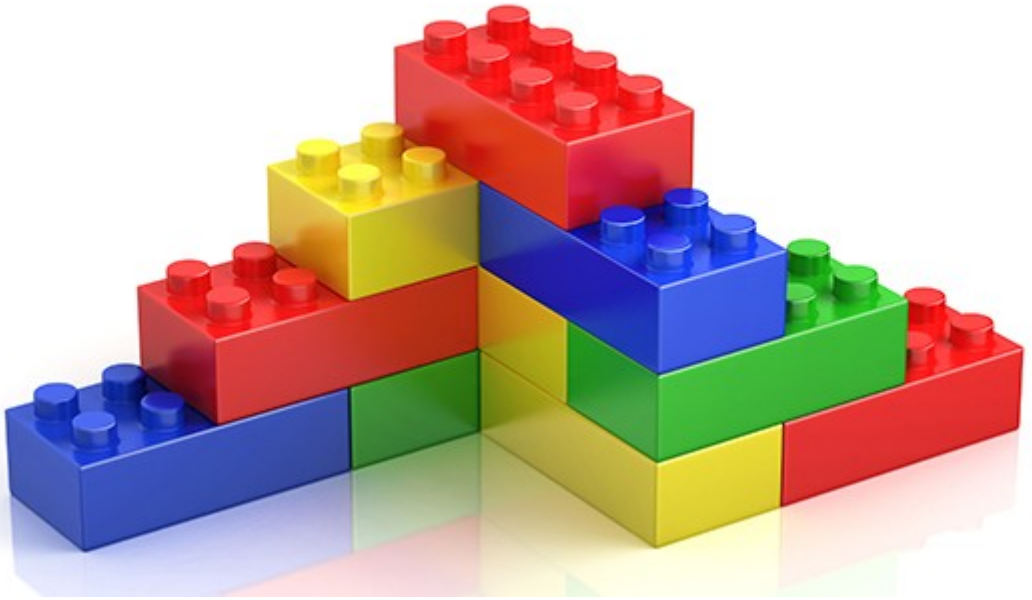
Domének

Sok fehérjében található különálló kompakt szerkezeti egység



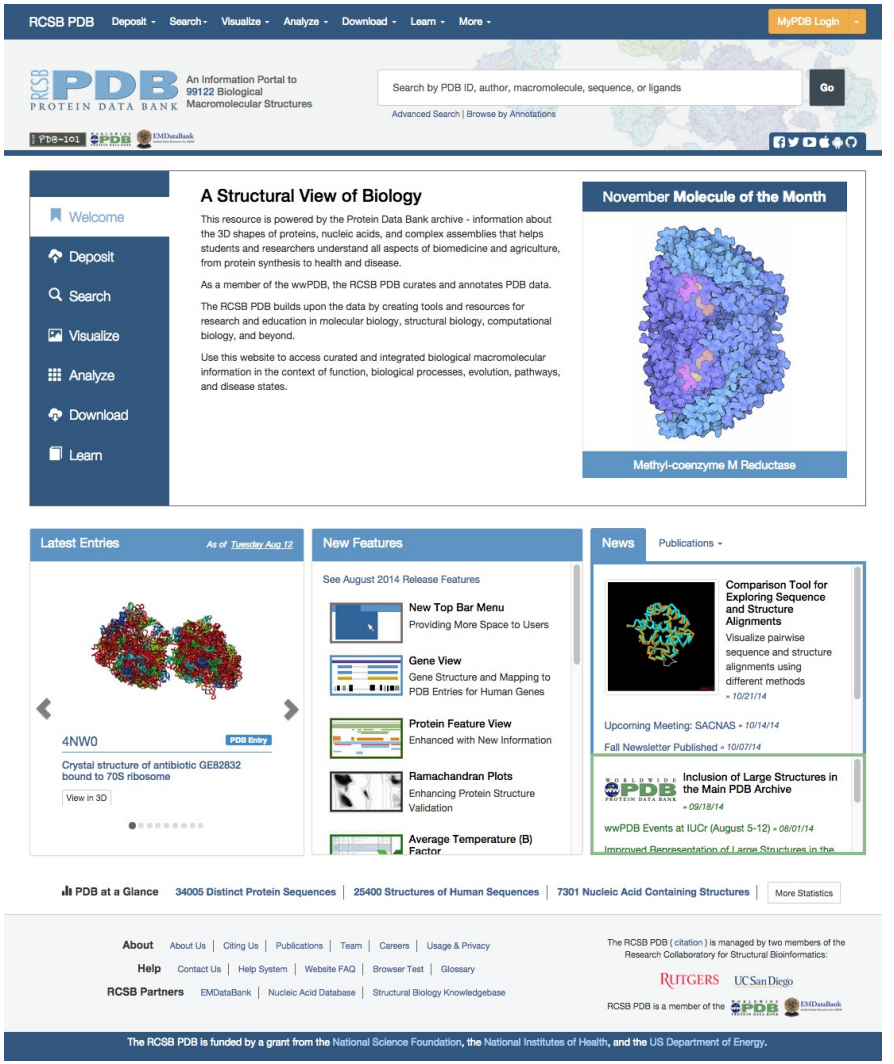
Domének

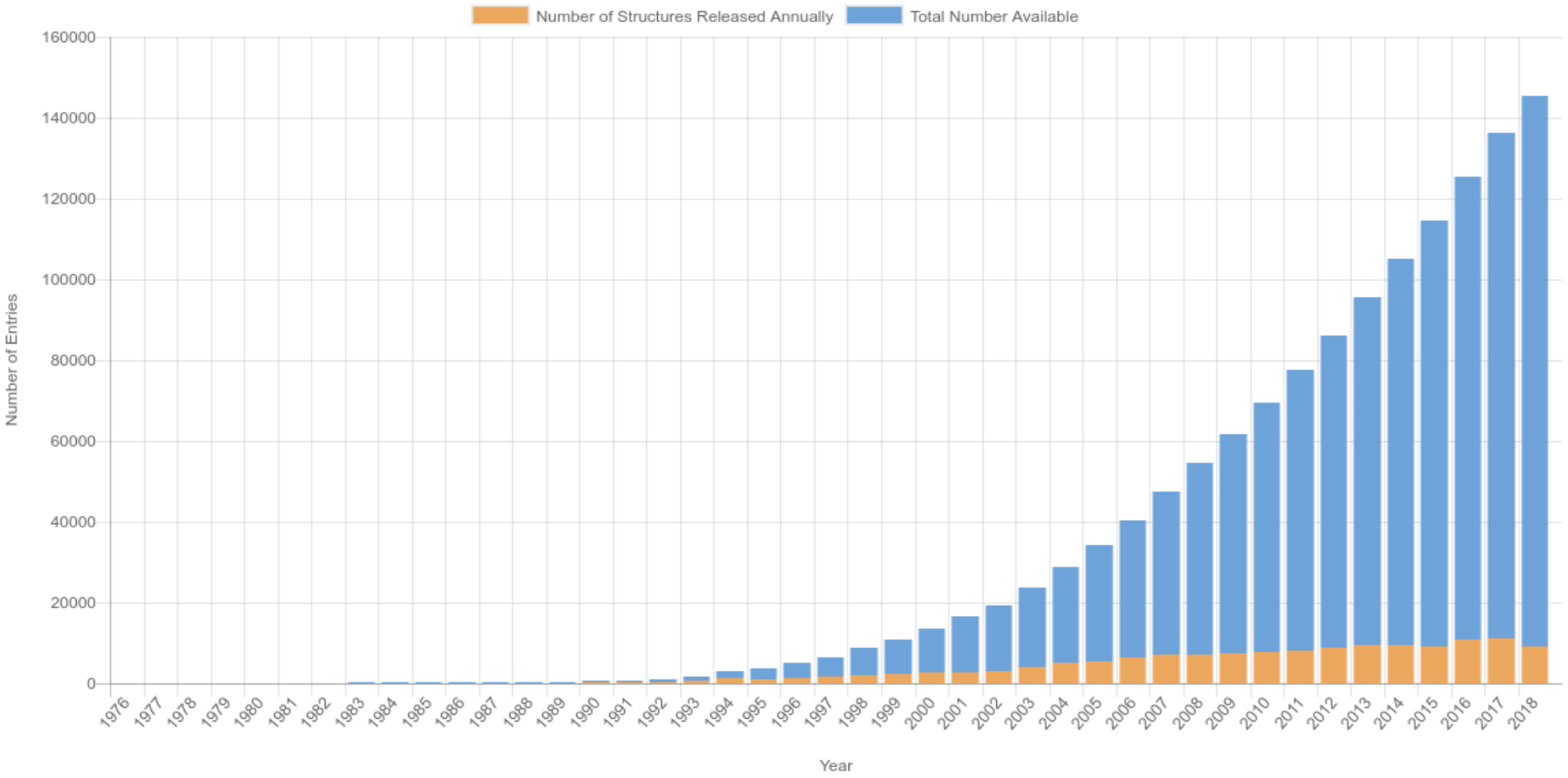
- Kompakt globuláris szerkezetek
- Felépítő egységek
- Önálló funkcionális egység



Hol találhatunk szerkezeteket?

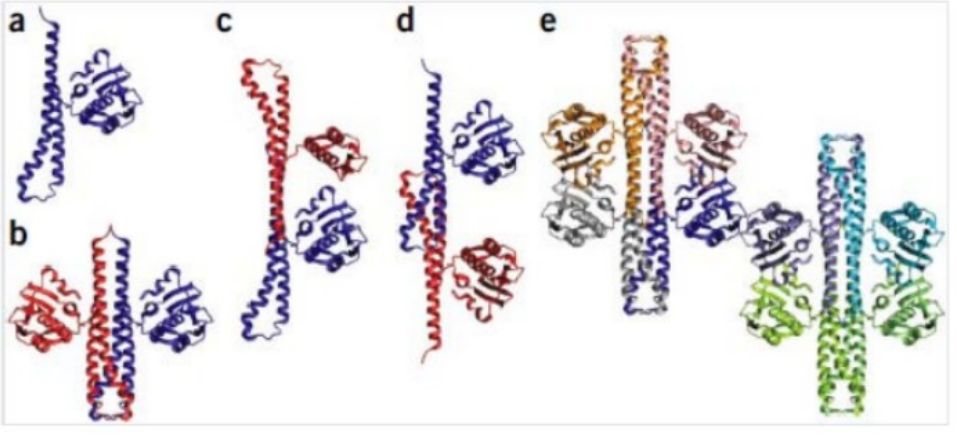
- Adatbázis
- Wordwide Protein Data Bank (wwPDB)
- 3 különböző:
  - PDB Europe
  - PDB Japan
  - **RSCB PDB**
    - Könnyen kezelhető
    - Jól fenntartott



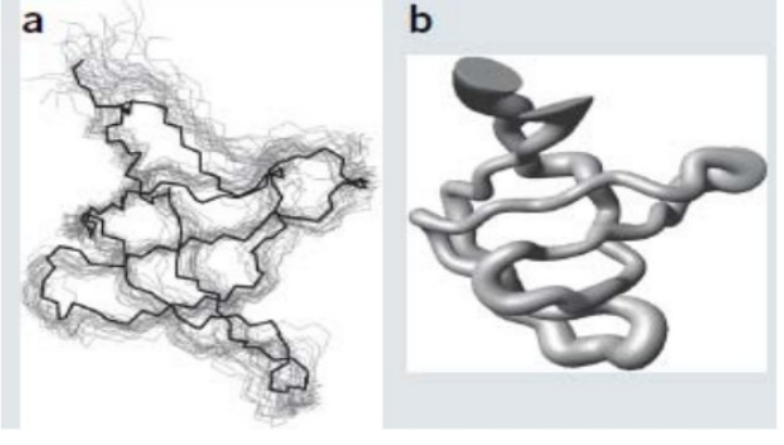

Number of Structures Released Annually
Total Number Available
Number of Entries
160000
140000
120000
100000
80000
60000
40000
20000
0
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year

Honnan jönnek a szerkezetek?

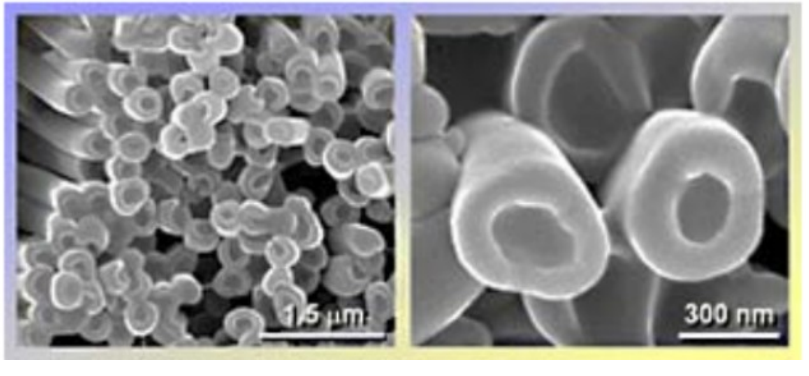
Röntgen krisztallográfia



NMR



Electron mikroszkóp



Honnan jönnek a szerkezetek?

Legtöbb szerkezetet röntgennel oldják meg (86%)

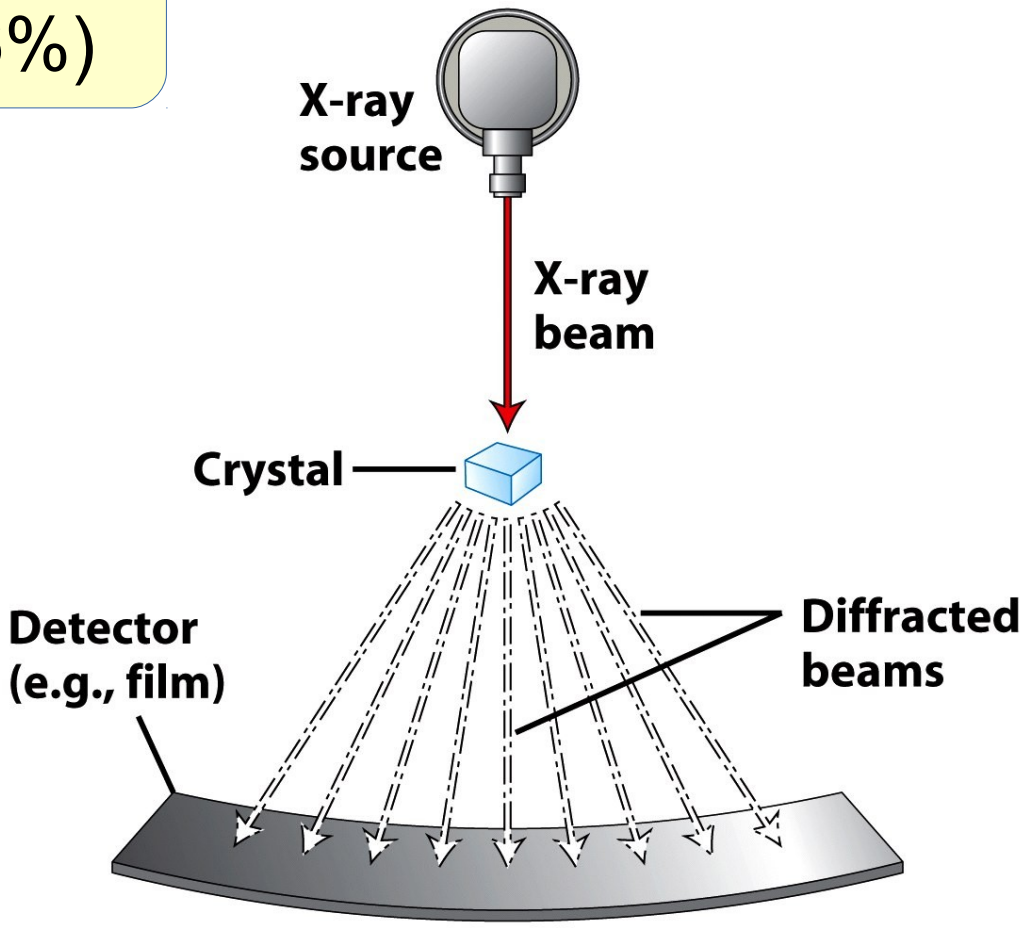
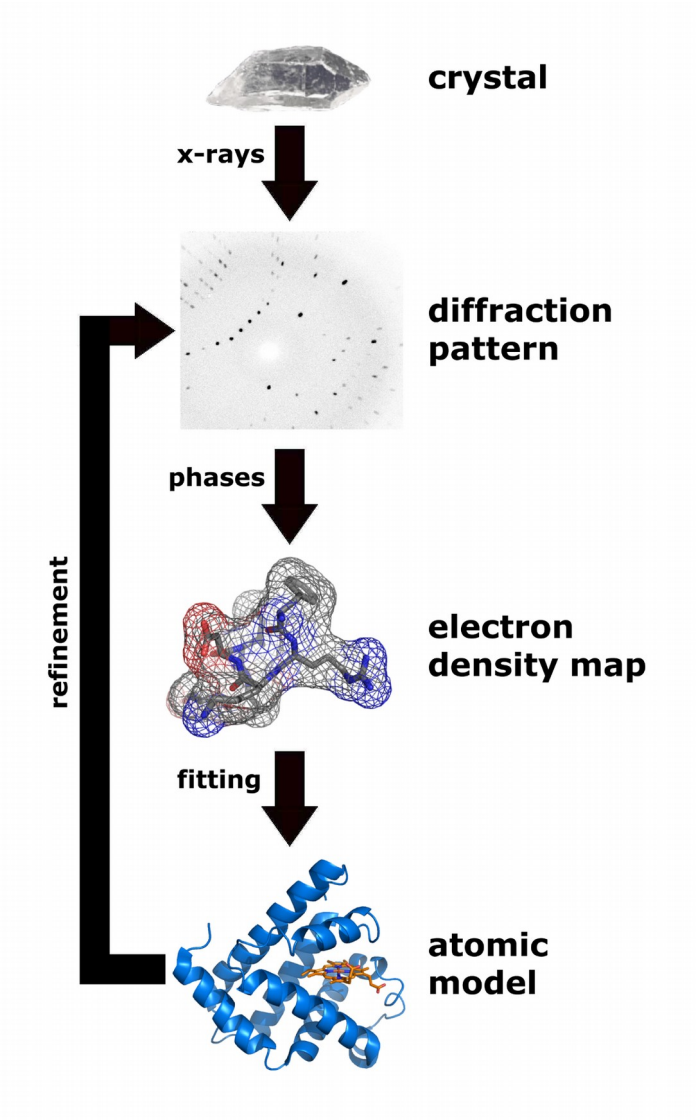
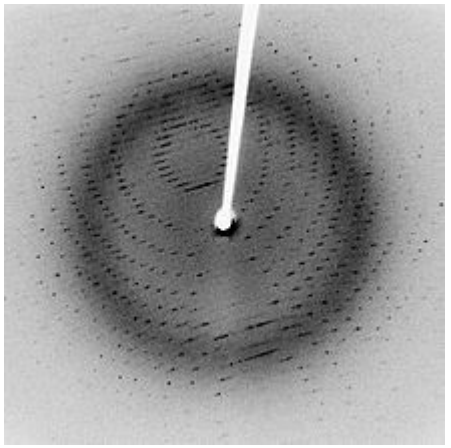


Figure 3-42a
Molecular Cell Biology, Sixth Edition

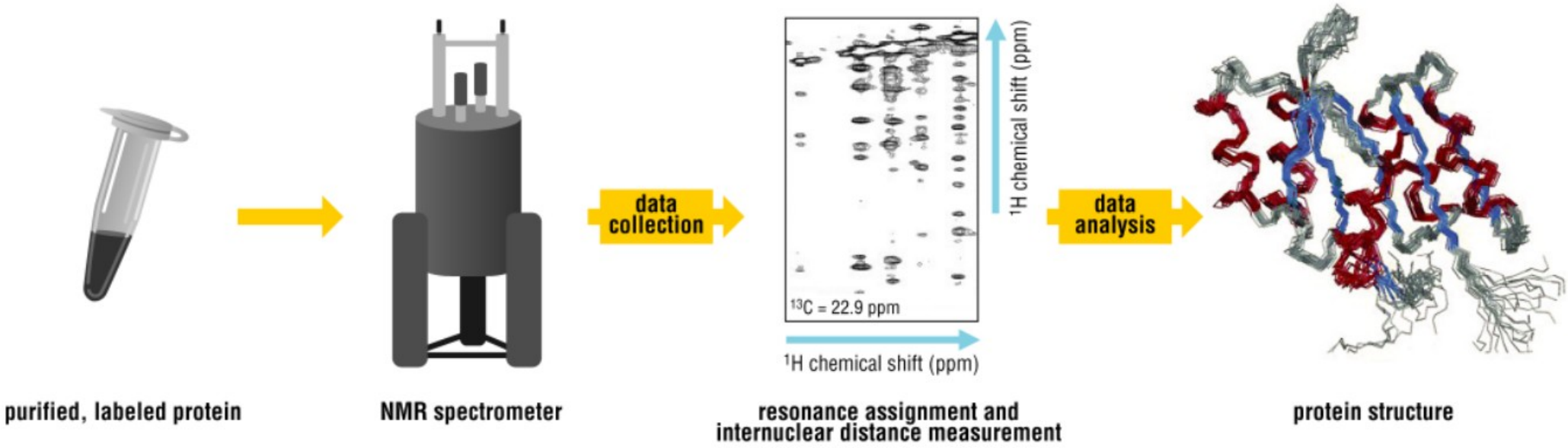
Where do these structures come from?



- Alacsony hullámhossz (ca. 1.5 Å),
- Brag egyenleg
- Elektron sűrűslget kapunk
- Crystallization artefaktok
- Nem fiziológiás
- Hidrogének nem látszanak



^{1}H-NMR (Proton Nuclear Magnetic Resonance)



NMR (13%)

- Oldatban
- Egy csokor szerkezetet kapunk amik a megfeelő kényszereket kielégítik
- Csak kis molekulák (> 35kDa)
- Kevésbé pontos
- Flexibilis molekulák is vizsgálhatóak

Elektron mikroszkóp (1%)

TEM (Transmission Electron Microscopy)

- Sub-angstrom felbontásra is képes
- Biológiai mintákra nem használható
- Két módszer született ami kiküszöböli a problémákat

Biológiai minték

- Magas vákum
- Sugárzás rezisztencia
- Bio atomok elektron szórása nagyon hasonló az oldószeréhez

Cryo – EM

- Normál körülmények között
- GYORS fagyasztás
- Víz nem kristályosodik → Vitrifikált víz

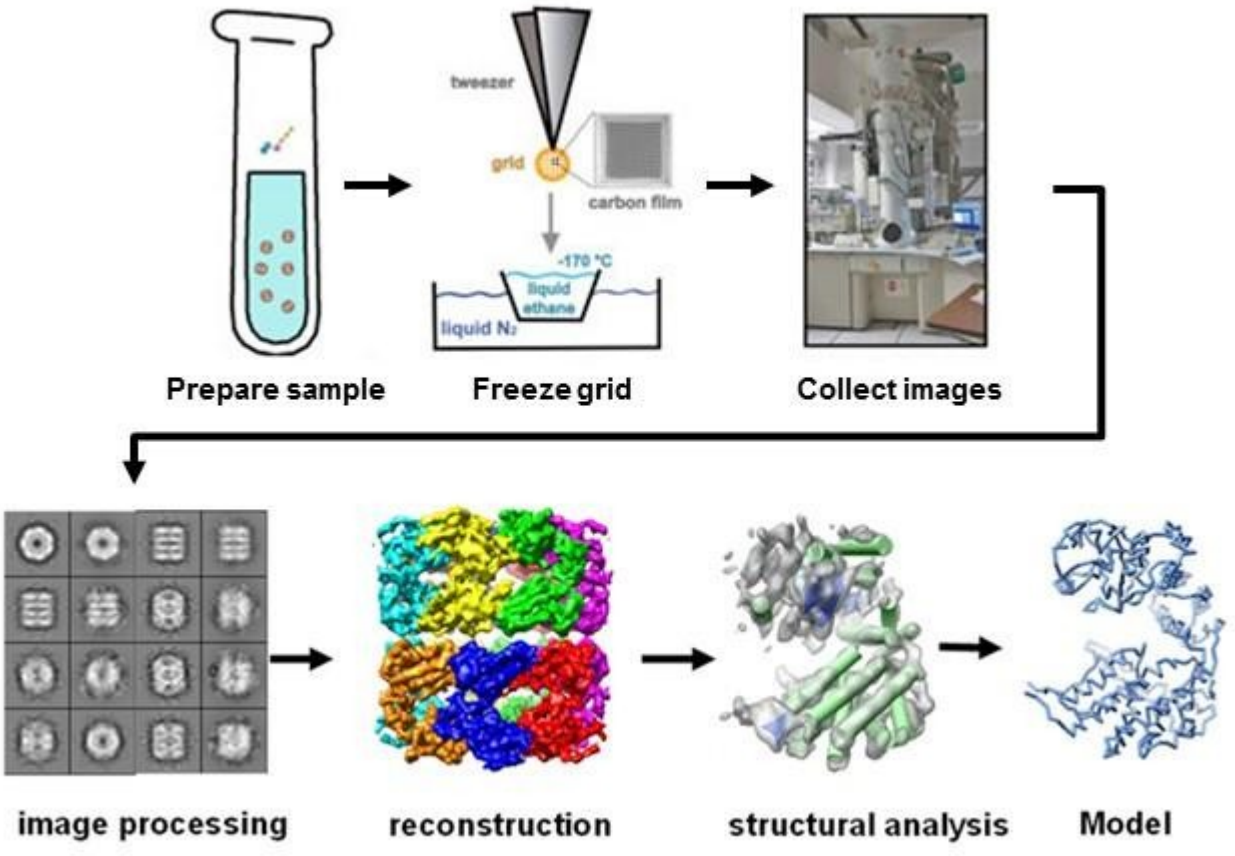
Vákum rezisztrens ✓

Sugárzás rezisztens ✓

Kontraszt ✗

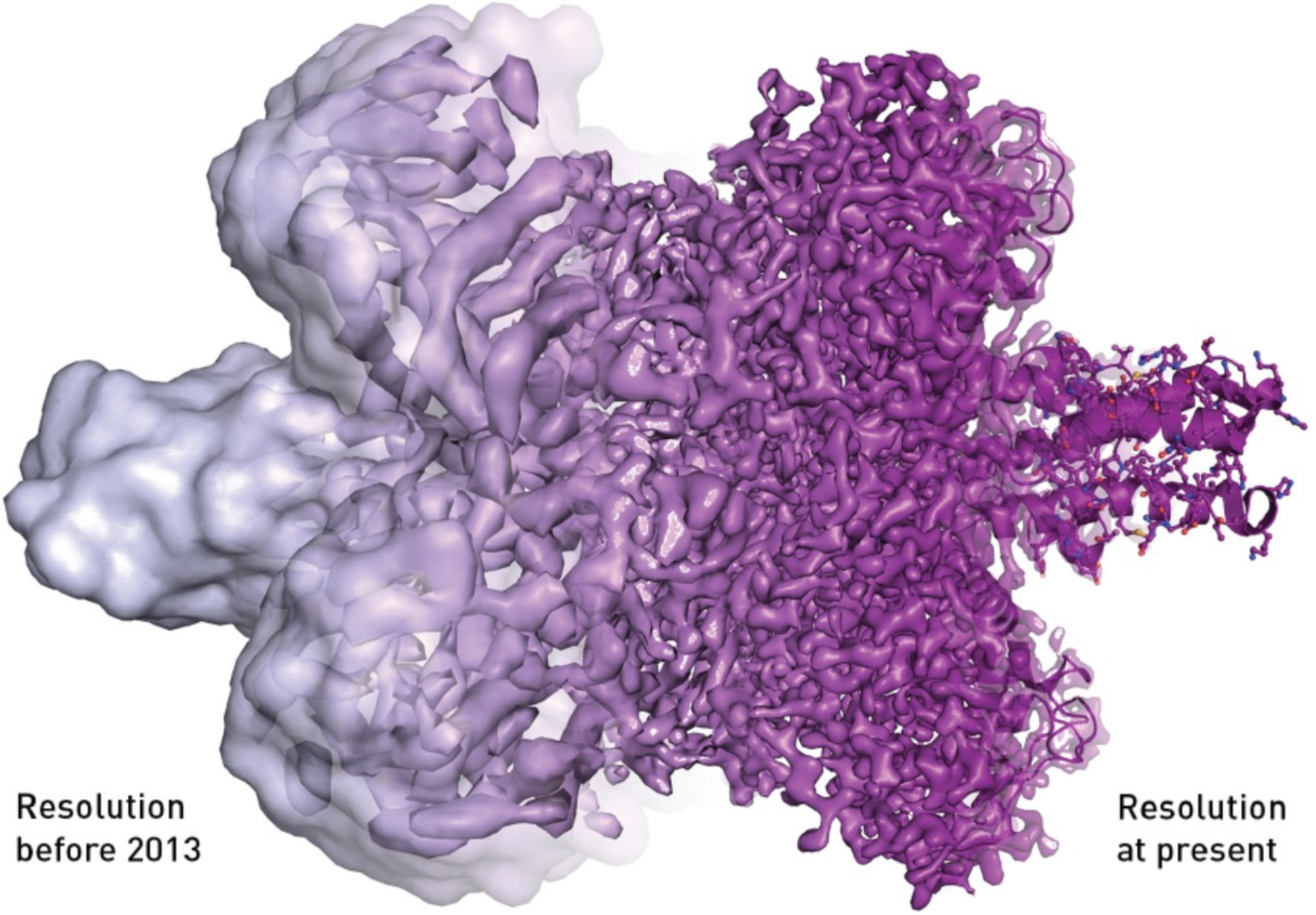
Negative Staining

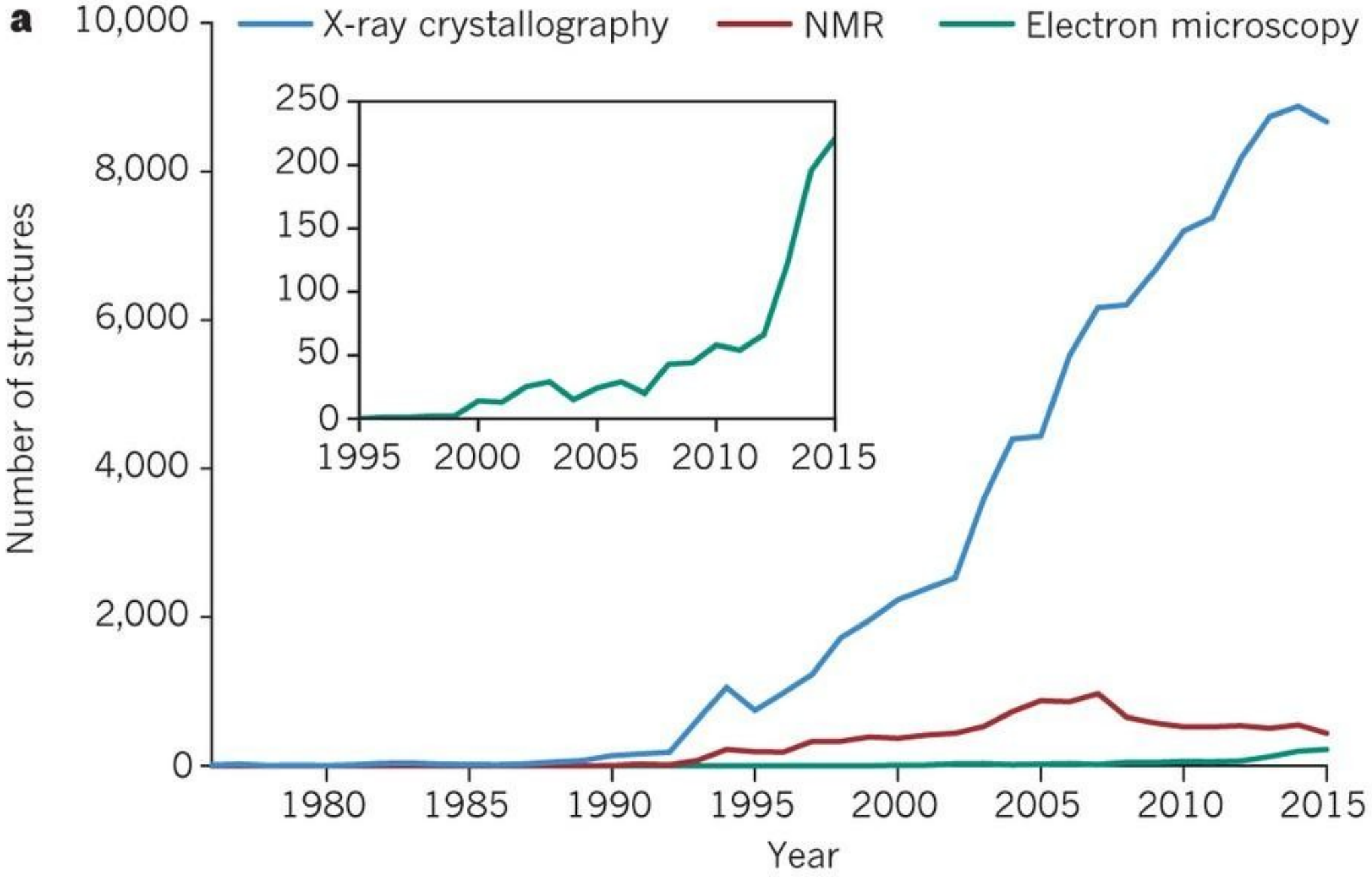
- A mintát egy nehés sóval kristályosítják (uranium)
- Vékony rétegben felviszik egy felületre



https://www.ibiology.org/ibioseminars/techniques/eva-nogales-part-1.html

Nobel Prize in Chemistry 2017




a
X-ray crystallography
NMR
Electron microscopy
Number of structures
10,000
8,000
6,000
4,000
2,000
0
1980
1985
1990
1995
2000
2005
2010
2015
Year
250
200
150
100
50
0
1995
2000
2005
2010
2015

Hogyan találunk meg egy szerkezetet?

A PDB-ben minden SZERKEZETNEK külön egyedi azonosítója van. 4 karakter, az első mindig egy szám, a többi 3 pedig bármi lehet

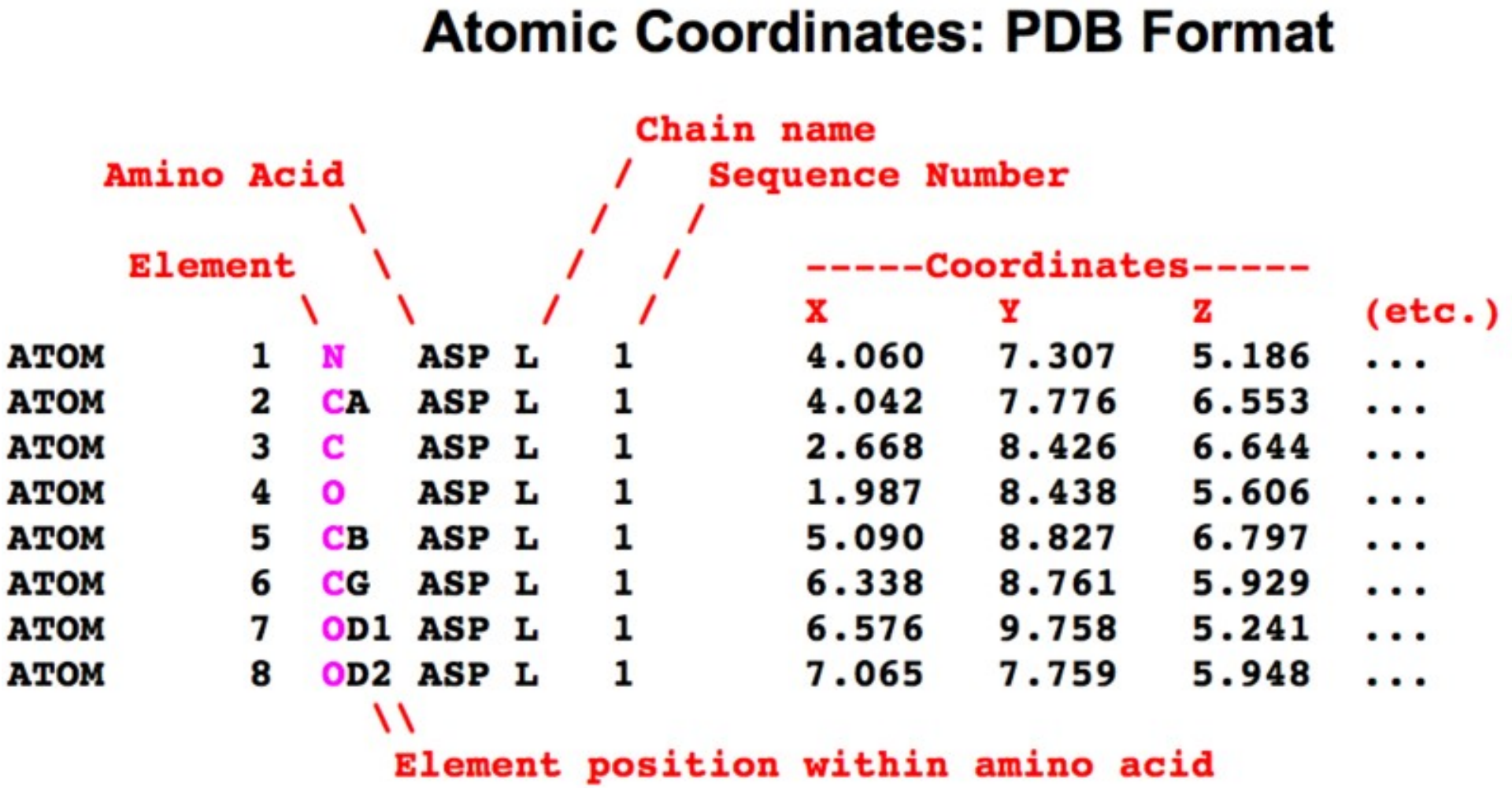
Például:

- 1mbn - The very first structure from 1973, the **myoglobin**
- 1tna - Form 1975 yeast phenylalanine transfer RNS, the first RNS structure
- 1bna - The first DNA double helix solved in 1980 with X-ray, (confirmation of the Watson & Crick modell from 1953).
- 2hhd - human **hemoglobin,** (deoxy form)
- 9ins - **insulin**

A .pdb file formátum

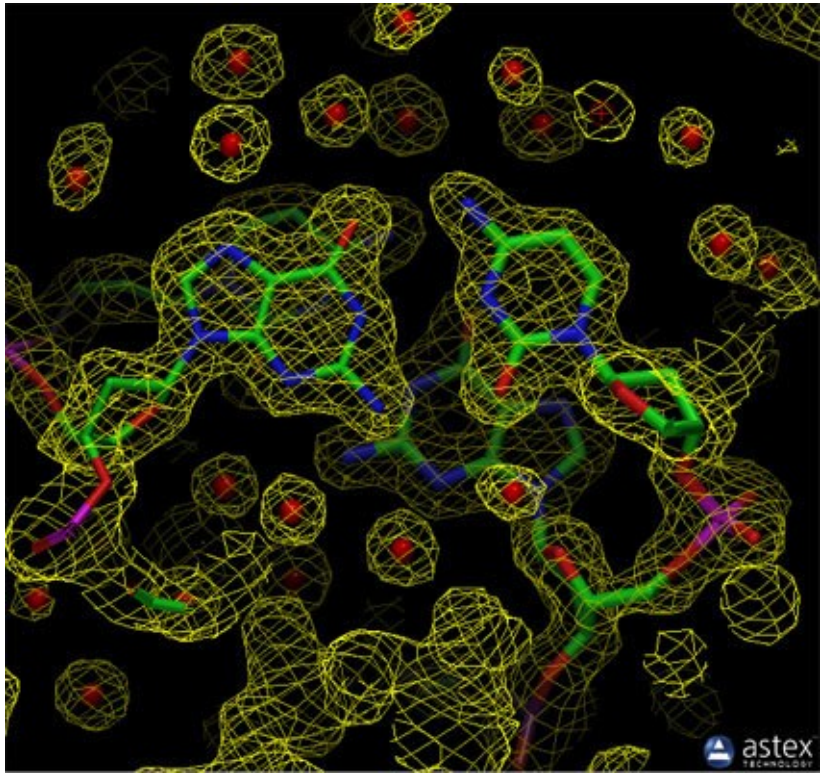
```
HEADER    EXTRACELLULAR MATRIX                      22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

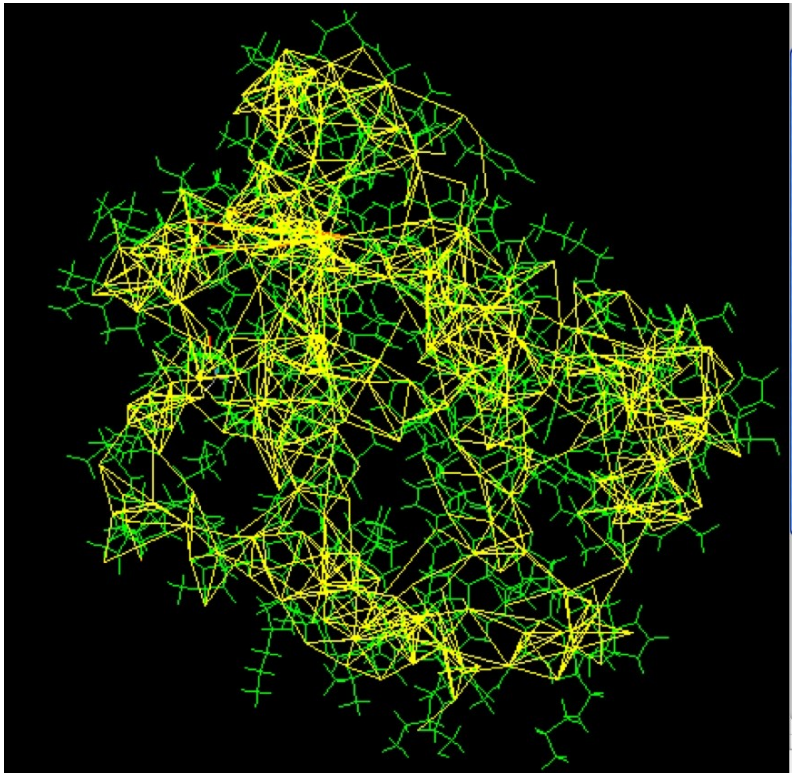
A .pdb file formátum



Model

A kísérletesen kapott szerkezetek nem közvetlenül a valóságot írjak le. Ezek a modellek, amik többé kevésbé pontosak, pontatlanok lehetnek.



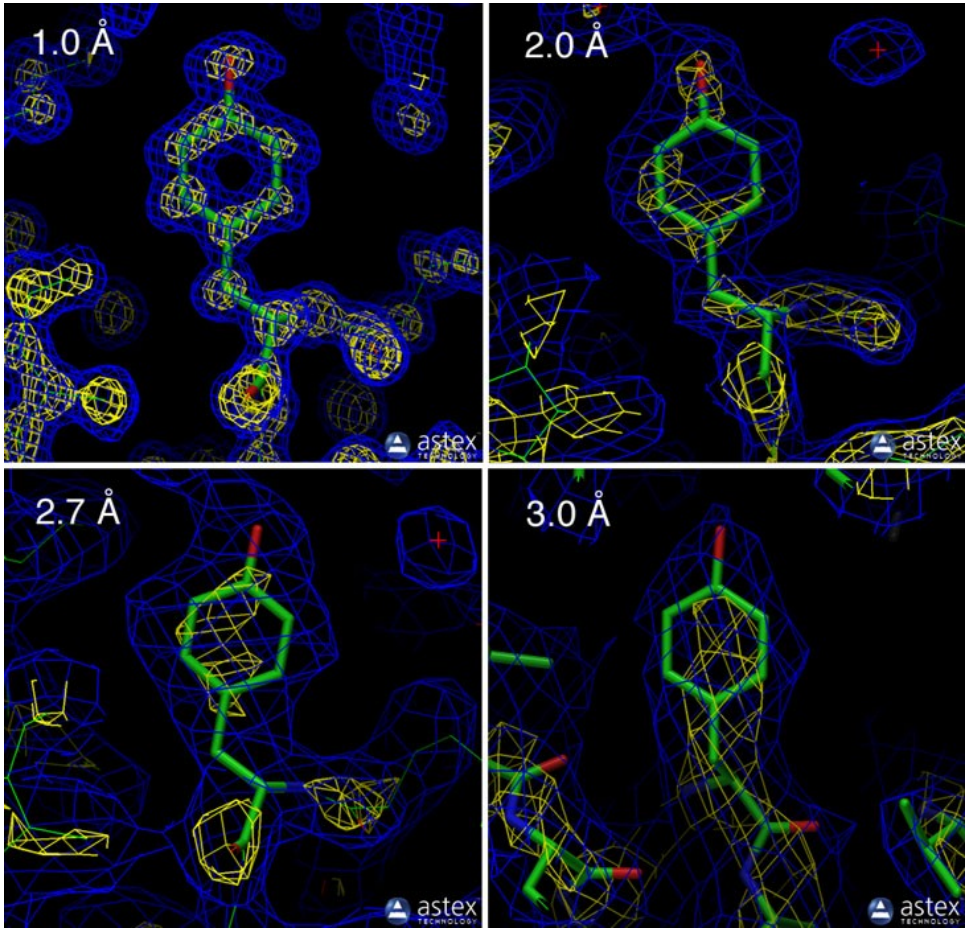


Felbontás

- Mennyire megbízható egy szerkezet

Alacsony: <3.0A
Átlagos: 1.8-3.0A
Jó: 1.0 – 1.8A
Atomi: >1.0A

A felbontás a szerkezet egyes részei között változhat!



Resolution

Resolution	Meaning
0.5 - 1.5	In general, structures have almost no errors at this resolution. Structures used for Rotamer libraries and geometry studies.
1.5 - 2.0	Few residues have wrong rotamer. Many small errors can normally be detected. Folds are extremely rarely incorrect.
2.0 - 2.5	Fold likely correct, low proportion of sidechains in wrong rotamer. Many small errors can normally be detected. Fold normally correct and number of errors in surface loops is small.
2.5 - 3.0	Fold likely correct except that some surface loops might be mismodelled. Several long, thin sidechains (lys, glu, gln, etc) and small sidechains (ser, val, thr, etc) likely to have wrong rotamers.
3.0 - 4.0	Fold possibly correct, but errors are very likely. Many sidechains placed with wrong rotamer.
>4.0	Individual coordinates meaningless

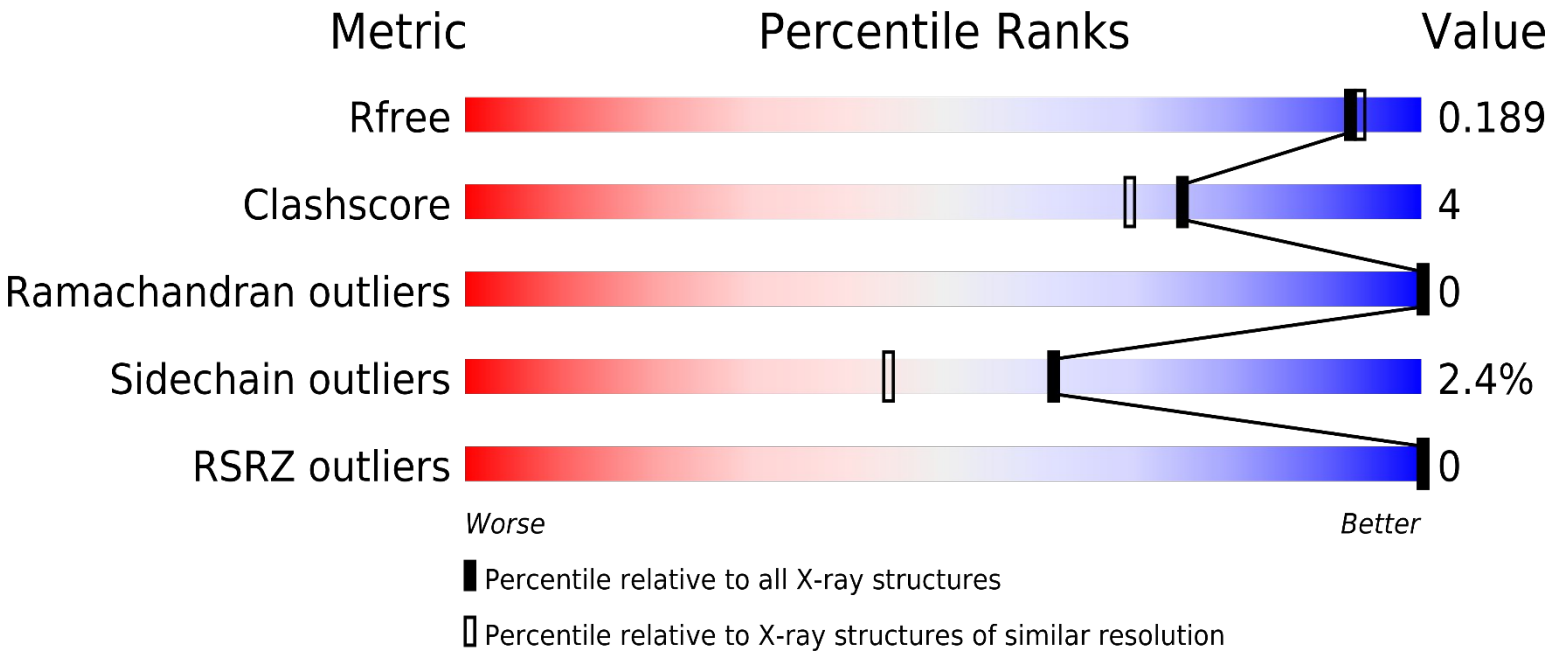
Mennyire jó egy szerkezet?

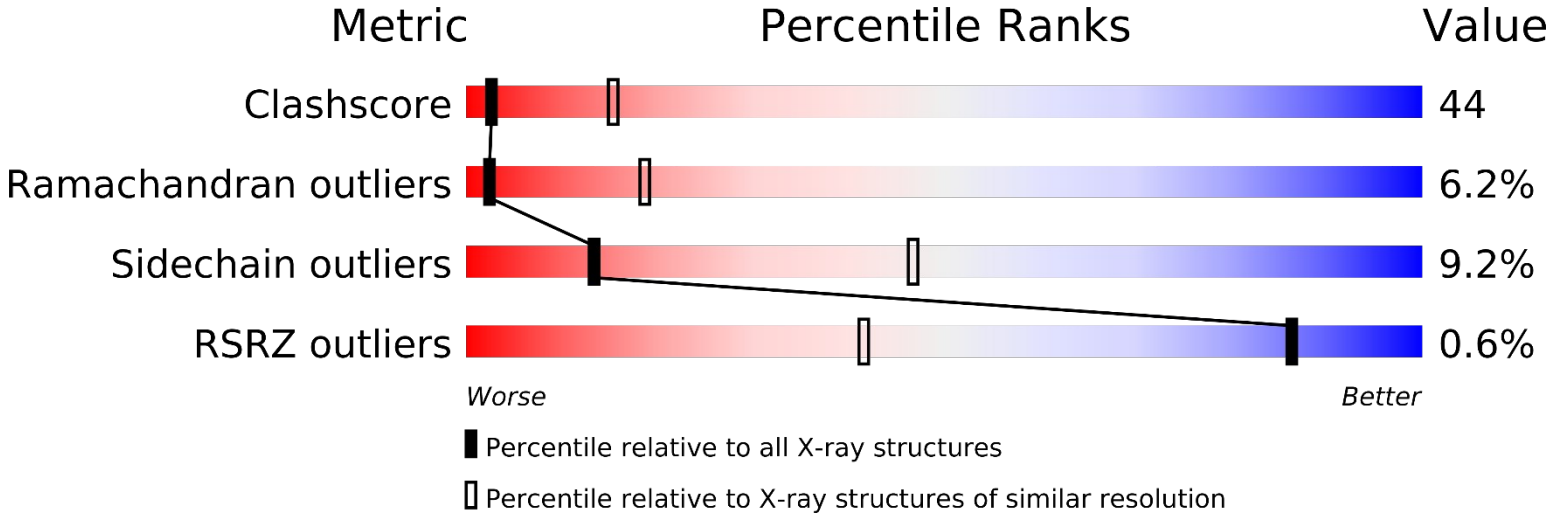
Felbontás? (In case of X-ray)

R-factor/FreeR-factor (X-ray)
Mennyire illik az adat korábbi ismert adatokra?

NMR mérésekre nincs standard

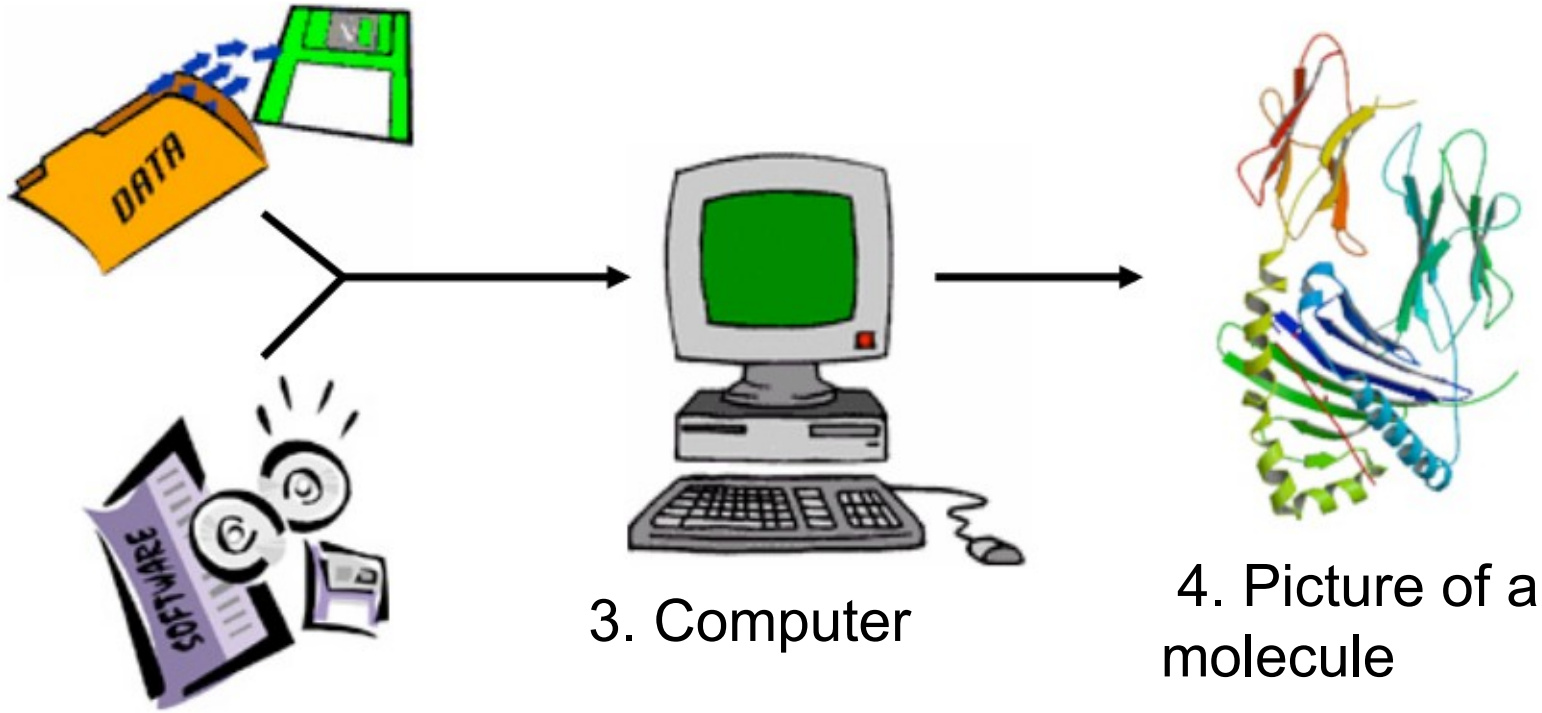
Ramachandran Plot
Clashes


Metric
Percentile Ranks
Value
Rfree
0.189
Clashscore
4
Ramachandran outliers
0
Sidechain outliers
2.4%
RSRZ outliers
0
Worse
Better
Percentile relative to all X-ray structures
Percentile relative to X-ray structures of similar resolution


Metric
Percentile Ranks
Value
Clashscore
44
Ramachandran outliers
6.2%
Sidechain outliers
9.2%
RSRZ outliers
0.6%
Worse
Better
Percentile relative to all X-ray structures
Percentile relative to X-ray structures of similar resolution

Visualization of structures

1. PDB file



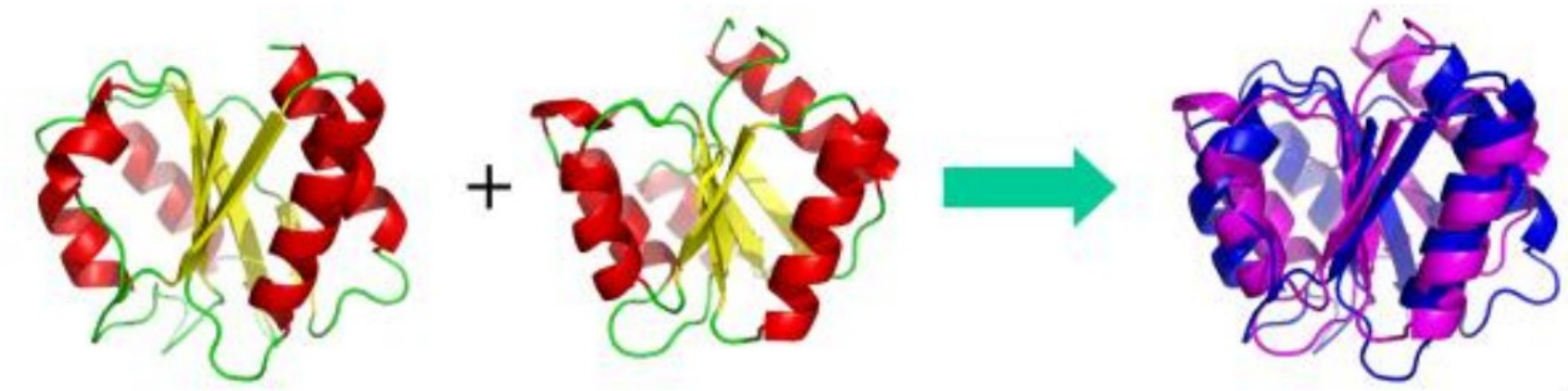
3. Computer

4. Picture of a molecule

2. Program for visualization

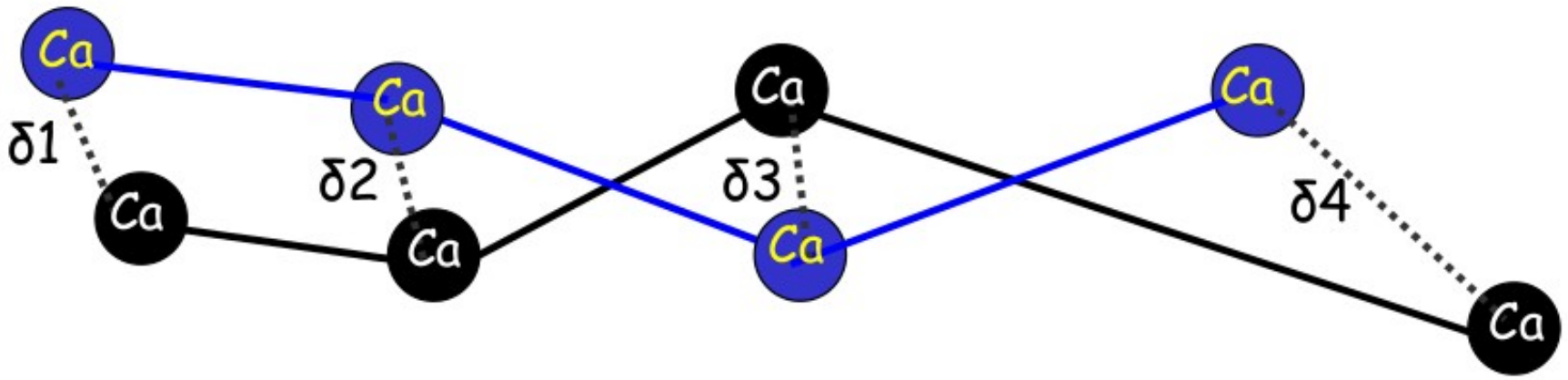
F.E.: Rasmol, Pymol, Chimera,
VMD, Jmol, Swiss PDB viewer

Mennyire hasonlóak szerkezetek?



Superposition: Minimizing the distance of positions

RMSD



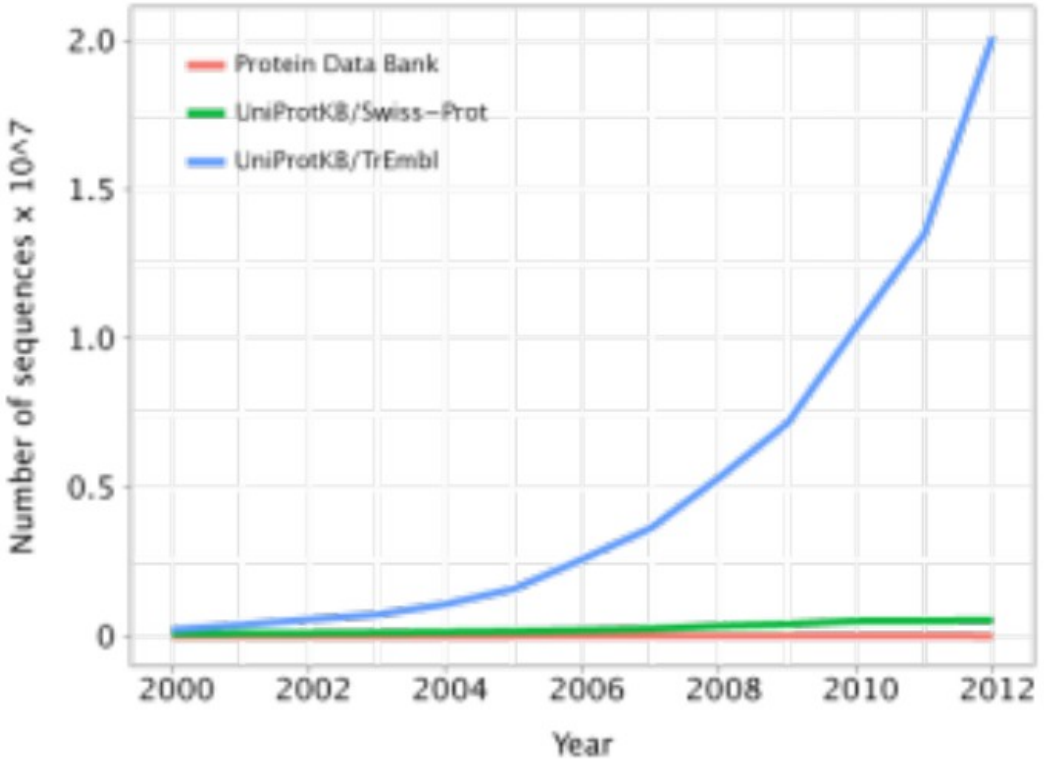
Root Mean Square Deviation (RMSD):

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2}$$

The most common function used to measure structural similarity

RMSD is the average distance between equivalent atoms of superimposed proteins (generally Cα).

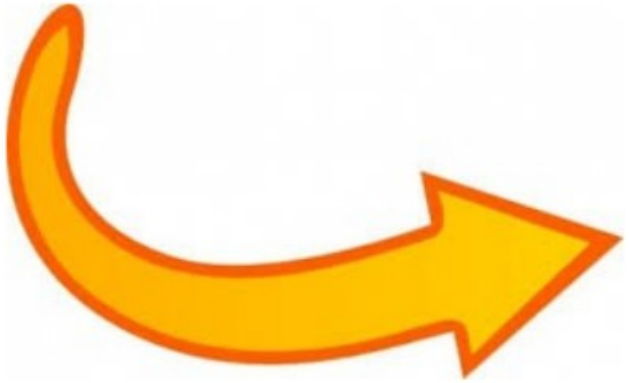
Szekvencia-szerkezet lyuk

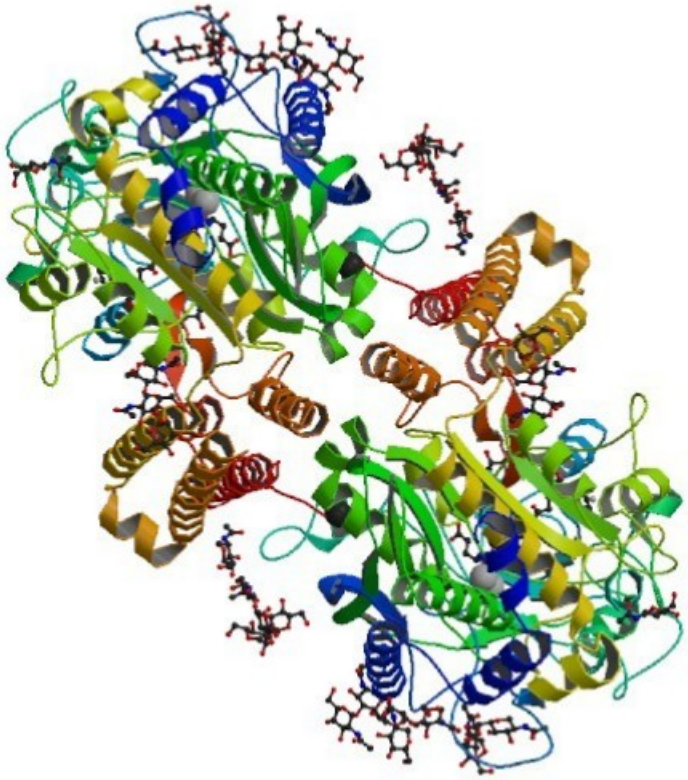


	2014
Sequences	50,000,000
Structures	100,000

Structure prediction

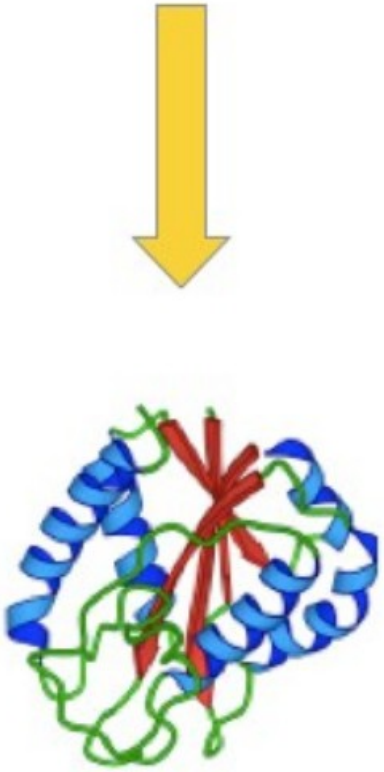
```
>Protein
RSKSSNEATNITPKHNMKAFLDELKAENIKKFLYNFTQIPHLAGTEQNFQLAKQIQSQWKEFGLDSVELAHYDVLLSYPN
KTHPNYISIINEDGNEIFNTSLFEPPPPGYENVSDIVPPFSAFSPQGMPEGDLVYVNYARTEDFFKLERDMKINCSGKIV
IARYGKVFRGNKVKNAQLAGAKGVILYSDPADYFAPGVKSYPDGWNLPGGGVQRGNILNLNGAGDPLTPGYPANEYAYRR
GIAEAVGLPSIPVHPIGYYDAQKLLEKMGGSAPPDSSWRGSLKVPYNVGPGFTGNFSTQKVKMHIHSTNEVTRIYNVIGT
LRGAVEPDRYVILGGHRDSWVFGGIDPQSGAAVVHEIVRSFGTLKKEGWRPRRTILFASWDAEEFGLLGSTEWAEENSRL
LQERGVAYINADSSIEGNYTLRVDCTPLMYSLVHNLTKELKSPDEGFEGKSLYESWTKKSPSPEFSGMPRISKLGSGNDF
EVFFQRLGIASGRARYTKNWETNKFSGYPLYHSVYETYELVEKFYDPMFKYHLTVAQVRGGMVFELANSIVLPFDCRDYA
VVLRKYADKIYSISMKHPQEMKTYSVSFDSLFSAVKNFTEIASKFSERLQDFDKSNPIVLRMMNDQLMFLERAFIDPLGL
PDRPFYRHVIYAPSSHNKYAGESFPGIYDALFDIESKVDPSKAWGEVKRQIYVAAFTVQAAAETLSEVA
```





Protein folding

GFCHIKAYTRLIMVG…



Folding

(physics)

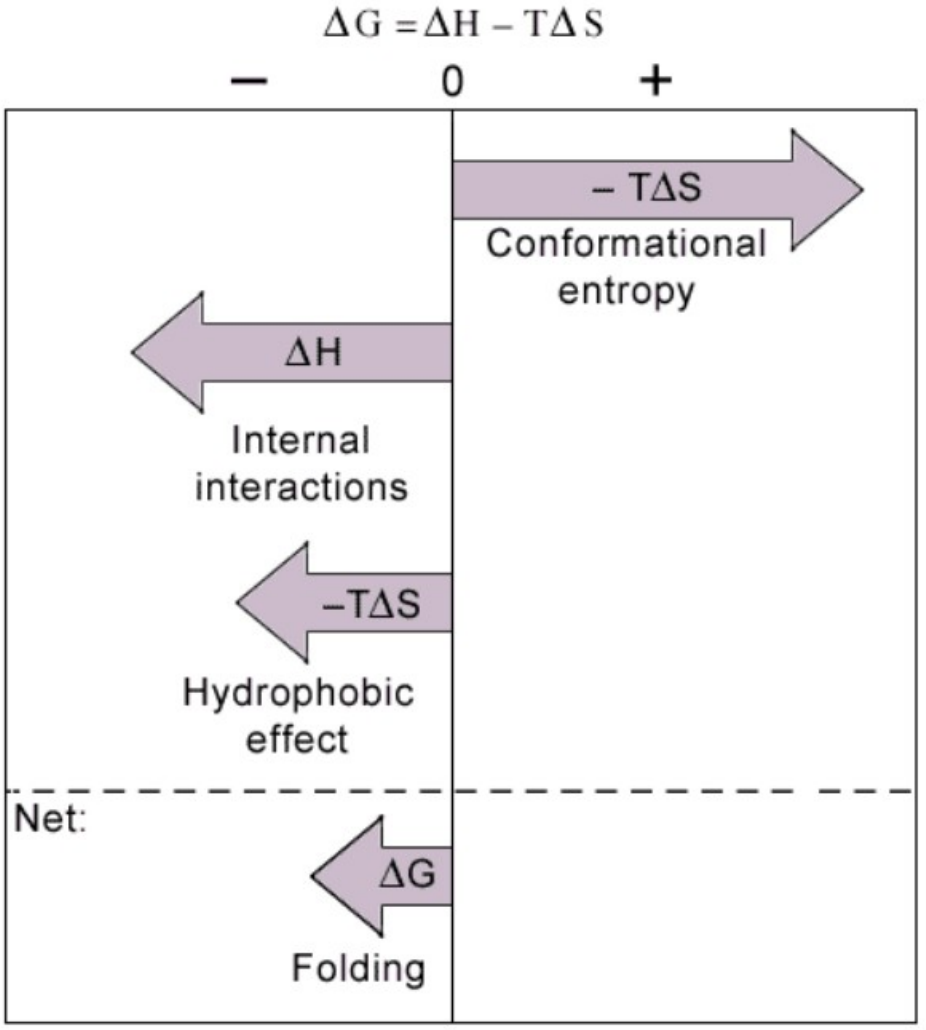
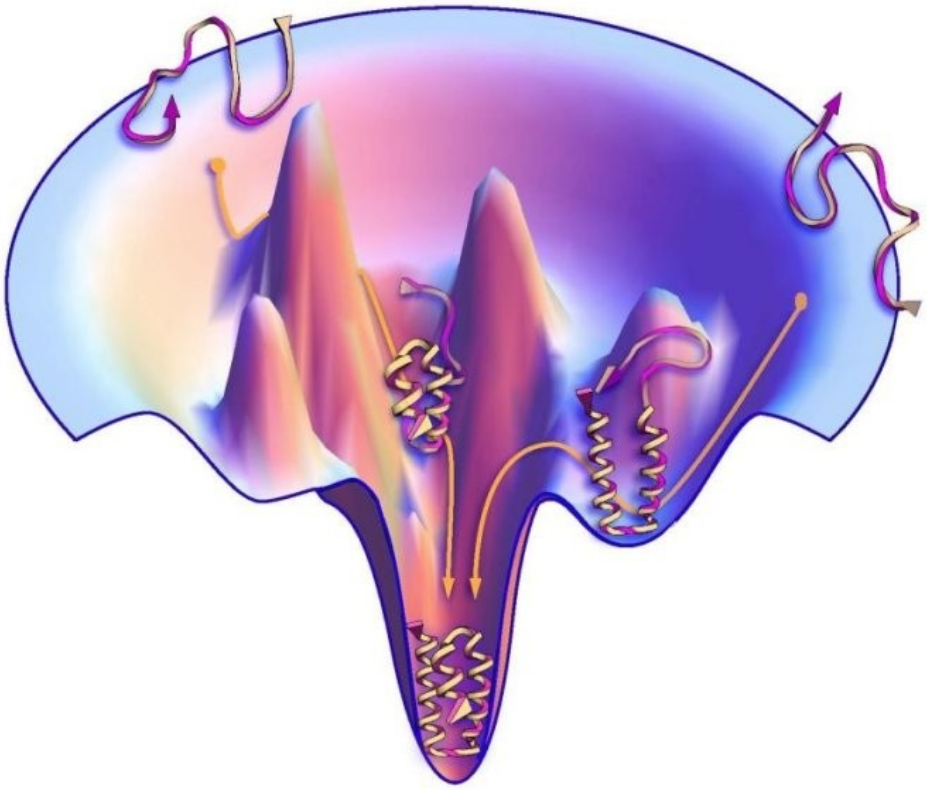


Figure 6.22, C.K. Mathews & K.E. van Holde, *Biochemistry*, 2nd edition (1996)

Fehérje folding

- A feháérjék szerkezete nem statikus
- Egy fehérje átmehet különböző konformációkba
- Szerkezet mindig a legalacsonyabb állapot felé mozog
- Lokális minimumok lassítják a feltekeredést



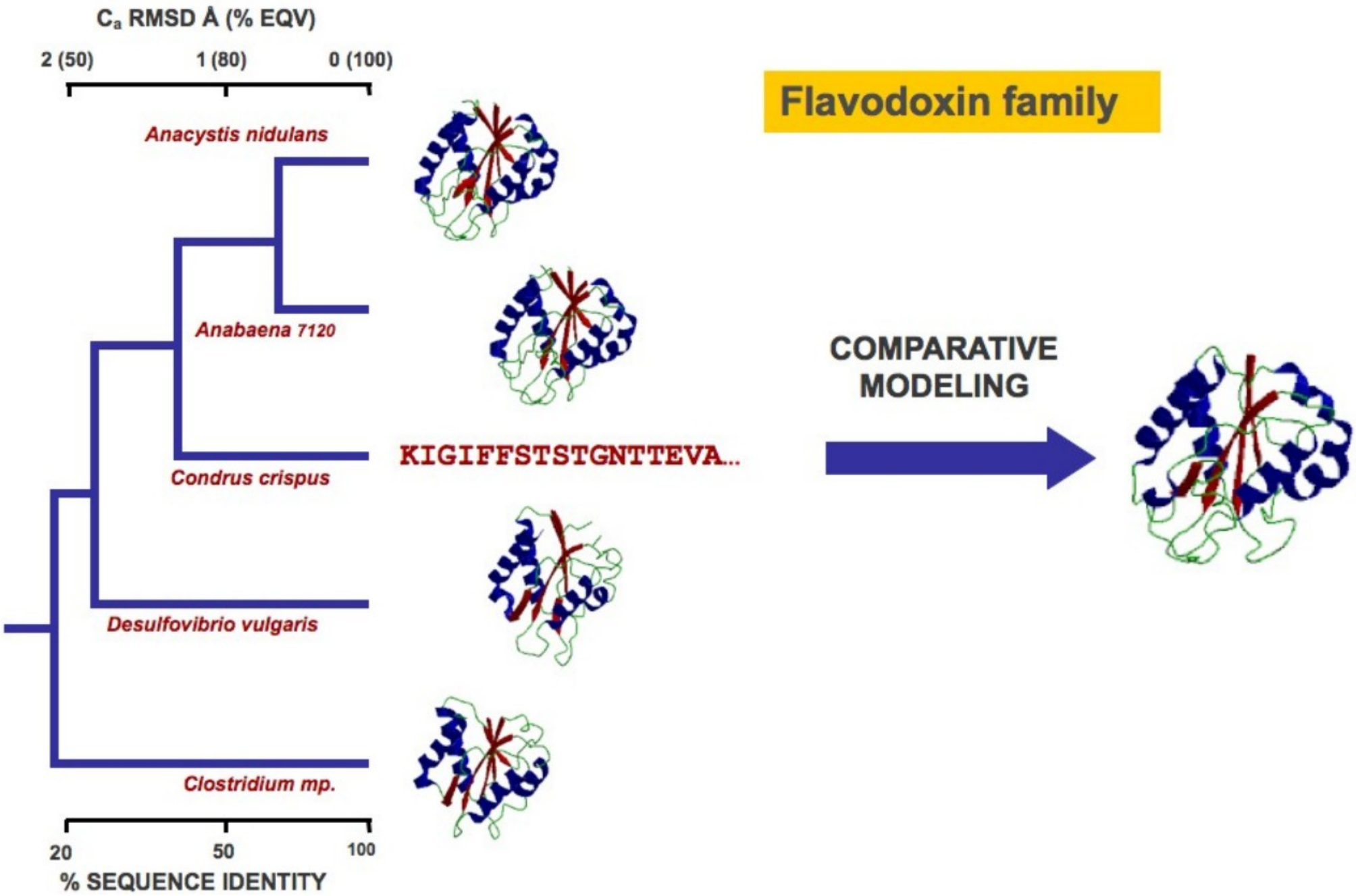
Fizikai modellek

- Rengetek konformáció, hatalmas konformációs tér
- A fizikai energia függvény rendszer (force field) nem teljesen ismert

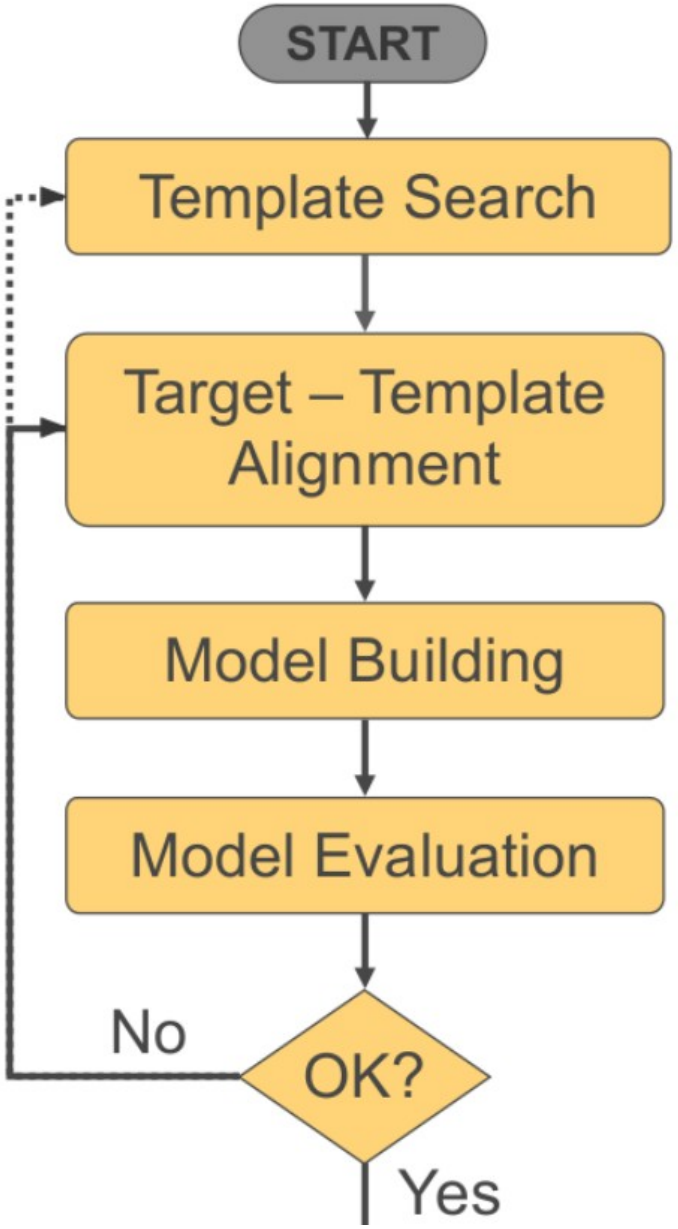
Szerkezet predikció

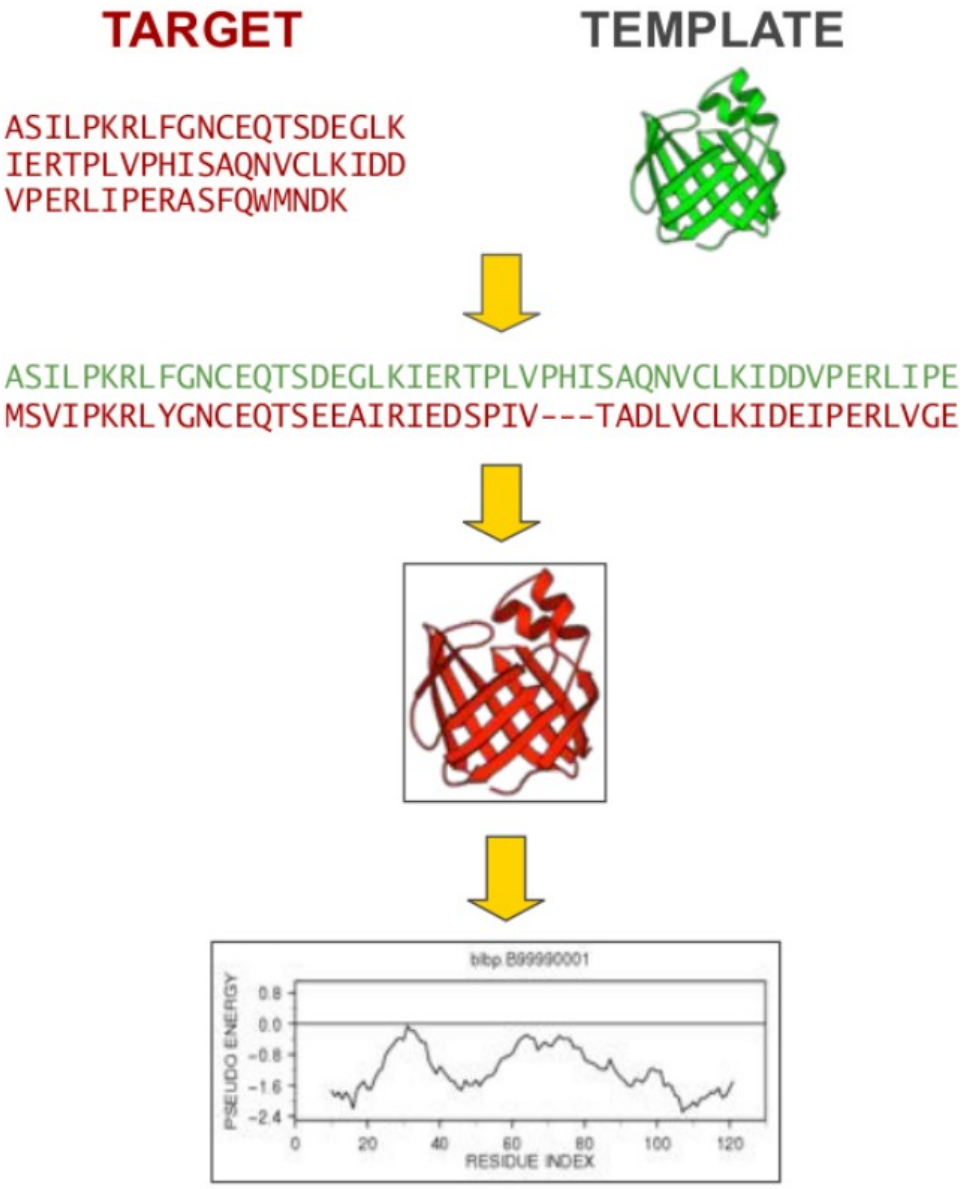
- Homologia modellezés
- (Threading)
- Ab-initio modellezés

Comparative Protein Structure Modeling



Steps of homology modelling

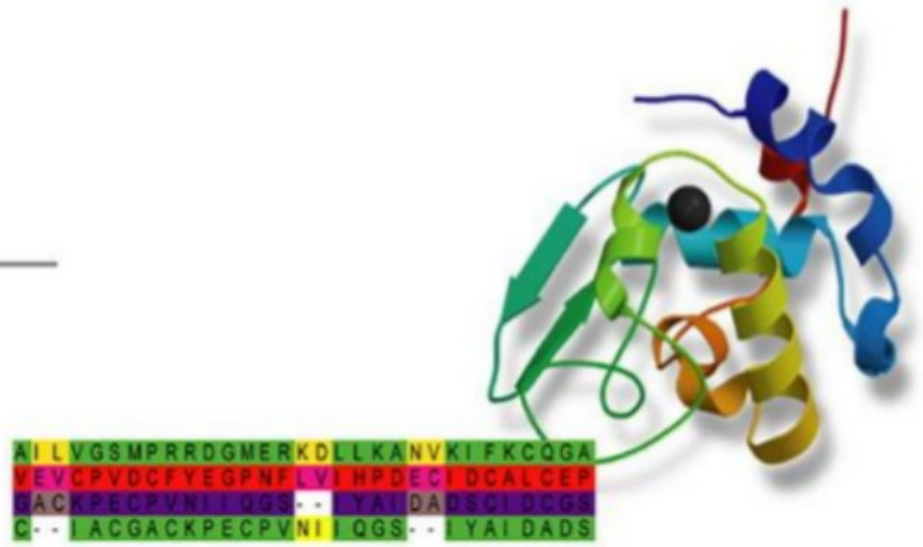




Modellező csomagok

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



SIB Swiss Institute of Bioinformatics | BIOZENTRUM Universität Basel | SWISS-MODEL

Modelling

myWorkspace

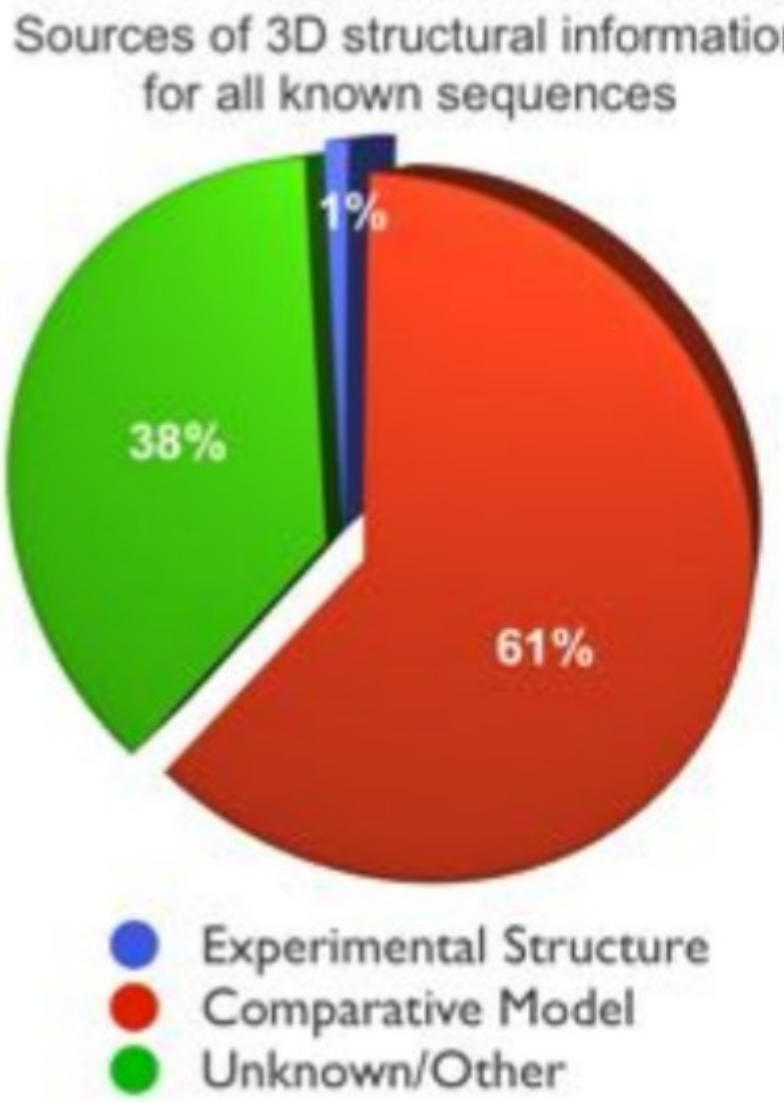
Automated Mode

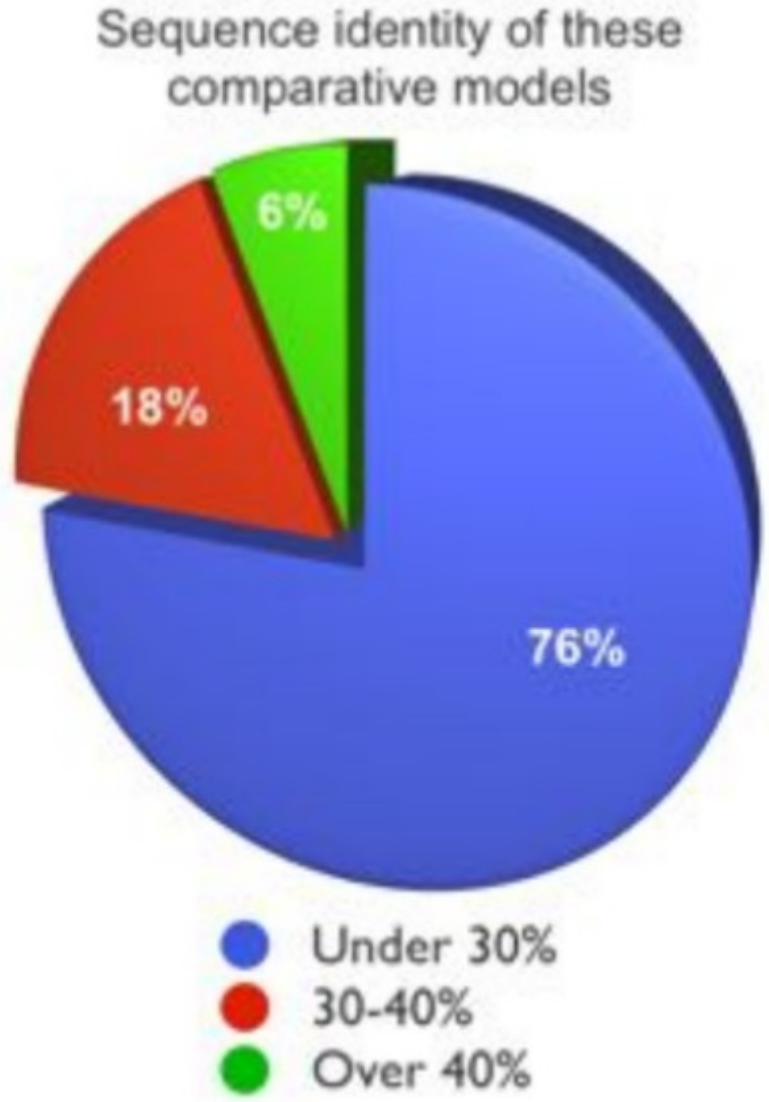
Alignment Mode

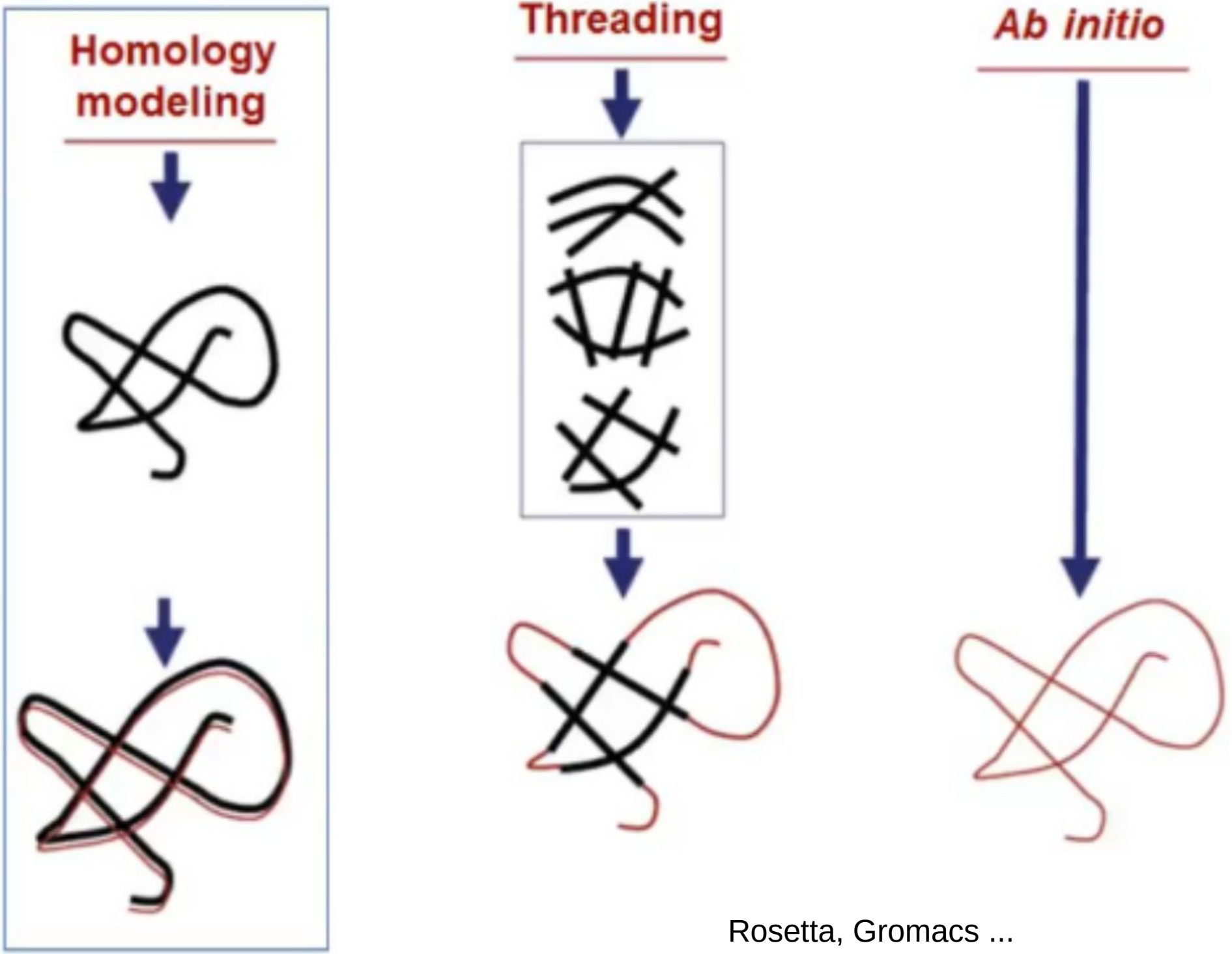
SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists worldwide.

SWISS-MODEL Team

Torsten Schwede:	Project Leader
Florian Kiefer:	SWISS-MODEL Repository
Lorenza Bordoli:	Method Development and user support
Konstantin Arnold:	SWISS-MODEL Workspace

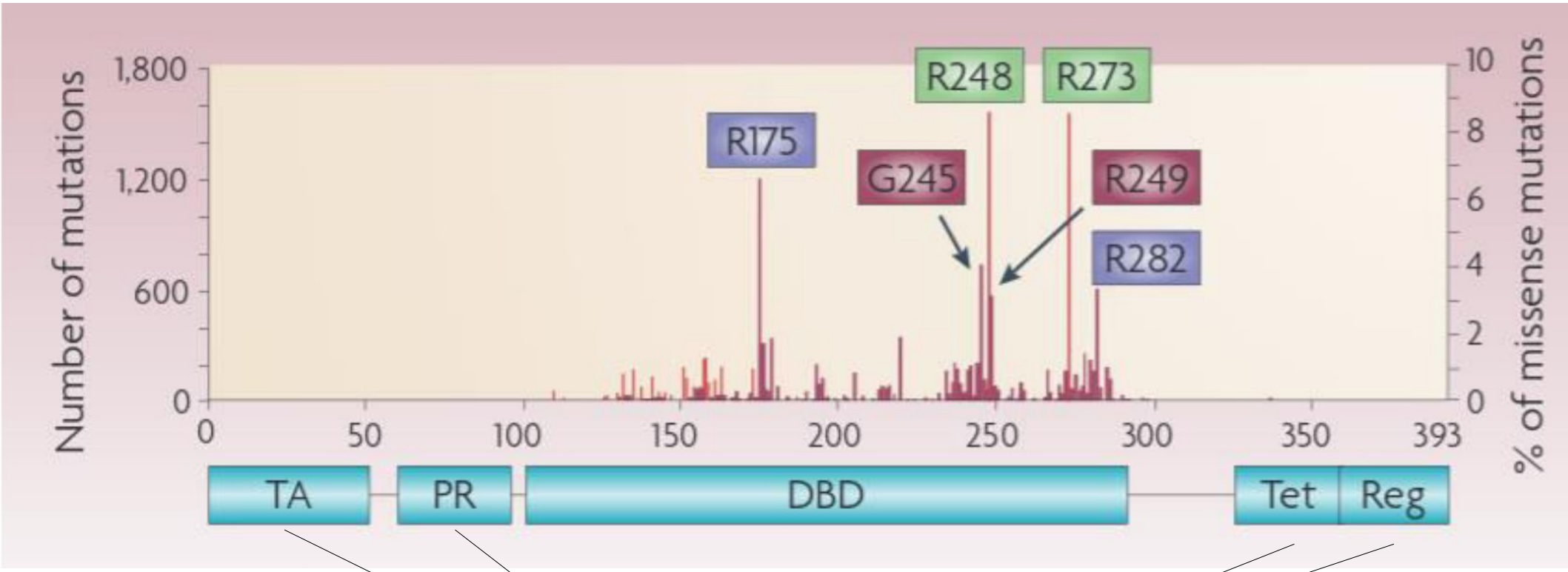

Sources of 3D structural information for all known sequences
1%
61%
38%
Experimental Structure
Comparative Model
Unknown/Other


Sequence identity of these comparative models
6%
18%
76%
Under 30%
30-40%
Over 40%

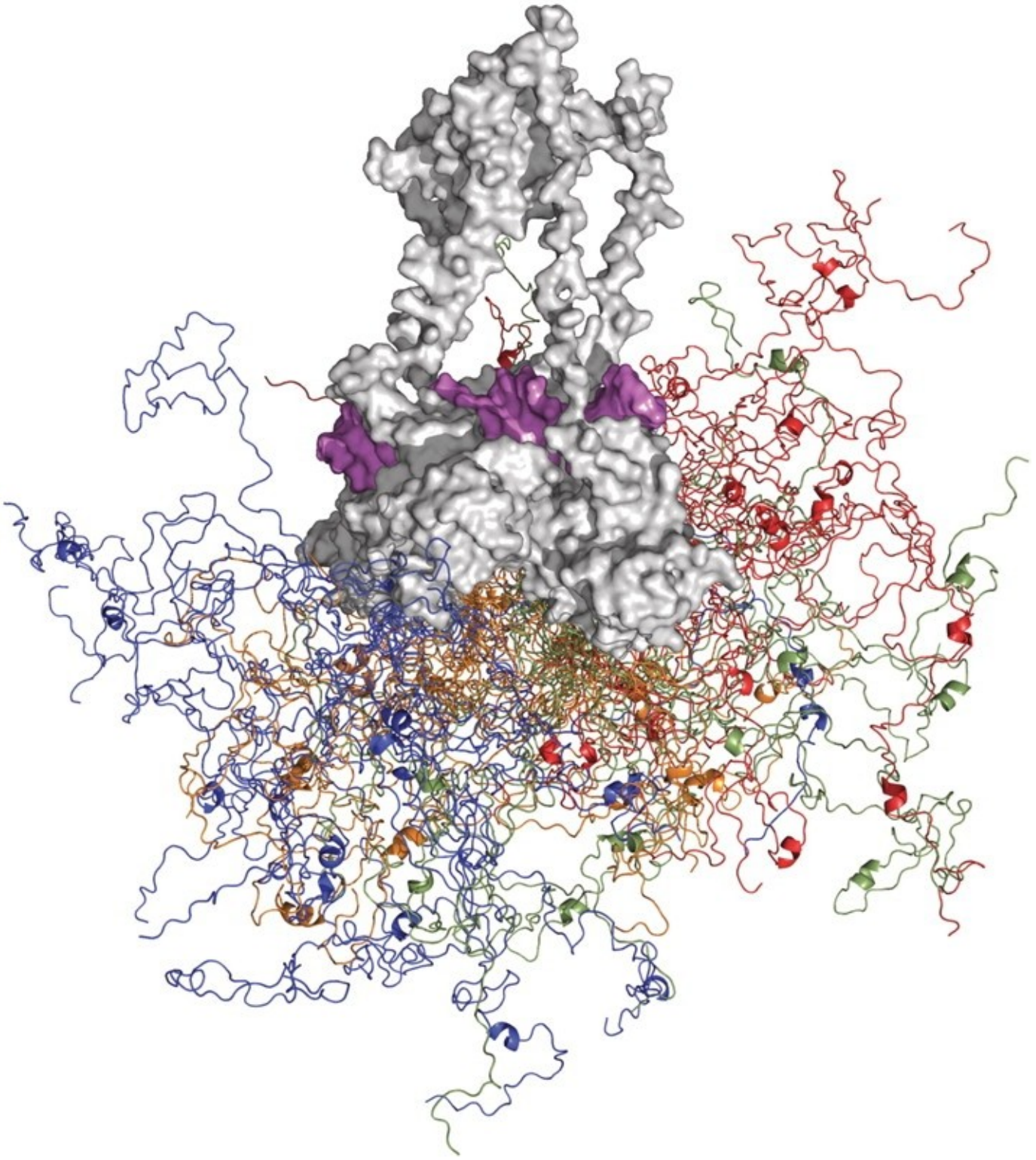


Rosetta, Gromacs ...

P53 fehérje



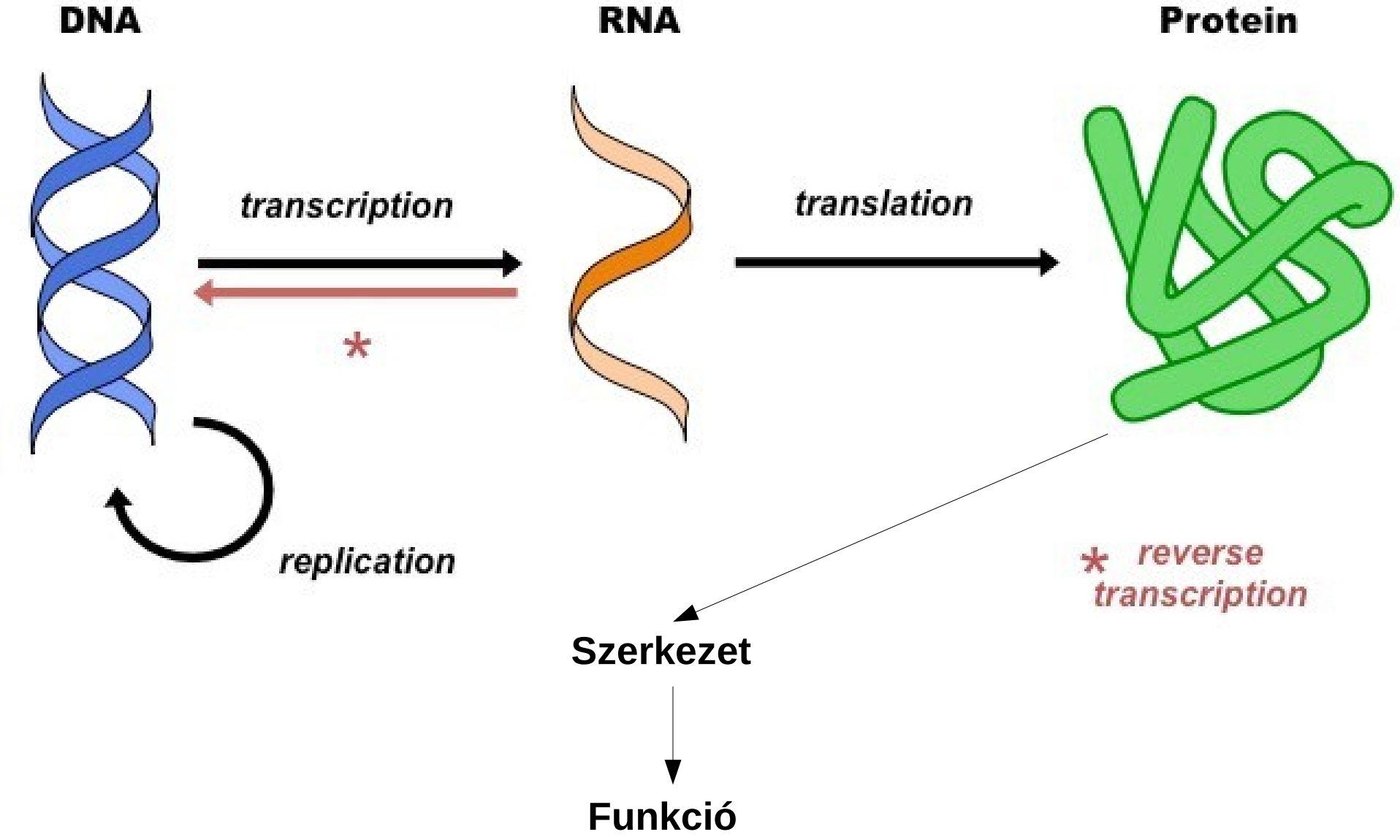
Ezekről a részekről csak olyan szerkezet van, ahol kölcsönhategy másik fehérjével.
Miért?



Structure function paradigm

- Nem mindig igaz
- Azon fehérjék is lehetnek funkcionálisak amelyek nem rendelkeznek jól definiált háromdimenziós szerkezettel
- A humán proteom több mint 50%-a teljesen, vagy részben rendezetlen
- Reguláció, apoptózis, jelátvitel...
- Fontos betegségekhez köthető
  - p53 → Rák
  - τ protein → Alzheimer
  - Synuclein → Parkinson

Molekuláris biológia centrális dogmája



Molekuláris biológia centrális dogmája