

# Bevezetés a szerkezeti bioinformatikába



**Gáspári Zoltán, 2019**

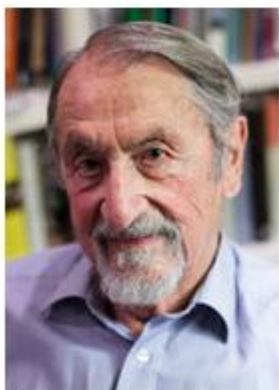
*gaspari.zoltan@itk.ppke.hu*



The Nobel Prize in Chemistry 2013

Martin Karplus, Michael Levitt, Arieh Warshel

# The Nobel Prize in Chemistry 2013



© Harvard University  
Martin Karplus



Photo: © S. Fisch  
Michael Levitt



Photo: Wikimedia  
Commons  
Arieh Warshel

The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel *"for the development of multiscale models for complex chemical systems"*.

But the achievements of bioinformatics in the field of structural biology are far more immense than the identification of the odd homolog, as has been recognized by the award of the 2013 Nobel prize in Chemistry to Michael Levitt, Arieh Warshel and Martin Karplus for achievements in computational modeling of complex chemical systems – especially notable for work on structural biology.

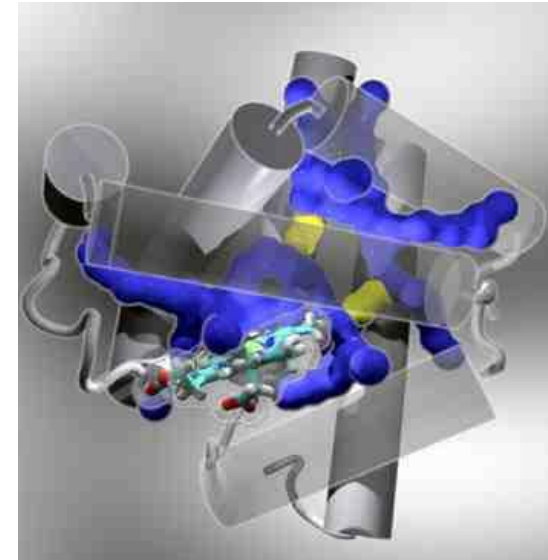
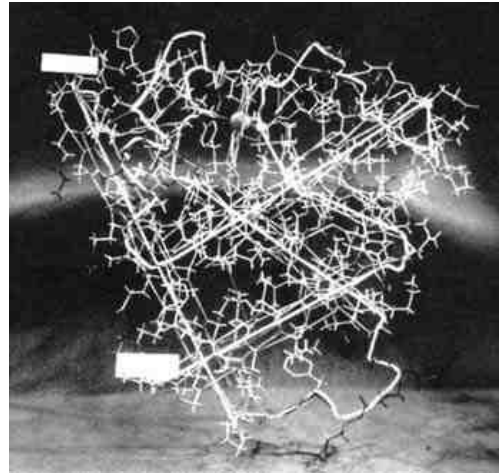
<http://blogs.biomedcentral.com>

Since the 1970s, Levitt has continuously been a strong pillar, visionary, and supporter of the growing field of computational biology. Like his co-laureates, Karplus and Warshel, he succeeded in the quest to move from the atomic and subatomic phenomena involved in chemical reactions to providing computational methods that are able to predict static and dynamical aspects of 3D structure for protein molecules with thousands to millions of atoms.

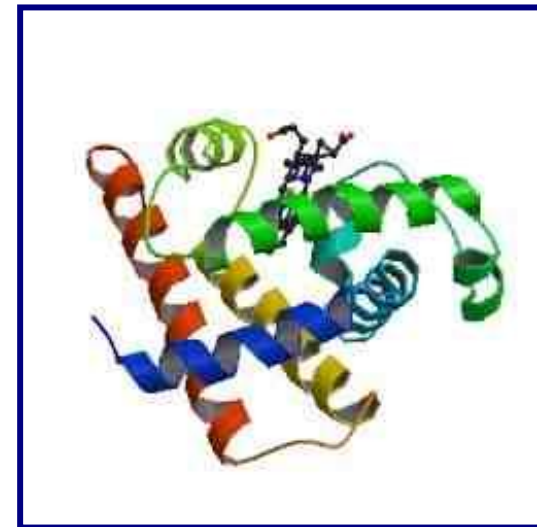
Levitt is an inspiring role model within the computational biology community who advocates for the good that science can bring to society. His enthusiasm and humility persist in the face of this recognition, as he stated in his own words, "The Nobel is recognition and it's fun to get it, and it's good for the university, it's good for Israel and it's good for our science. Levitt also told Associated Press, "It's sort of nice in more general terms to see that computational science, computational biology is being recognized." He added, "It's become a very large field and it's always in some ways been the poor sister, or the ugly sister, to experimental biology."

<http://www.iscb.org/>

# Minden fehérjeszerkezet modell!



a mioglobin  
'evolúciója'



- A szerkezetkutatásban a modell szót általában a kísérleti adatok nélkül, predikciós módszerekkel előállított szerkezetekre használják



# A fehérjék szerkezeti szintjei

## Elsődleges szerkezet (=szekvencia):

VDCSEYPKPACPKDYRPVCGSDNKTYSNKCNFCNAVV

ESNGTLTLNHFGKC

(turkey ovomucid inhibitor III domain )

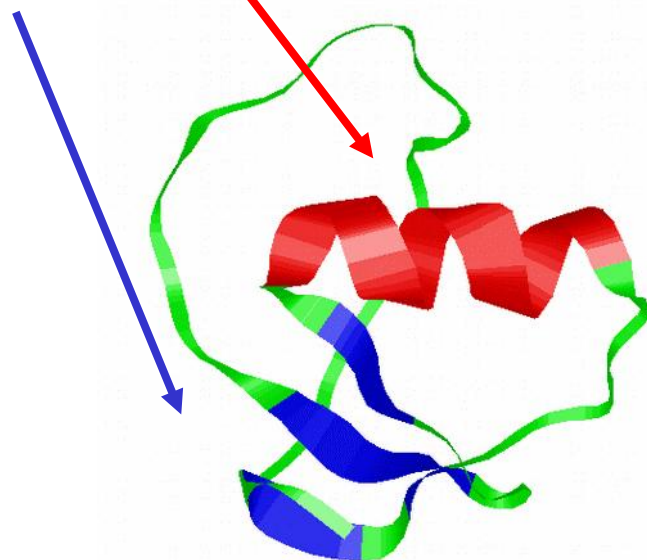
## Másodlagos szerkezet:

hurkok vagy  $\beta$ -kanyarok

$\alpha$ -hélix

$\beta$ -redő

A polipeptidlánc lokális konformációja



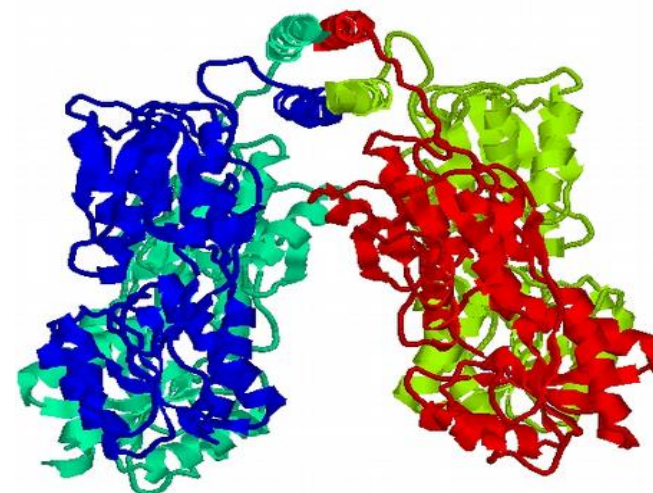
## Harmadlagos szerkezet:

a molekula 3D-szerkezete



## Negyedleges szerkezet:

alegységösszetétel



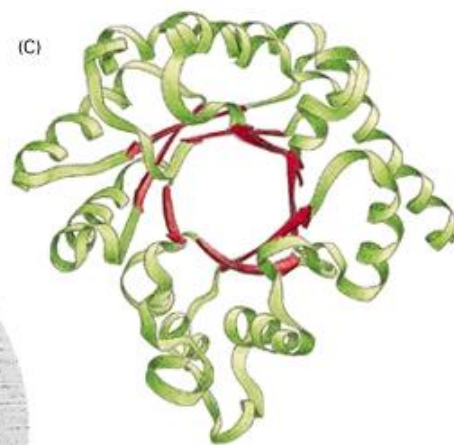
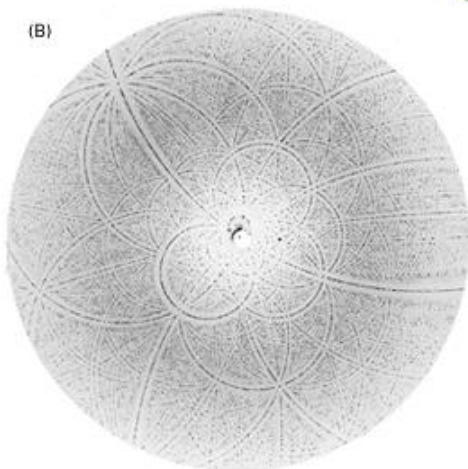
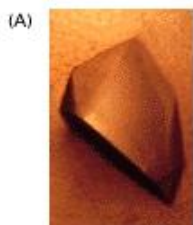
# A fehérjék térszerkezetének meghatározása

## röntgenkristallográfia

- Egy vagy néhány jól meghatározott konformer,
- A kristályban / adatbázisban nem feltétlenül a biológiailag releváns szerkezet/komplex van (biological assembly sok esetben külön megadva)



Kémiai Nobel-díj, 1962



fehérjekristály,  
diffrakciós kép és  
térszerkezet



Szinkrotron

# A fehérjék térszerkezetének meghatározása

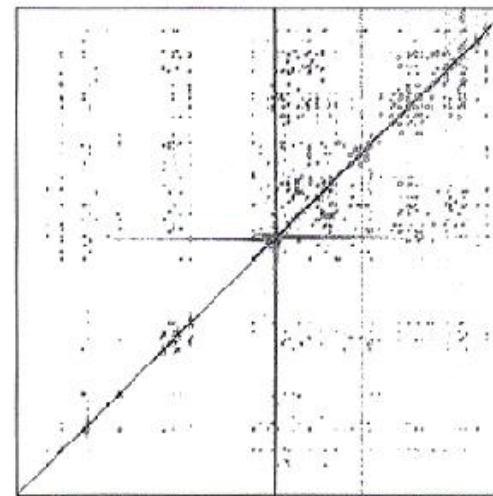
## NMR-spektroszkópia

- Nagyobb bizonytalanság: sok konformer vagy egy “átlagszerkezet” az adatbázisban
- Alapesetben nem tükrözi az oldatbeli dinamikát!



Kémiai Nobel-díj, 2002

NMR-spektrométer



(A)

2D NMR-spektrum és térszerkezeti sokaság



(B)



# A 'fehérjék állatkertje'


  
**PDB**
  
 PROTEIN DATA BANK
   
[rcsb.org](http://rcsb.org)

## Molecular Machinery: A Tour of the Protein Data Bank

Cells build many complex molecular machines that perform the biological jobs needed for life. Some of these machines are molecular scissors that cut food into digestible pieces. Others then use these pieces to build new molecules when cells grow or tissues need to be repaired. Some molecular machines form sturdy beams that support cells, and others are motors that use energy to crawl along these beams. Some recognize attackers and mobilize defenses against infection.

Researchers around the world are studying these molecules at the atomic level. These 3D structures are freely available at the Protein Data Bank (PDB), the central storehouse of biomolecular structures. A few examples from the ~100,000 structures held in the PDB are shown here, with each atom represented as a small sphere. The enormous range of molecular sizes is illustrated here, from the water molecule (H<sub>2</sub>O) with only three atoms (shown at the left) to the ribosomal subunits with hundreds of thousands of atoms.

### Digestive Enzymes: breaking food into small nutrient molecules

1. Amylase 1am1
2. Phospholipase 1lpe
3. Deoxyribonuclease 2dnq
4. Lysozyme 1lzl
5. Trypsin 3trp
6. Trypsin 2trp
7. Carboxypeptidase 3cpa
8. Ribonuclease 5rta

### Blood Plasma Proteins: transporting nutrients and defending against injury

9. Factor X 1dka, 1dod
10. Thrombin 1tpb
11. Fibrin 1fmj, 2baf
12. Serum Albumin 1e1r

### Viruses and Antibodies: engaging in constant battle in the bloodstream

13. Antibody 1igt
14. Rhinovirus 4rhv

### Hormones: carrying molecular messages through blood

15. Calcitonin 1gcn
16. Insulin 2ins
17. Epidermal Growth Factor 1egf

### Channels, Pumps and Receptors: getting back and forth across the membrane

18. Ras Protein 3p2l
19. Beta2-Adrenergic Receptor/Gs Protein 3b6c
20. Acetylcholine Receptor 2bcg
21. Epidermal Growth Factor Receptor 1em, 2lva, 2p5l
22. Rhodopsin 1f88
23. Polyporphyrin 3p5l
24. Potassium Channel 3kt4
25. Calcium Pump 1ca4
26. Cyclooxygenase 1p9h

### Photosynthesis: harvesting energy from the sun

27. Photosystem II 1psf
28. Light harvesting Complex 1mvt
29. Photosynthetic Reaction Center 1prt

### Enzymes: cutting and joining the molecules of life

39. Fatty Acid Synthase 2znh, 2znc
40. Rubisco: Ribulose Biphosphate Carboxylase/Oxygenase 1rex
41. Green Fluorescent Protein 1gfl
42. Lactase 2lca
43. Glutamine Synthetase 2gls
44. Alcohol Dehydrogenase 2ahx
45. Dihydrofolate Reductase 1dfr
46. Nitrogenase 1n2c
47. Leucine Aminopeptidase 1lap
48. Beta Lactamase 4blm
49. Catalase 1agw
50. Thymidylate Synthase 2tnc
51. Tryptophan Synthase 1twg
52. Aspartate Carboxyltransferase 4at1
53. Hexokinase 1dkg
54. Phosphoglucose isomerase 1thx
55. Phosphofruktokinase 4pfs
56. Aldolase 4ald
57. Triosephosphate isomerase 2tpil
58. Glyceraldehyde-3-phosphate Dehydrogenase 3pfd
59. Phosphoglycerate Kinase 3pks
60. Phosphoglycerate Mutase 3pym
61. Enolase 5enl
62. Pyruvate Kinase 1a3w

### Energy Production: powering the processes of the cell

30. Cytochrome c Oxidase (Complex IV) 1coo
31. Cytochrome c 3cyl
32. Cytochrome bc1 (Complex III) 1bcy
33. Succinate Dehydrogenase (Complex II) 1sdc
34. NADH:Quinone Oxidoreductase (Complex I) 5nbs, 3tko
35. ATP Synthase 1at9, 1kt7, 1l2p, 2a7u
36. Myoglobin 1mbd
37. Hemoglobin 4hbh

### Storage: containing nutrients for future consumption

38. Ferritin 1ftr

### Infrastructure: supporting and moving cells

63. Actin 1mbg
64. Myosin 1mbg
65. Microtubule 1tub
66. Collagen 1h4w (far left, reverse)

### Protein Synthesis: building new molecular machines

67. Transfer RNA 4tra
68. Valyl-tRNA Synthetase 1gat
69. Thionyl-tRNA Synthetase 1gt6
70. Gutamyl-tRNA Synthetase 1eug
71. Isoleucyl-tRNA Synthetase 1lly
72. Phenylalanyl-tRNA Synthetase 1eiy
73. Acamyl-tRNA Synthetase 1aay
74. Ribosome 1l5e, 1l2j
75. Elongation Factor Tu/tRNA 1ttb
76. Elongation Factor G 1dar
77. Elongation Factor Ts and Tu 1eku
78. Peptidyl 1ha
79. Chaperonin GroEL/ES 1aon
80. Proline cis/trans isomerase 2cpl
81. Heat Shock Protein Hsp90 2c9p
82. Proteasome 4b4t
83. Ubiquitin 1ubq

### DNA: storing and reading genetic information

84. DNA 1bna
85. Restriction Endonuclease EcoRI 1eri
86. DNA Topoisomerase 1top
87. Topoisomerase 1top
88. RNA Polymerase 2o3b
89. 3c Receptor 1b4k
90. Catabolite Gene Activator Protein 1cgp
91. DNA-binding Protein' Transcription Factor 1b1at
92. DNA Helicase 4hw
93. DNA Polymerase 1haa
94. Nucleosome 1aoo
95. HU Protein 1p94
96. Single-stranded DNA-binding Protein 3af5a

Scale:  
1nm 5nm 10nm  
1nm (nanometer) = 10<sup>-9</sup> millimeters

Extracellular Proteins
Membrane Proteins
Intracellular Proteins: Cytosol
Intracellular Proteins: Cytosol
Intracellular Proteins: Nucleus



# A Protein Data Bank



- www.rcsb.org
- A legfontosabb weboldalak egyike a szerkezeti biológusok számára
- Ún. elsődleges adatbázis – adatokat a kutatók teszik be
- 2019. február: közel 150 000 szerkezet (nukleinsavak, fehérjék és komplexeik)
- Sok oktatási jellegű információ (pl. hónap molekulája)

The screenshot shows the top navigation bar of the RCSB PDB website. It includes a dark blue header with white text for navigation: "RCSB PDB", "Deposit", "Search", "Visualize", "Analyze", "Download", "Learn", and "More". On the right side of the header is a yellow "MyPDB" button. Below the navigation bar is a white search bar with the placeholder text "Search by PDB ID, author, macromolecule, sequence, or ligands" and a black "Go" button. To the left of the search bar is the RCSB PDB logo and the text "148969 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education". Below the search bar are links for "Advanced Search" and "Browse by Annotations". At the bottom of the header are several partner logos: "PDB-101", "Worldwide Protein Data Bank", "EMDataResource", "Nucleic Acid Database", and "Worldwide Protein Data Bank Foundation". On the far right are social media icons for Facebook, Twitter, YouTube, and LinkedIn.

The screenshot shows the main content area of the RCSB PDB website. On the left is a dark blue sidebar with white text and icons for navigation: "Welcome", "Deposit", "Search", "Visualize", "Analyze", "Download", and "Learn". The main content area has a white background. At the top is a section titled "A Structural View of Biology" with a blue header. Below the title is a paragraph: "This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease." Below this is another paragraph: "As a member of the wwPDB, the RCSB PDB curates and annotates PDB data. The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond." Below the text is a section titled "Job Opportunity with RCSB PDB at Rutgers" with a blue header. Below the title is a banner image showing a desk with a computer monitor, a red chair, and a protein structure. To the left of the banner is a blue box with white text: "JOIN OUR TEAM". On the right side of the main content area is a section titled "February Molecule of the Month" with a blue header. Below the title is a large 3D molecular model of a protein structure, colored in shades of blue, green, and red. Below the model is a blue box with white text: "Initiation Factor eIF4E".



# A Protein Data Bank



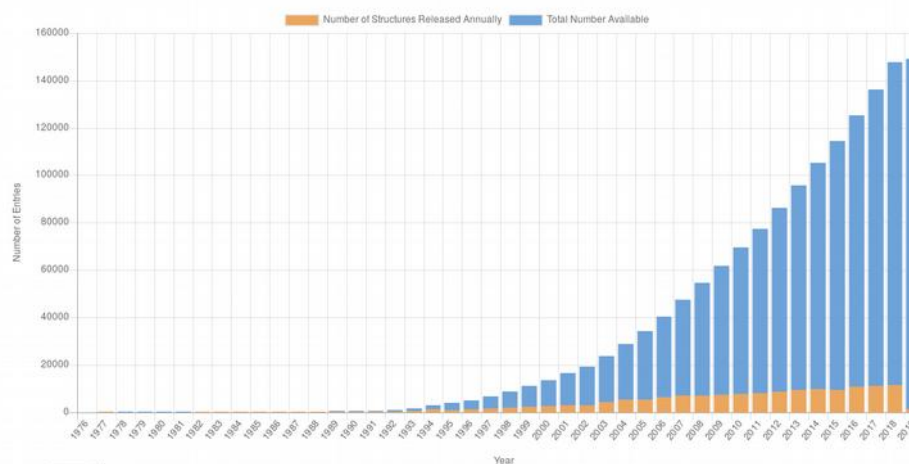
A Protein Data Bank (PDB) az RCSB (Research Collaboratory for Structural Bioinformatics) által fenntartott adatbázis, mely fehérjék és nukleinsavak térszerkezetét tartalmazza. Az egyetlen valóban átfogó biológiai térszerkezeti adatbázis, elvileg az összes ismert fehérjeszerkezetet tartalmazza (gyakorlatilag csak azokat, amelyeket a szerkezetet meghatározó kutatók ténylegesen hozzáférhetővé tettek). Ma már a legtöbb tudományos folyóirat megköveteli, hogy a benne ismertetett fehérjeszerkezeteket a szerzők eljuttassák a PDB-be a közléssel egyidejűleg.

Elsődleges adatbázis, azaz a "nyers" adatokat tartalmazza osztályozás stb. nélkül. 1971-ben alapította Walter Hamilton a Brookhaveni Nemzeti Laboratóriumban.

## A PDB néhány fontos jellemzője röviden

- Webes elérhetősége: [www.rcsb.org](http://www.rcsb.org), [www.wwpdb.org](http://www.wwpdb.org)
- A szerkezeteket atomi koordináták formájában tartalmazza
- A térszerkezeteket 4 karakteres kódokkal azonosítja (pl. 1wdc, 1mup stb.), az első karakter kötelezően szám, a következő 3 szám vagy betű
- A benne található szerkezetek számának növekedése a többi elsődleges biológiai adatbáziséhoz hasonlóan exponenciális
- Hetenként frissítik

PDB Statistics: Overall Growth of Released Structures Per Year



# A PDB formátum

Fejléc (annotáció): a molekula adatai, szerzők, kísérletes technika, stb.

NINCS: taxonómia (↔ pl. GenBank)

```
HEADER      MUSCLE PROTEIN                               19-JAN-96    1WDC
TITLE       SCALLOP MYOSIN REGULATORY DOMAIN
COMPND      MOL_ID: 1;
COMPND      2 MOLECULE: SCALLOP MYOSIN;
COMPND      3 CHAIN: A, B, C;
COMPND      4 FRAGMENT: PROTEOLYTIC FRAGMENT, REGULATORY DOMAIN;
COMPND      5 OTHER_DETAILS: PH 7.0
SOURCE      MOL_ID: 1;
SOURCE      2 ORGANISM_SCIENTIFIC: AEQUIPECTEN IRRADIANS;
SOURCE      3 ORGANISM_COMMON: BAY SCALLOP;
SOURCE      4 TISSUE: SKELETAL MUSCLE
KEYWDS      MYOSIN, CALCIUM BINDING PROTEIN, MUSCLE PROTEIN
EXPDTA      X-RAY DIFFRACTION
AUTHOR      A.HOUDUSSE,C.COHEN
REVDAT      1   11-JUL-96 1WDC      0
JRNL        AUTH    A.HOUDUSSE,C.COHEN
JRNL        TITL    STRUCTURE OF THE REGULATORY DOMAIN OF SCALLOP
JRNL        TITL 2  MYOSIN AT 2 A RESOLUTION: IMPLICATIONS FOR
JRNL        TITL 3  REGULATION
JRNL        REF     TO BE PUBLISHED
JRNL        REFN
```

Egyéb információk: szekvencia, másodlagos szerkezeti elemek, diszulfidhidak...



# A PDB formátum

Törzs: atomi koordináták

ATOM	1	N	ARG	A	774	-31.629	7.797	92.108	1.00	71.22	N
ATOM	2	CA	ARG	A	774	-31.385	8.882	91.101	1.00	71.34	C
ATOM	3	C	ARG	A	774	-29.888	9.183	90.975	1.00	70.56	C
ATOM	4	O	ARG	A	774	-29.474	10.339	91.042	1.00	71.34	O
ATOM	5	CB	ARG	A	774	-32.139	10.161	91.504	1.00	71.23	C
ATOM	6	CG	ARG	A	774	-33.305	10.546	90.582	1.00	74.01	C
ATOM	7	CD	ARG	A	774	-34.689	10.154	91.158	1.00	79.24	C
ATOM	8	NE	ARG	A	774	-35.050	10.873	92.400	1.00	83.50	N
ATOM	9	CZ	ARG	A	774	-36.024	10.514	93.259	1.00	84.49	C
ATOM	10	NH1	ARG	A	774	-36.947	9.596	92.924	1.00	80.65	N
ATOM	11	NH2	ARG	A	774	-36.117	11.134	94.447	1.00	84.25	N
ATOM	12	N	ASP	A	775	-29.076	8.139	90.843	1.00	70.24	N
ATOM	13	CA	ASP	A	775	-27.627	8.302	90.918	1.00	71.14	C
ATOM	14	C	ASP	A	775	-27.042	8.263	89.530	1.00	68.46	C
ATOM	15	O	ASP	A	775	-26.321	9.177	89.118	1.00	67.65	O

atom  
sorszáma, neve

aminosav típusa, láncazonosítója, sorszáma

3D koordináták

betöltöttség

B-faktor

atom  
típusa

# Mi van a PDB-ben?

## **Amit a kutatók beleraknak!**

- Az adatok minőségéért, megbízhatóságáért, teljességéért nem a PDB, hanem az adatokat ott elhelyező kutatók a felelősek!
- Egy térszerkezet meghatározása túl nagy erőfeszítés ahhoz, hogy bárkit elutasítsanak ⇒ széles skálán mozog a szerkezetek minősége
- A PDB ma már előírja, hogy a koordináták mellett a kutatók néhány járulékos kísérleti adatot is hozzáférhetővé tegyenek, ilyenek a krisztallográfiai szerkezeti faktorok vagy az NMR-szerkezeteknél a szerkezetszámoláshoz használt kényszerfeltételek (általában atomi távolságok)
- Sokszor nem frísítik a kapcsolódó közlemény adatait, ha az megjelent
- PDB nagytakarítás (formátumegységesítési kísérlet): 2007 .

## **Koordináták elhelyezése a PDB adatbázisban**

- Ma a koordináták elhelyezése weben keresztül, automatikus feldolgozás segítségével történik.
- A koordinátákat a PDB formátumnak megfelelően kell megadni, a feltöltésnél szintaktikai (megfelel-e a feltöltött fájl a PDB formátumnak) és (szerkezet)minőségi ellenőrzés is történik, utóbbi eredménye bekerül a végleges állományba.
- A PDB kód nem választható, annak hozzárendelése automatikus a feltöltési eljárás megkezdésekor
- A feltöltési eljárás befejezése után a beküldő szerző e-mailen megkapja az elkészített állományt és lehetősége van azon módosítani a PDB személyzet közreműködésével.
- A végleges változat elfogadása után a koordináták nem feltétlenül válnak azonnal hozzáférhetővé mindenki számára, a szerzők ugyanis megszabhatják, hogy mikor váljanak azzá (azonnal, a kapcsolódó közlemény megjelenésekor vagy adott időpontban). A visszatartási idő azonban nem lehet több, mint egy év.



# Precizitás és szabatoság

- Szabatoság (accuracy): valóságnak való megfelelés
- Precizitás (precision): kísérleti hiba



precíz & szabatos



szabatos, de nem  
precíz  
(átlag ~ OK)



precíz de nem  
szabatos

- Szabatoság: CSAK a független kísérleti ellenőrzés dönti el egyértelműen!

## Vigyázat, (már itt is) csálnak!

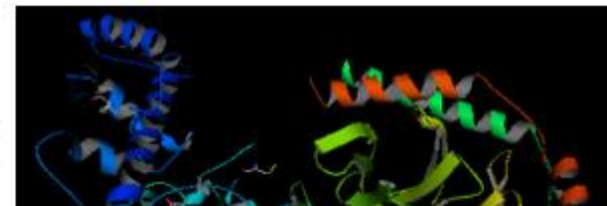
# Crystallographer faked data

A protein researcher at the University of Alabama at Birmingham (UAB) has been found guilty of falsifying data that he used to construct 12 fraudulent protein structures that made it into the scientific literature and an international archive of protein structures. A G-protein image based on crystal structure data image: S. Jahnichen  
After investigating the misconduct -- with the help of a committee of independent protein scientists -- the university has [linkurl:asked;http://main.uab.edu/Sites/rep](http://main.uab.edu/Sites/rep)

By Bob Grant | December 18, 2009



A protein researcher at the University of Alabama at Birmingham (UAB) has been found guilty of falsifying data that he used to construct 12 fraudulent protein structures that made it into the scientific literature and an international archive of protein structures. After investigating the misconduct -- with the help of a committee of independent protein scientists -- the university has [linkurl:asked;http://main.uab.edu/Sites/reporter/articles/71570/](http://main.uab.edu/Sites/reporter/articles/71570/) that the structures be removed from the database and that ten research papers, authored by former



### **Csak alapos minőségellenőrzés vezetett rá a szerkezetek hamis mivoltára!**

- Szerkezetek elemzésével professzionális szinten foglalkozó csoport „leplezte le” a csalást



# A minőségellenőrzés elvei

**Alapelv: adott paraméterek összevetése ismert/elvárt értékekkel**

**Honnan vannak az ismert/elvárt értékek?**

- Ismert egyéb szerkezetek:

- Kismolekulák (ezek esetében kötéshosszak stb. sokszor közvetlenül és több független módszerrel is meghatározhatók) → feltételezzük, hogy hasonló kémiai jellegű részletek hasonló geometriával rendelkeznek a fehérjékben is (pl. amidkötés)
- Ismert fehérjék → körkörös érvelésnek tűnik, de az ismert szerkezetek számának növekedésével (és azok egyre részletesebb leírásával) már rendelkezésre áll elegendő mennyiségű és minőségű adat függetlennek tekinthető vizsgálathoz)

- Elméleti számítások (amiket szintén kismolekulák segítségével paramétereznek és/vagy ellenőriznek)

**Egy fontos (de sokszor nem kellően hangsúlyozott) szempont:**

- Mind a röntgenkristallográfiában, mind az NMR-spektroszkópiában a szerkezetmeghatározás adatszegénynek számít, azaz a meghatározandó koordináták számához képest kevés kísérleti adatpont van → fel kell használni ismert geometriai jellemzőket a modellépítés során

- Emiatt a makromolekuláris szerkezetek a mért/ismertnek vett paramétereket vegyesen alkalmazó módszerekkel születnek, a kétféle adattípus relatív súlya pedig esetenként változó lehet:

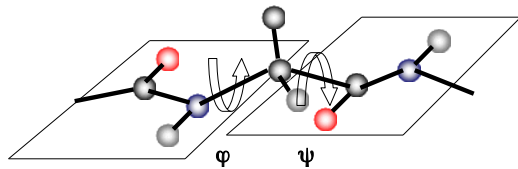
- kiváló geometria: lehet, hogy a kísérleti adatokat nem teljesíti maradéktalanul a szerkezet
- maradéktalan megfelelés a kísérleti adatoknak: lehet, hogy a geometria rovására?

# A minőségellenőrzés elvei

## A Ramachandran-felület

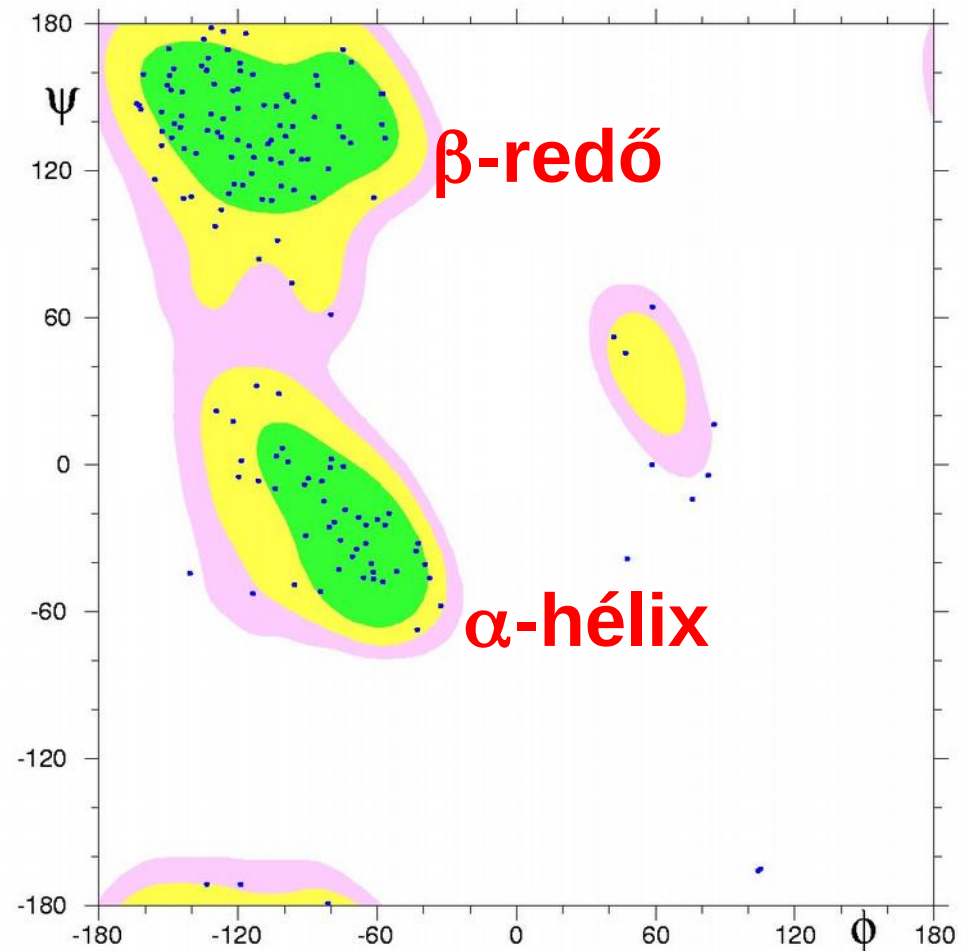
- Alapja, hogy a fehérjékben az aminosavak molekulagerincének  $\varphi$  és  $\psi$  torziós szögei nem vehetnek fel akármilyen értékeket, vannak preferált konformációk.
- E két torziós szöget X és Y-koordinátaként használva a kapott koordinátarendszerben ábrázolhatjuk az egyes aminosavaknak megfelelő szögpárokat.
- Általában a glicin és prolin aminosavak nem szerepelnek a térképen, ezek konformációs preferenciái ugyanis egyediek. Mint az várható, az  $\alpha$ -hélixeknek és a  $\beta$ -szálaknak megfelelő régiókba esik a legtöbb aminosav, de a hurkokban és kanyarokban lévők többsége is a megengedett tartományokon belül van.

Az 1MUP (major urinary protein) fehérje Ramachandran-térképe



## Egyéb paraméterek

- Kötéshosszak, kötésszögek megfelelő volta
- Ne legyenek atom-atom ütközések
- Eltemetett amidcsoportok többsége H-kötésben legyen

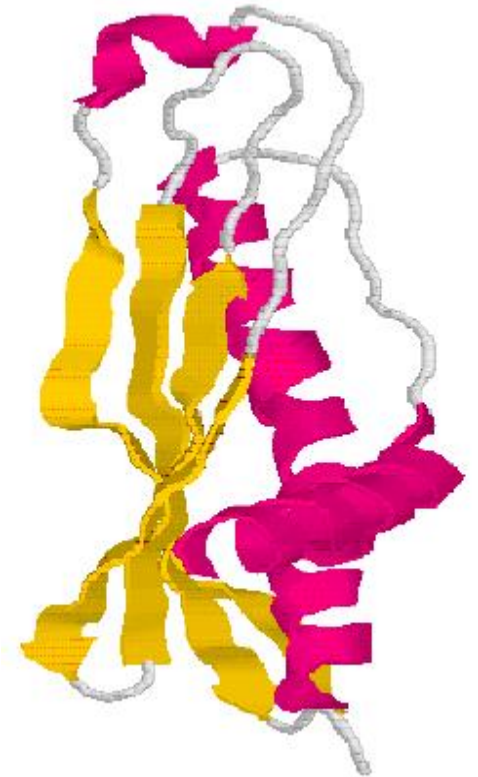
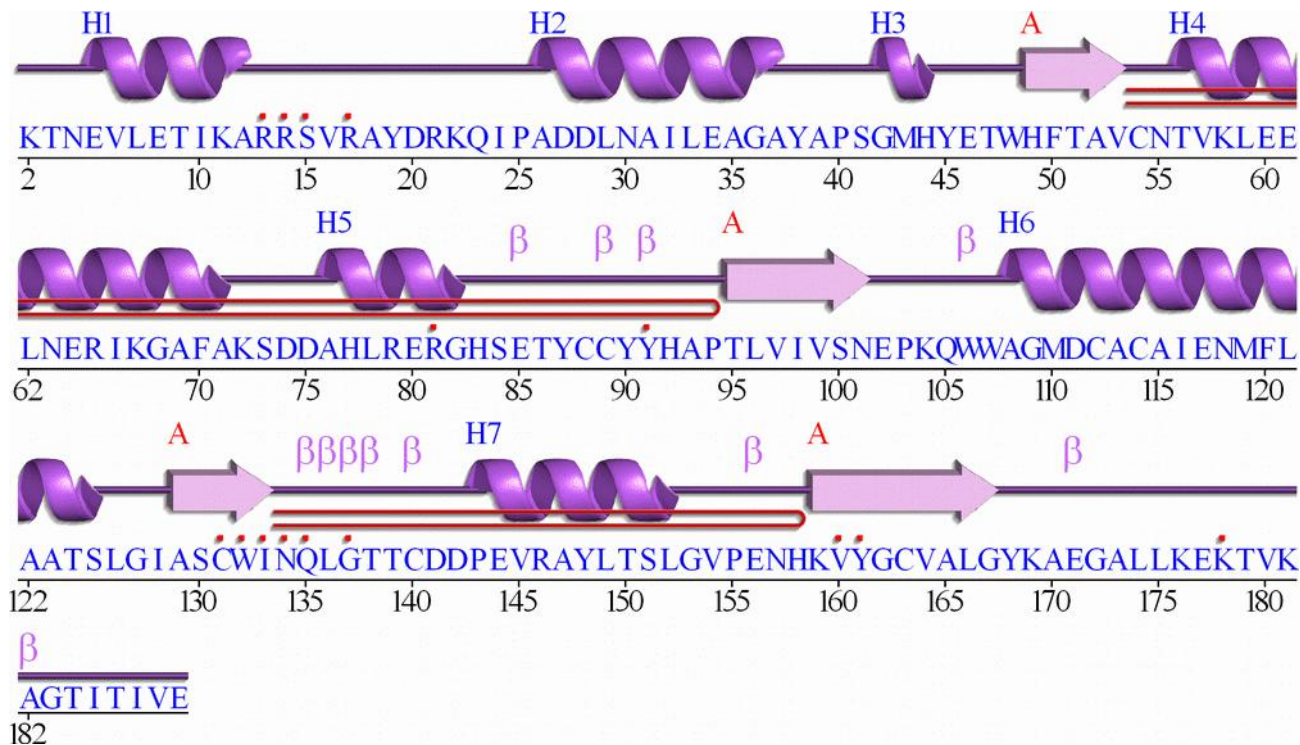




# Másodlagos szerkezeti elemek hozzárendelése

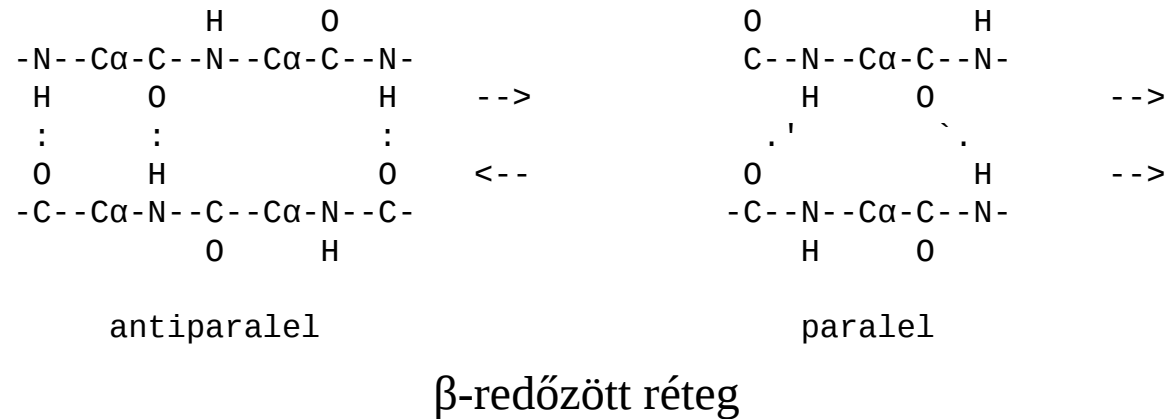
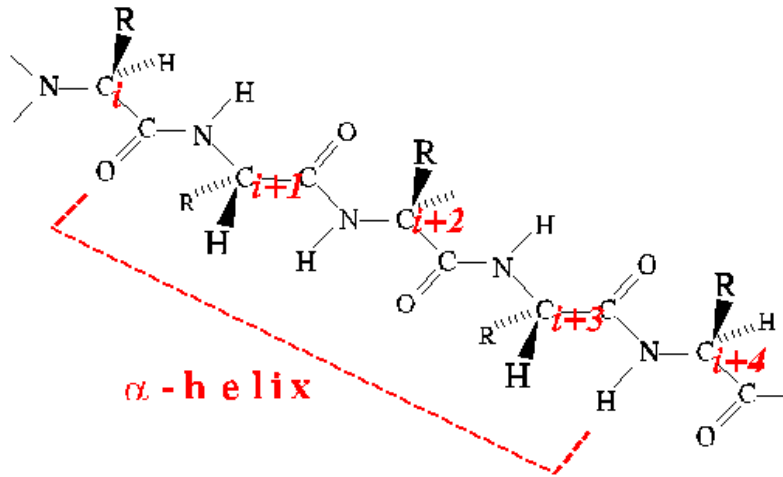
## Definíciók és megfontolások

- Másodlagos szerkezet a polipeptidlánc helyi (lokális) konformációja, minden szakasznak van!
- Sokszor csak az  $\alpha$ -hélixekre és  $\beta$ -lemezekre használják!
- A filozófiai kérdés az, hogy ezek a szerkezeti elemek valóban létező kategóriák-e (hol kezdődnek, hol végződnek?)
- Fehérjék flexibilitása: a szerkezeti elemek időben állandóak-e?
- A másodlagos szerkezeti elemek "valóságába" vetett hitet a homológ szerkezetek összevetéséből nyert kép erősíti
- Miért fontos? Megjelenítés, jellemzés, térszerkezetjósítás (ennek egyik kulcslépése a másodlagos szerkezeti elemek predikciója, ami egyre pontosabb)



# Másodlagos szerkezeti elemek hozzárendelése

- A másodlagos szerkezeti elemek eredeti definíciója (Linus Pauling!) a hidrogénkötés-mintázatokon alapul
- *De facto* standard: DSSP (Dictionary of Secondary Structure of Proteins): csak hidrogénkötés-mintázatokat néz:
  - Az  $\alpha$ -hélix kritériuma, hogy két  $i \rightarrow i+4$  hidrogénkötéssel kezdődjön (azaz két egymás utáni aminosav CO csoportja alkosson a négyvel utána következő NH csoportjával hidrogénkötést), és két  $i-4 \leftarrow i$  hidrogénkötéssel végződjön

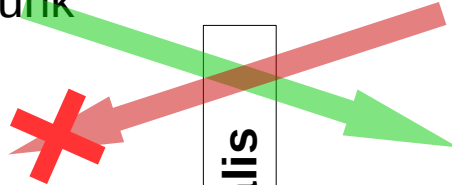


- A  $\beta$ -szálba sorolás feltétele, hogy vagy az adott aminosav létesítsen két megfelelő H-kötést, vagy két H-kötéssel legyen körülvéve. A szálak helyzete lehet antiparalel vagy paralel.
- Egyéb programok a molekulagerinc torziós szögeit is figyelembe vehetik (ezáltal pl. hosszabb hélixeket/szálakat tudnak azonosítani)

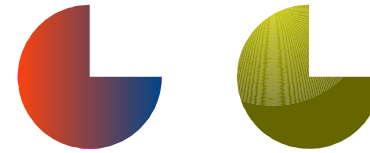
# Lokális és globális hasonlóság a bioinformatikában

Az evolúciós rokonság jelének tekintjük (valószínűtlen, hogy egymástól függetlenül ennyire hasonló dolgok alakuljanak ki - globuláris fehérjékre igaz)

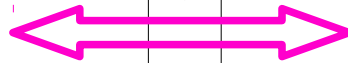
Hasonló térszerkezetet várunk



A szekvenciák közötti hasonlóság nem feltétlenül könnyen detektálható (divergencia)



Általában a teljes szekvenciát tekintjük



Szerkezeteknél általában a domének szintjén értelmezzük

szekvencia

**hasonlóság**

3D szerkezet

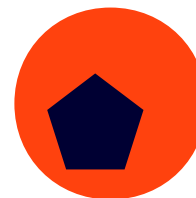
Általában doméneket/motívumokat vizsgálunk



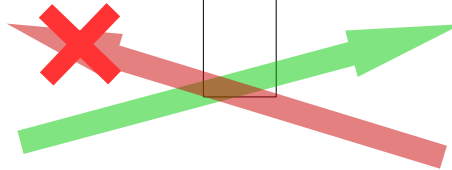
Doméneknél kisebb egységekre értjük általában



Lokális



Jelezhet hasonló lokális szerkezetet



A szekvenciában nem feltétlenül folytonos szegmens (pl. aktív centrum)

A lokális hasonlóság sokszor hasonló funkcióra utal (pl. aktív centrum, partnerkötőhely)  
Nem minden esetben feltételez evolúciós rokonságot (konvergencia)



# Definíciók

- **Szerkezetek fedésbe hozása:**
  - Két, *egymásnak megfeleltethető/megfeleltetett* ponthalmaz között a legjobb átfedés megtalálása (pl. két konformer C $\alpha$  atomjai)
  - Van egzakt matematikai megoldása (legkisebb négyzetes illesztés)
- **Szerkezetillesztés:**
  - Találjuk meg két ponthalmaz között azt a megfelelést, ami a legjobb illesztést fogja adni
  - NP-nehéz probléma ( $\Rightarrow$  sok különböző, heurisztikus megoldás)
- **Szerkezetek összehasonlítása:**
  - Két szerkezet között valamilyen hasonlósági mérőszám megadása
  - Az eljárásnak rész lehet szerkezetillesztés, de ez nem feltétlenül szükséges
- Mindhárom lehet **lokális vagy globális** (utóbbi általában domént szintet jelent itt!)

**1. megjegyzés:** (a szekvenciaillesztéssel ellentétben) ezen műveletek egyikében sem vesszük figyelembe az aminosavak típusát(!). Egyes esetekben az illesztett atomok kémiai jellege természetesen fontos lehet (pl. lokális illesztés kötőhelyek azonosításához)

**2. megjegyzés:** az olyan fedésbe hozás, ami szekvenciaillesztés segítségével meghatározott pozíciómegfeleltetésen alapul, nem tekinthető szerkezetillesztésnek ( $\leftrightarrow$  sok szerkezetmanipuláló program így hívja, figyeljünk oda!). Nagyon hasonló szekvenciák és szerkezetek esetében természetesen ez megfelelő lehet.

**3. megjegyzés:** mint MINDEN bioinformatikai probléma (ideértve a szekvenciaillesztést is!) esetében, a legjobb matematikai megoldás nem feltétlenül értelmes biológiailag! (Pl. két evolúciósan rokon enzim illesztésekor az aktív centrumoknak illik ugyanott lenniük, különben értelmetlen lehet az egész).

**4. megjegyzés:** a szekvenciák esetéhez hasonlóan itt is van többszörös illesztés, ill. hasonlóságkeresés adatbázisban

# Analógiák a szekvenciák köréből

- Szekvenciák „fedésbe hozása”:
  - Triviálisan hasonlító szekvenciák esetében (pl. 1 pozícióban különböznek) egyszerűen egymás alá írjuk őket:

**EVTCEPGFKD**

**EVTCEPGFAD**

- Szekvenciaillesztés:
  - Nem annyira triviális hasonlóságnál megkeressük a legjobb illesztést (ennek van egzakt megoldása!)

**EVTCEPGFK-D**

**D-SCKP-YAID**

A gyakorlatban a szekvenciaillesztés annyira gyors, hogy mindig azt használjuk...

- Szekvenciák összehasonlítása illesztés nélkül:
  - Pl. aminosavösszetétel vagy egyéb jellemzők alapján, ha az illesztés nem praktikus:

**EEVTKRKAIDEEERRKS**

**RRREEESSKKREEDYK**

# Térszerkezetek illesztése és fedésbe hozása



## Fedésbe hozás (superposition)

(pl. ugyanazon fehérje NMR- és kristályszerkezete)  
csak matematikai probléma, az egymásnak megfelelő  
atomok adottak. Egyértelmű (?) mérőszám: RMSD  
(root mean square deviation, atomi pozíciók átlagos  
négyzetes eltéréseinek gyöke)

## Szerkezetillesztés (structure alignment)

az egymásnak megfelelő atomokat keressük  
(pl. kanonikus szerinproteáz-inhibitorok  
enzimkötő régiójának illesztése)  
jószág megítélése??

A szekvenciaillesztéssel ellentétben nincs  
egyértelmű megoldása (NP-nehéz probléma,  
sok különböző megvalósítás)

