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"***.

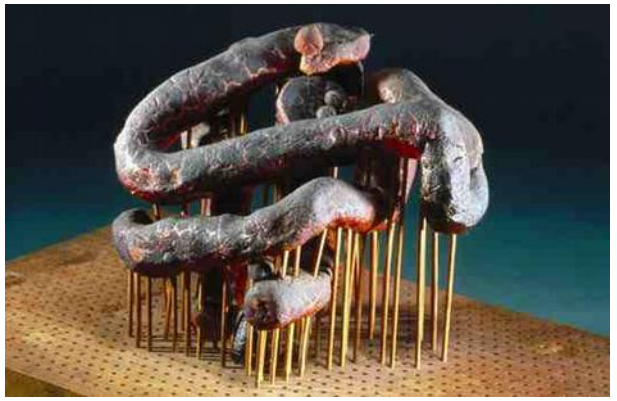
But the achievements of bioinformatics in the field of structural biology are far more immense than the identification of the odd homolog, as has been recognized by the award of the 2013 Nobel prize in Chemistry to Michael Levitt, Arieh Warshel and Martin Karplus for achievements in computational modeling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

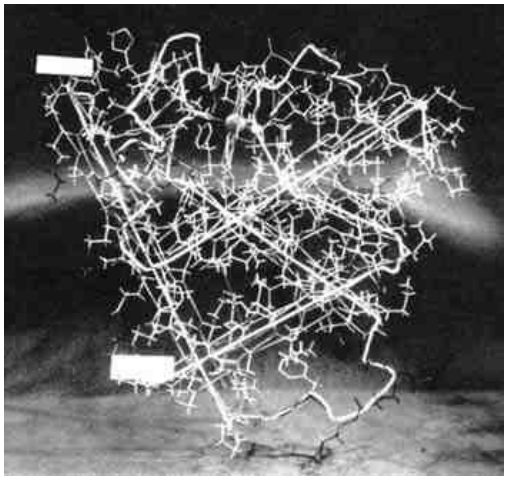
Since the 1970s, Levitt has continuously been a strong pillar, visionary, and supporter of the growing field of computational biology. Like his co-laureates, Karplus and Warshel, he succeeded in the quest to move from the atomic and subatomic phenomena involved in chemical reactions to providing computational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

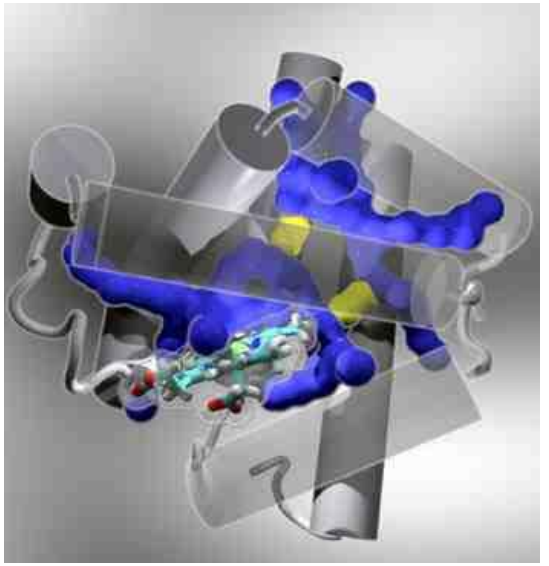
Levitt is an inspiring role model within the computational biology community who advocates for the good that science can bring to society. His enthusiasm and humility persist in the face of this recognition, as he stated in his own words, “The Nobel is recognition and it's fun to get it, and it's good 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/

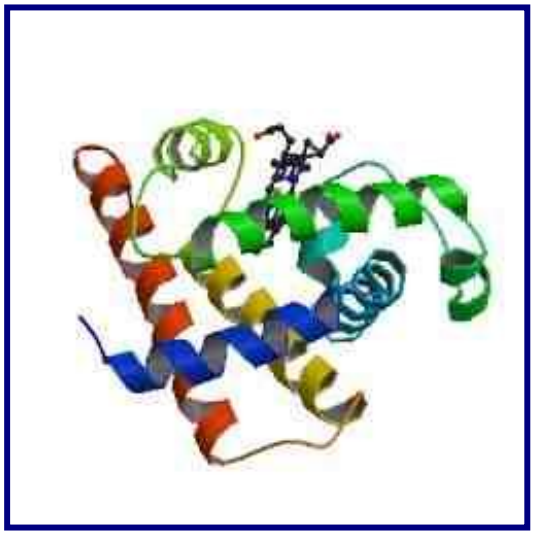
Minden fehérjeszerkezet modell!







a mioglobin
'evolúciója'

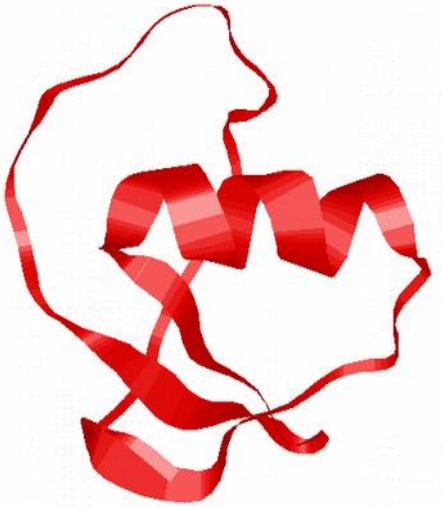


• A szerkezetkutatásban a modell szót általában a kísérleti adatok nélkül, predikciós módszerekkel előállított szerkezetekre használják

A fehérjék szerkezeti szintjei

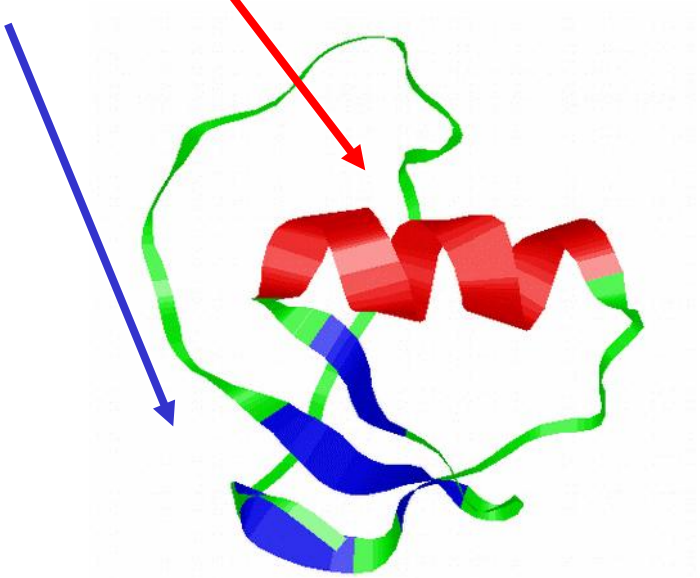
Elsődleges szerkezet (=szekvencia):
VDCSEYPKPACPKDYRPVCGSDNKTYSNKCNFCNAVV
ESNGTLTLNHFGKC
(turkey ovomucid inhibitor III domain)

Harmadlagos szerkezet:
a molekula 3D-szerkezete

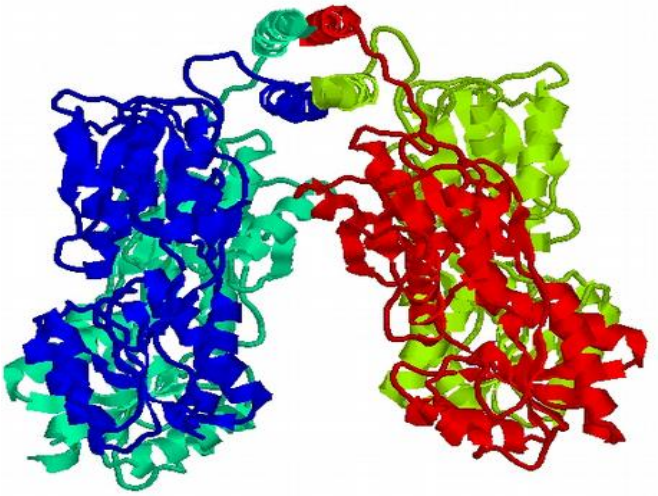


Másodlagos szerkezet:
hurkok vagy β-kanyarok
α-hélix
β-redő

A polipeptidlánc lokális konformációja



Negyedleges szerkezet:
alegységösszetétel

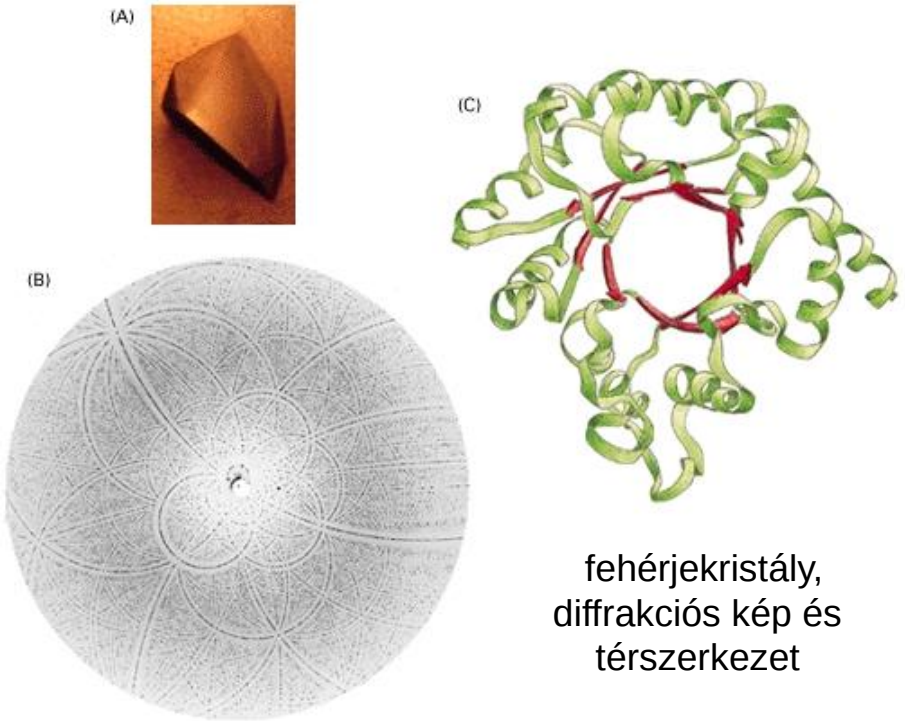


A fehérjék térszerkezetének meghatározása

röntgenkrisztallográfia

- 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

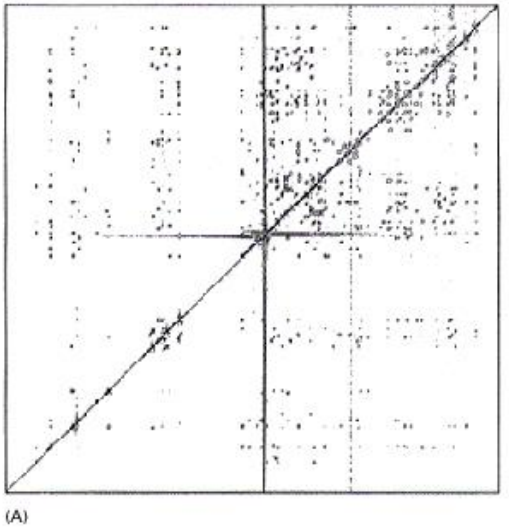
A fehérjék térszerkezetének meghatározása

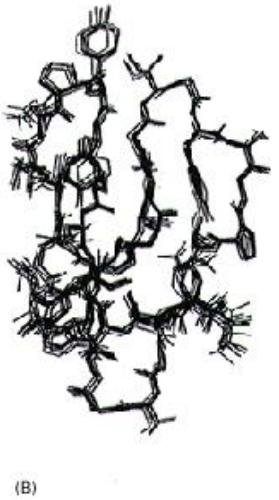
NMR-spektroszkópia

Kémiai Nobel-díj, 2002

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

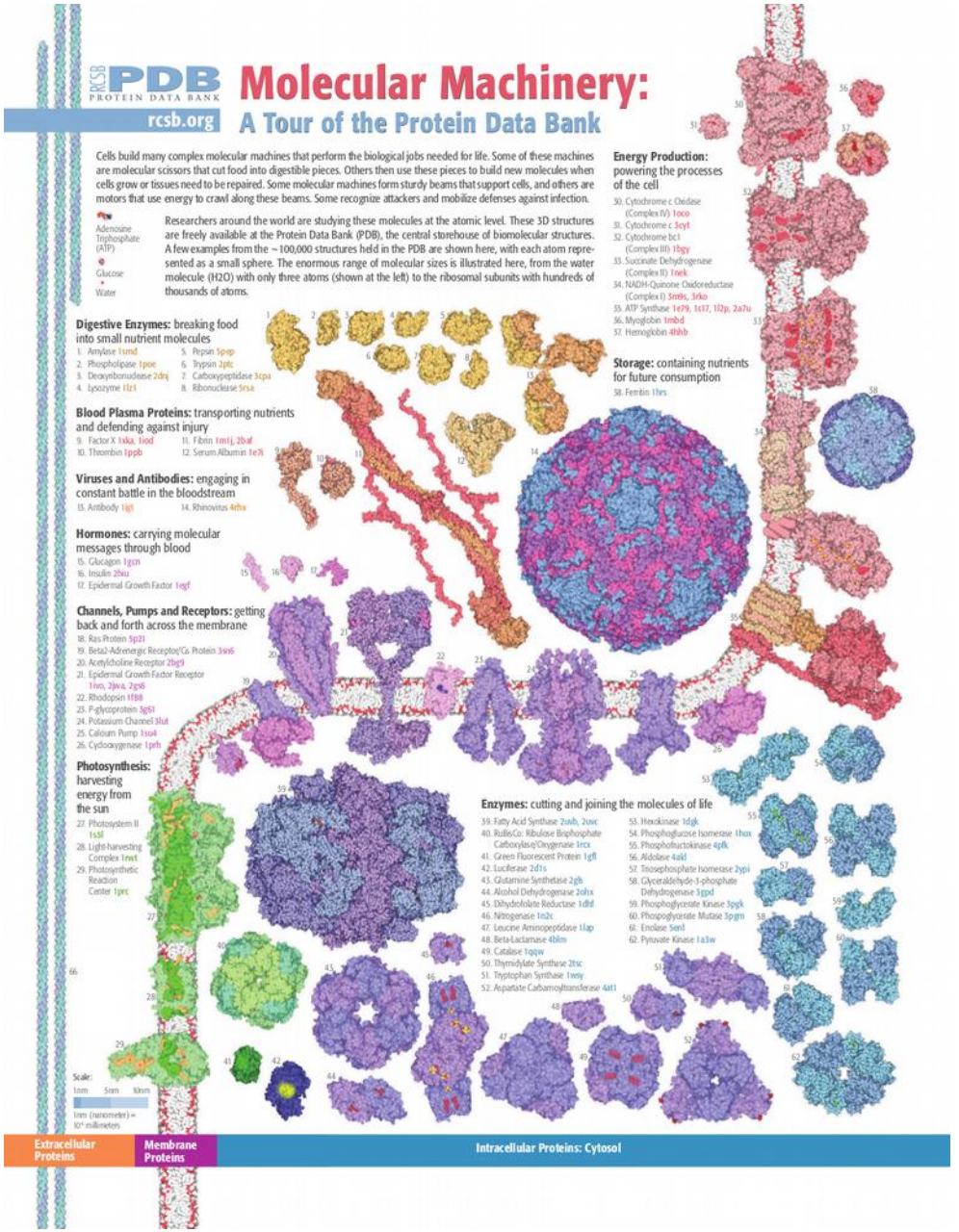
NMR-spektrométer

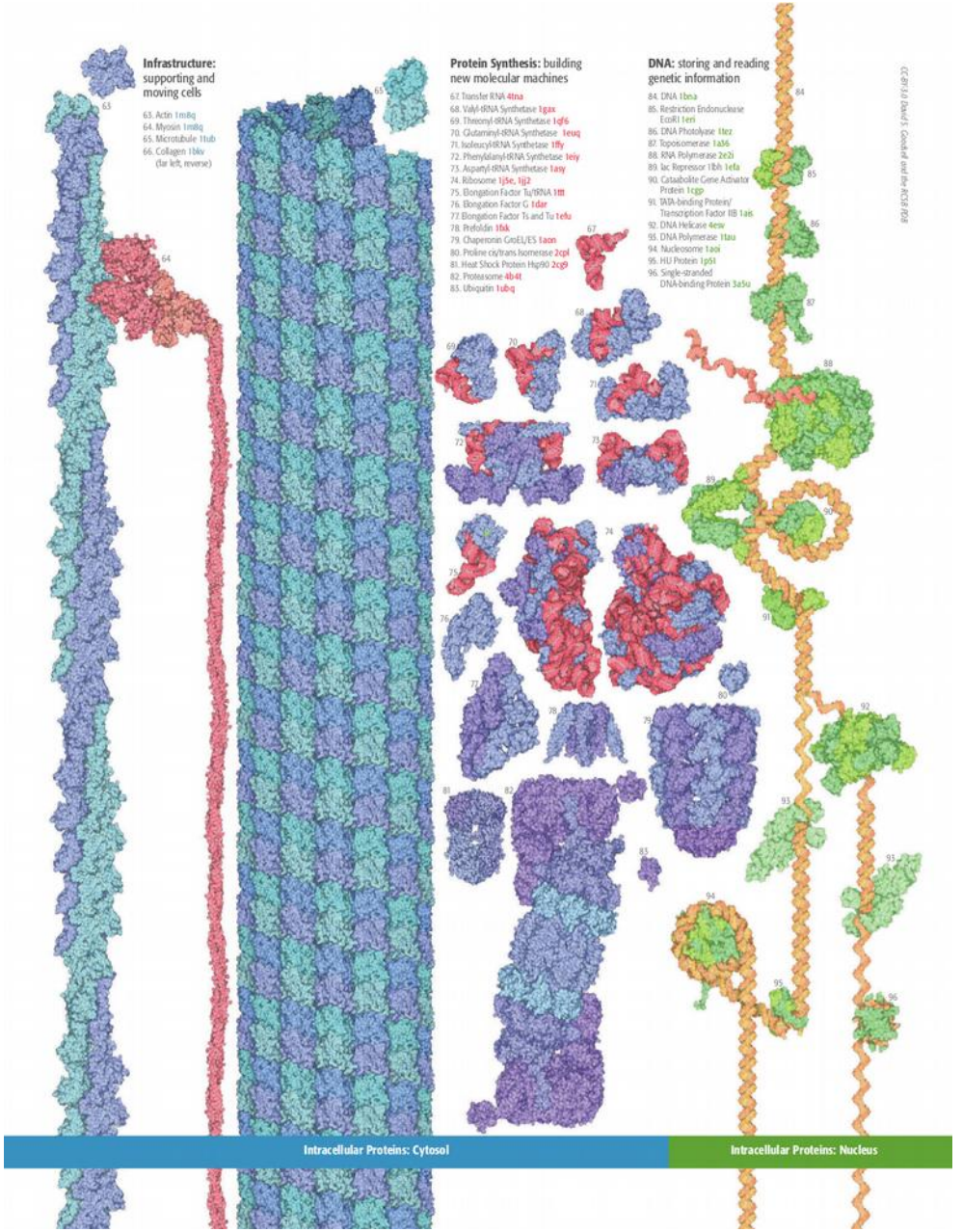




2D NMR-spektrum és térszerkezeti sokaság

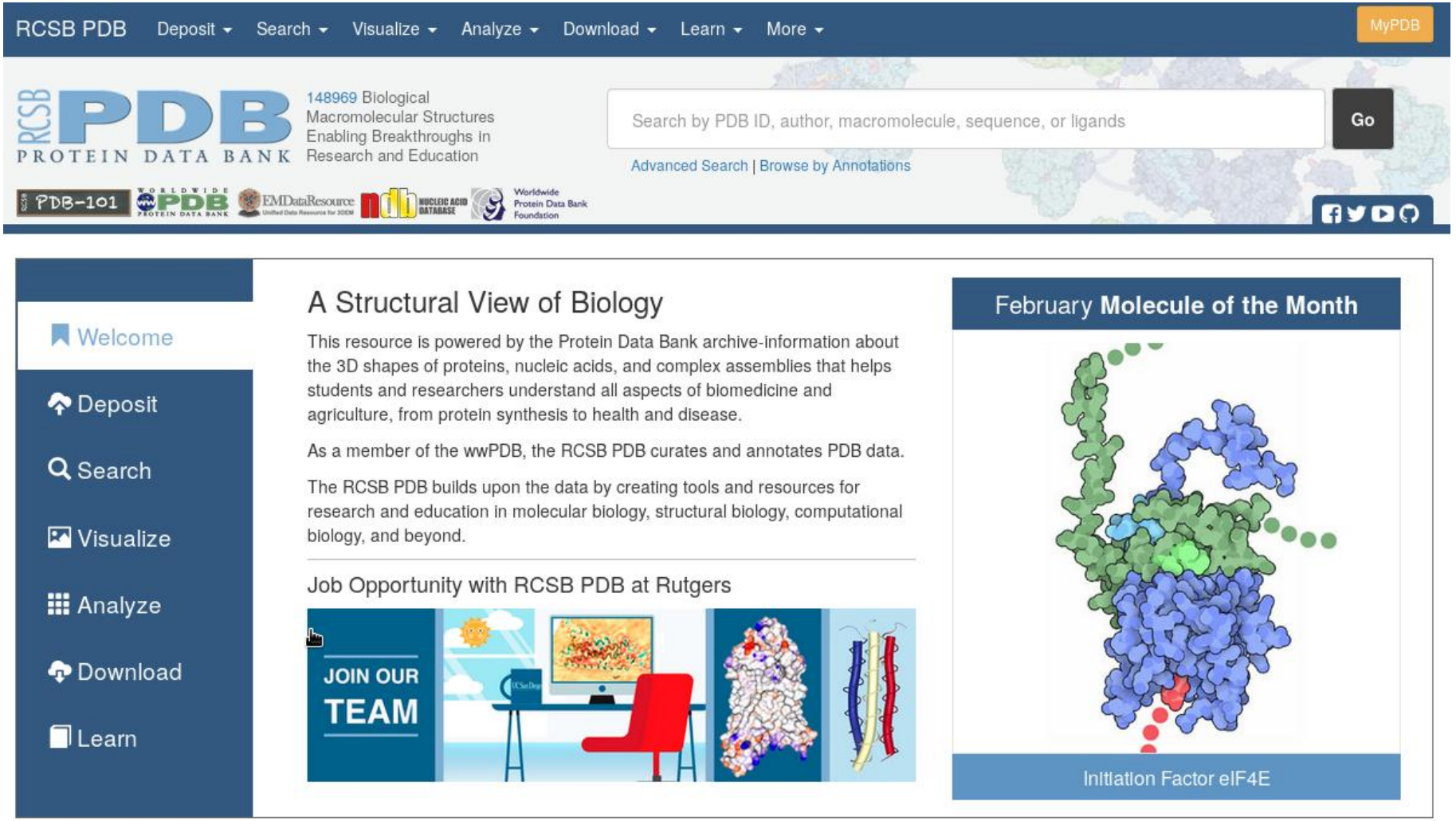
A 'fehérjék állatkertje'





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)

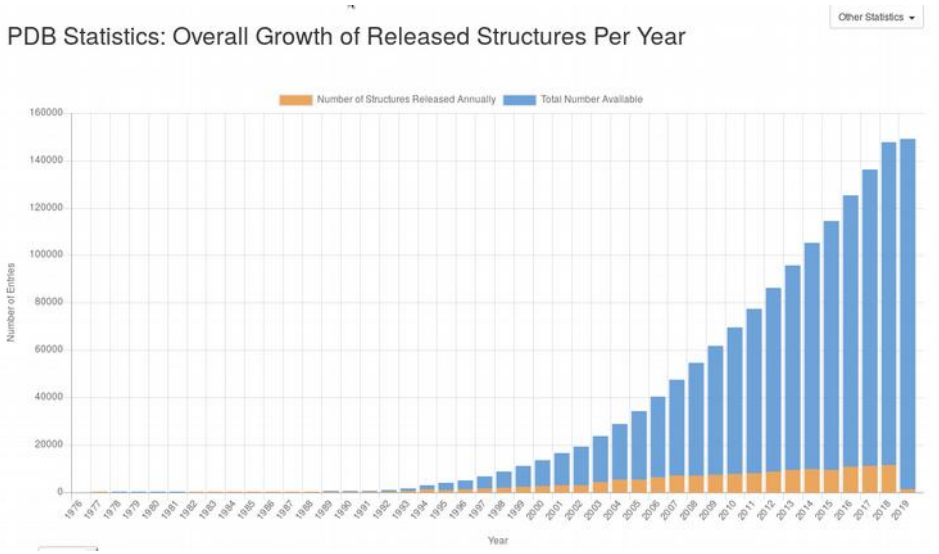


A Protein Data Bank

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, 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



A PDB formátum

Fejléc (annotáció): a molekula adatai, szerzők, kísérletes technika, stb.
NINCS: taxonómia (↔ pl. GenBank)

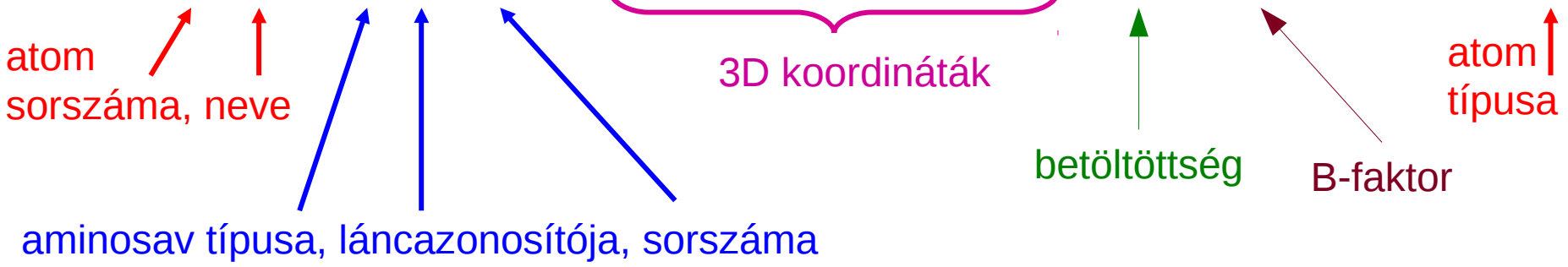
```
HEADER    MUSCLE PROTEIN                          19-JAN-96   1WDC
TITLE     SCALLOP MYOSIN REGULATORY DOMAIN
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: SCALLOP MYOSIN;
COMPND   3 CHAIN: A, B, C;
COMPND   4 FRAGMENT: PROTEOLYTIC FRAGMENT, REGULATORY DOMAIN;
COMPND   5 OTHER_DETAILS: PH 7.0
SOURCE    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
```



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 frisítik 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 szabatosság

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

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

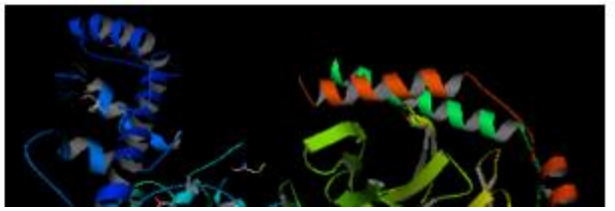
Vigyázat, (már itt is) csalnak!

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 oncrystal structure dataImage: 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

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/ 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 röntgenkrisztallográfiában, 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

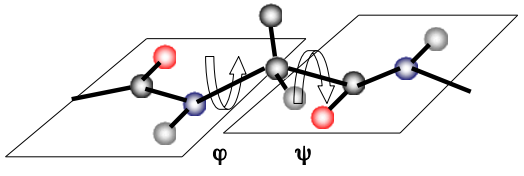
A Ramachandran-felület

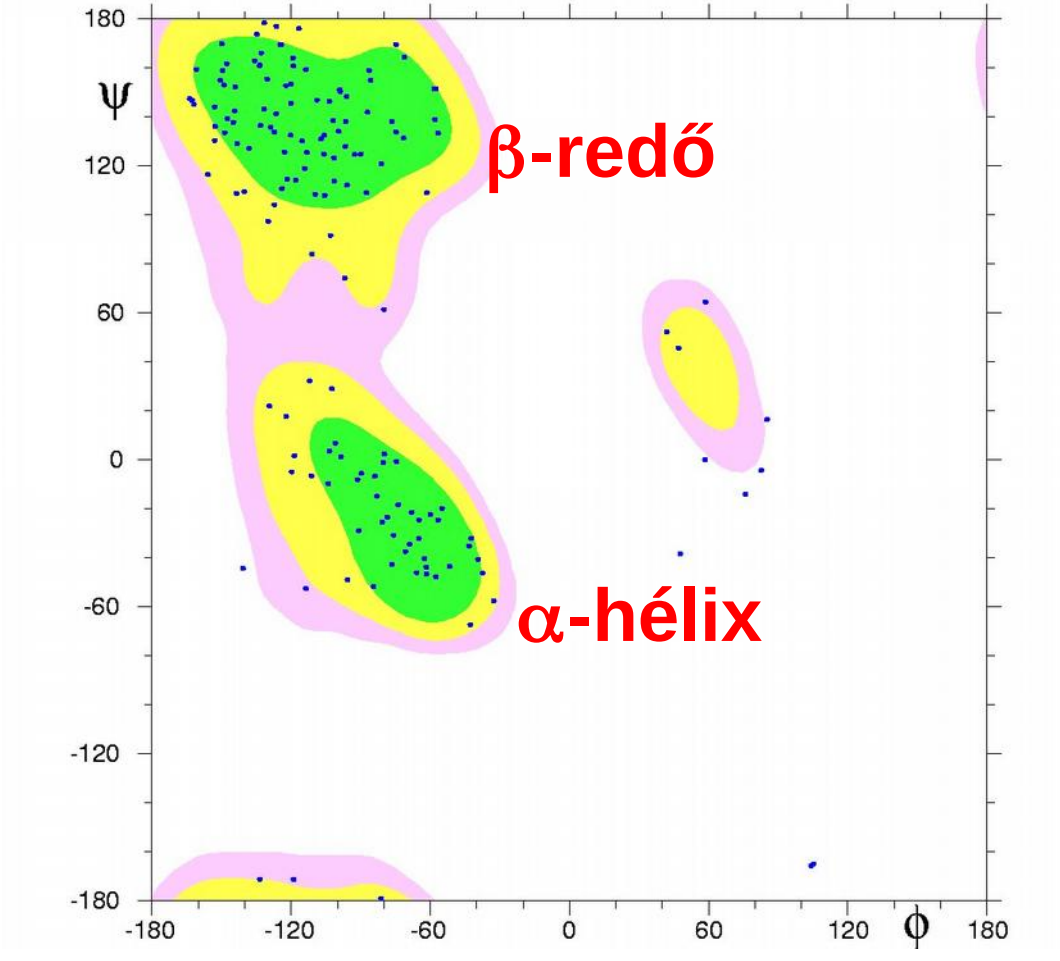
• Alapja, hogy a fehérjékben az aminosavak molekulagerincének φ és ψ torziós szögei nem vehetnek fel akármilyen értékeket, vannak preferált konformációk.

• 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 α-hélixeknek és a β-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

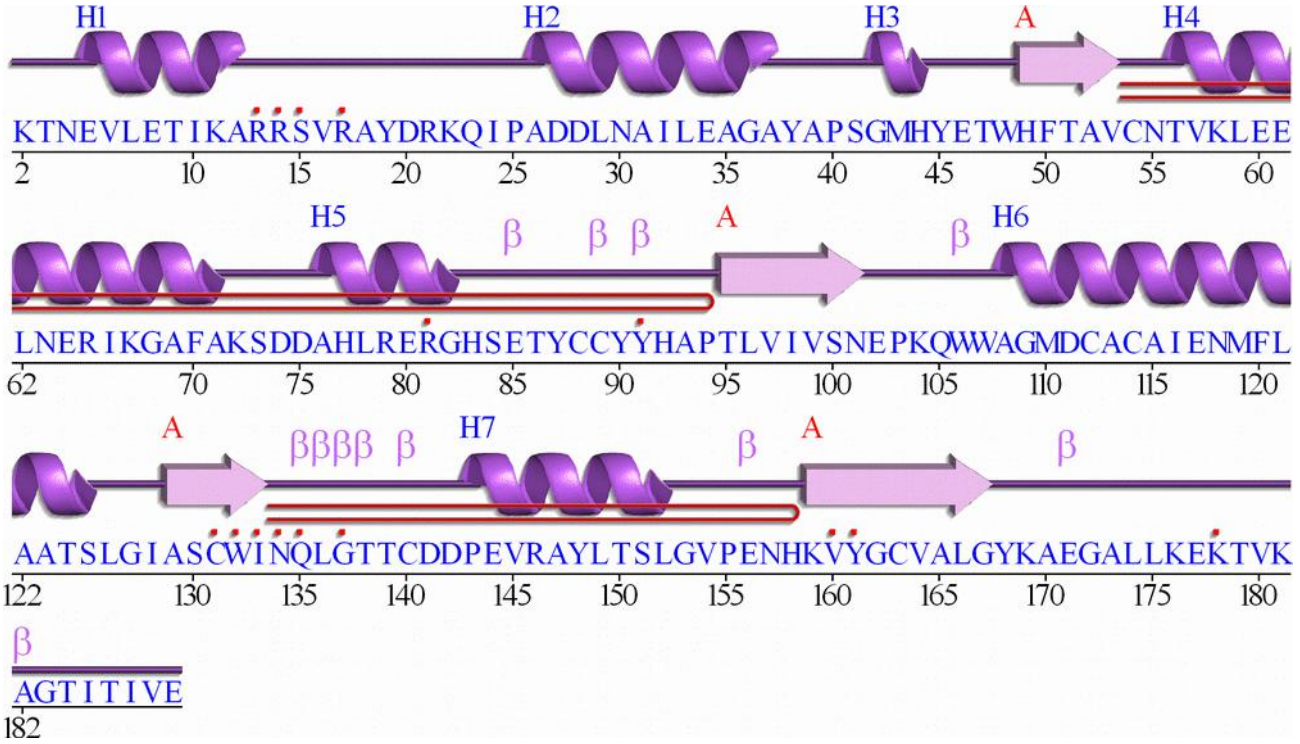
• Ne legyenek atom-atom ütközések

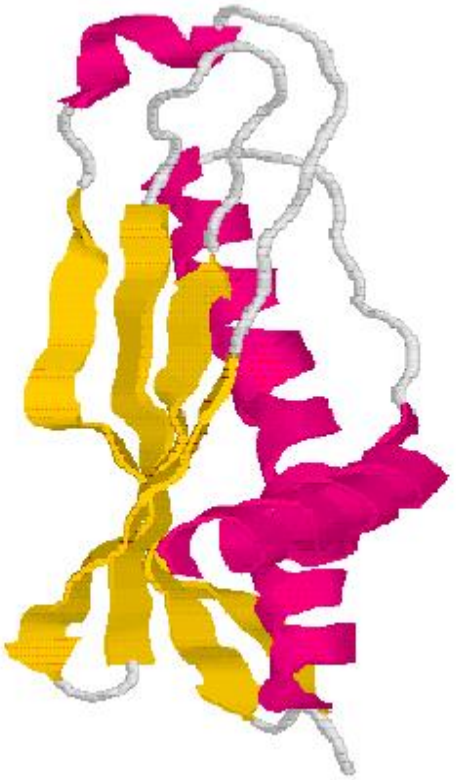
• Eltemetett amidcsoportok többsége H-kötésben legyen

Másodlagos szerkezeti elemek hozzárendelése

Definíciók és megfontolások

- Másodlagos szerkezet a polipeptidlánc helyi (lokális) konformációja, minden szakasznak van!
- Sokszor csak az α-hélixekre és β-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 szerkezetei elemek "valóságába" vetett hitet a homológ szerkezetek összevetéséből nyert kép erősíti
- Miért fontos? Megjelenítés, jellemzés, térszerkezetjóslás (ennek egyik kulcslépése a másodlagos szerkezeti elemek predikciója, ami egyre pontosabb)



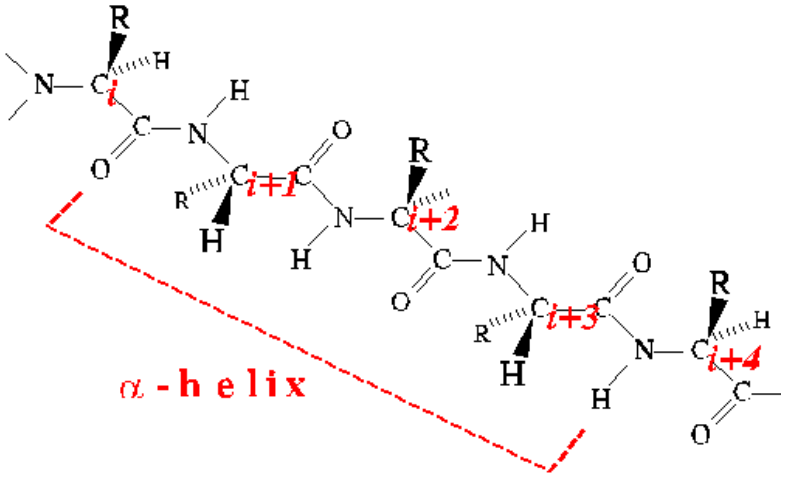


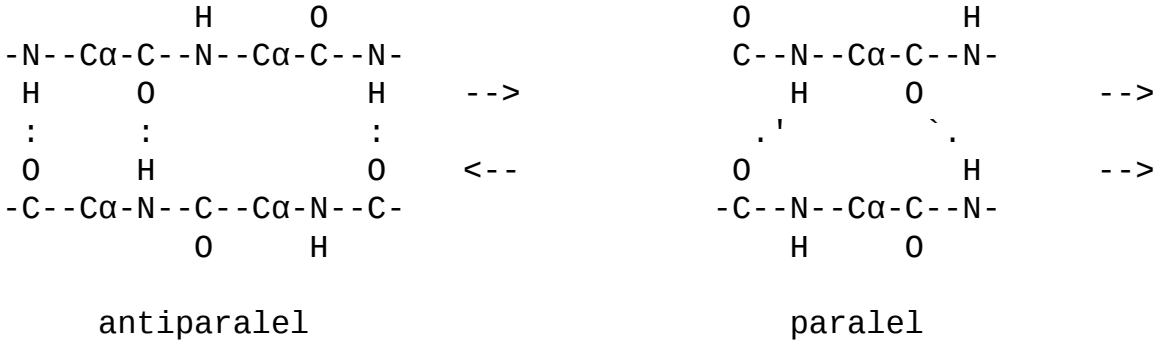
Másodlagos szerkezeti elemek hozzárendelése

• A másodlagos szerkezeti elemeg eredeti definíciója (Linus Pauling!) a hidrogénkötés-mintázatokon alapul

• *De facto* standard: DSSP (Dictionary of Secondary Structure of Proteins): csak hidrogénkötés-mintázatokat néz:

  • Az α-hélix kritériuma, hogy két *i* -> *i+4* hidrogénkötéssel kezdődjön (azaz két egymás utáni aminosav CO csoportja alkosson a néggyel utána következő NH csoportjával hidrogénkötést), és két *i-4* <- *i* hidrogénkötéssel végződjön





β-redőzött réteg

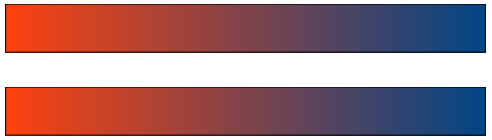
  • A β-szálba sorolás feltétele, hogy vagy az adott aminosav létesítsen két megfelelő H-kötést, vagy két H-kötéssel legyen körülvéve. A szálak helyzete lehet antiparalel vagy paralel.

• Egyéb programok a molekulagerinc torziós szögeit is figyelembe vehetik
(ezáltal pl. hosszabb hélixeket/szálakat tudnak azonosítani)

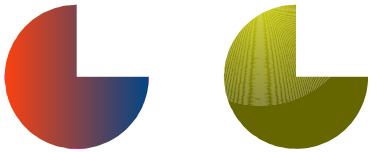
Lokális és globális hasonlóság a bioinformatikában

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

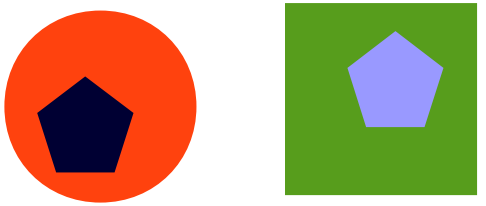
Globális

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 Cα 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 (=> 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 (↔ 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

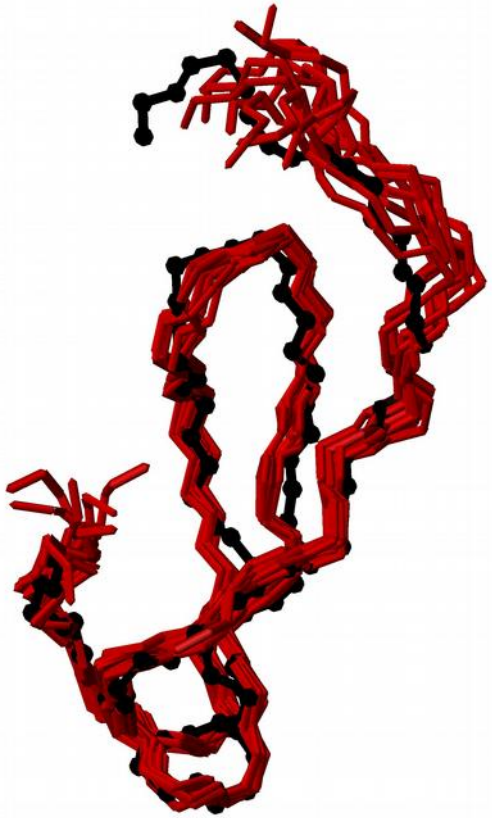
A gyakorlatban a szekvenciaillesztés annyira gyors, hogy mindig azt használjuk...

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

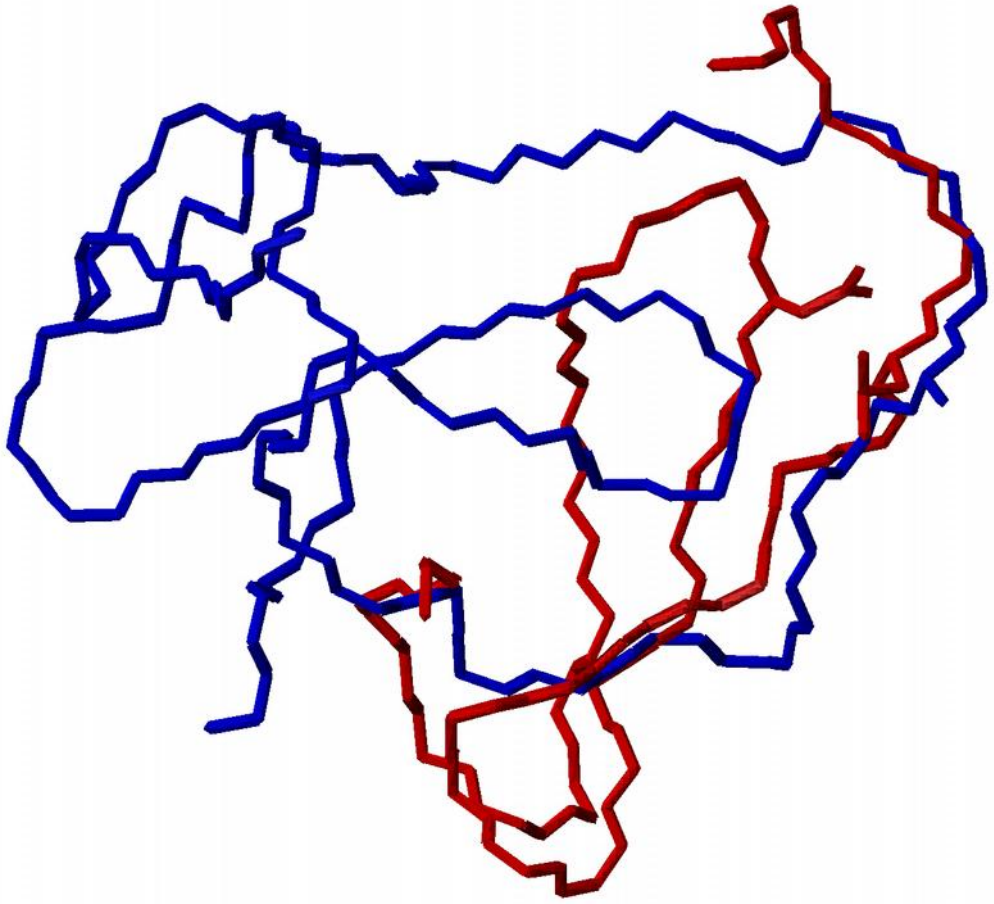
EEEVTKRKAIDEEEERRRKS

RRREEEESKKREEDEYK

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 eltérésének gyöke)



Szerkezetillesztés (structure alignment)
az egymásnak megfelelő atomokat keressük
(pl.kanonikus szerinproteáz-inhibitorok enzimkötő régiójának illesztése)
jóság megítélése??
A szekvenciaillesztéssel ellentétben nincs egyértelmű megoldása (NP-nehéz probléma, sok különböző megvalósítás)