Bioinformatika: bevezetés

Gáspári Zoltán, 2020

gaspari.zoltan@itk.ppke.hu

Bioinformatika, 2019/20. tavaszi félév

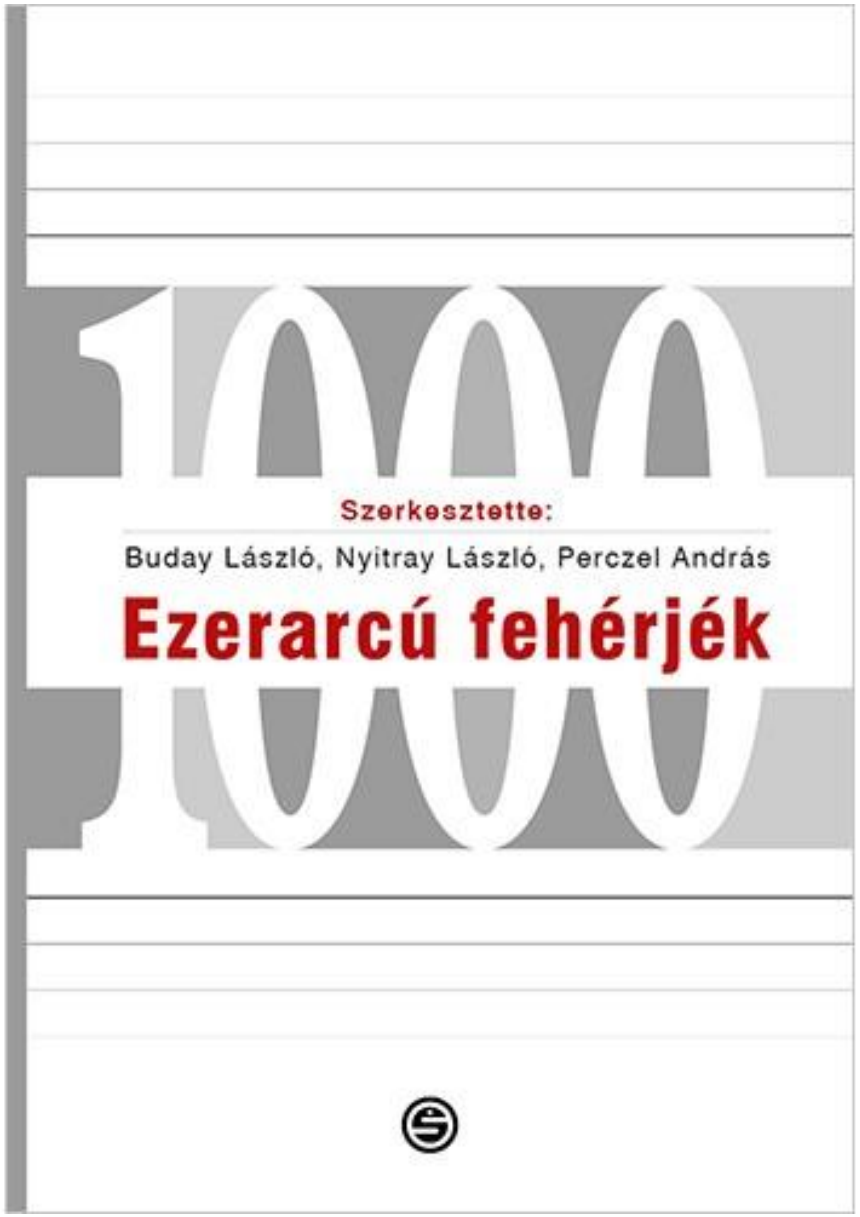
dátum	előadó	téma
feb 12	Gáspári Zoltán	Bevezetés, szekvenciaillesztés
feb 19	Gáspári Zoltán	Proteomikai adatbázisok
feb 26	Gáspári Zoltán	Szerkezeti bioinfo
már 4	Tantos Ágnes	rendezetlen fehérjék I
már 11	Tantos Ágnes	rendezetlen fehérjék II
már 18	Grolmusz Vince	Hálózatok I
már 25	Grolmusz Vince	Hálózatok II
ápr 1		*ZH I (az első 5 óra anyagából)*
ápr 8	Békési Angéla	NGS bevezető, technikák és alkalmazási területek
ápr 22	Ligeti Balázs	NGS adatprocesszálás I
ápr 29	Ligeti Balázs	NGS adatprocesszálás II és esettanulmányok
máj 6	Békési Angéla	“mindennapi bioinformatika” - rutin alkalmazások online eszközökkel
máj 13		*ZH II (az utolsó 6 óra anyagából)*
máj 20		*pót ZH (mindkét ZH-ra)*

A tárgy teljesítéséhez mindkét ZH legalább elégséges jegyre megírása szükséges. A pótZH ideje 2 óra, .

Tárgyfelelős: Vértessy Beáta, tanszékvezető egyetemi tanár
Kapcsolattartó: Békési Angéla, *angela.bekesi@gmail.com*

Tantárgy weboldala:
http://oktatas.ch.bme.hu/oktatas/konyvek/mezgaz/BMEVEMBM103_Bioinformatika/

Bioinformatikai tankönyvek



Buday, Nyitray, Perczel (szerk): Ezerarcú fehérjék (Bioinformatika fejezet)

Antal Péter (szerk): Bioinformatika

Mi a bioinformatika?

Sokféle meghatározás létezik, melyek általában említik a számítógépeket és speciális területeket. Személyes véleményem szerint a legjobb meghatározás az, hogy a bioinformatika **a biológiai adatok feldolgozásának és értelmezésének a tudománya**.

Ehhez ma, a XXI. sz. elején **számítástechnikai eszközök**et használunk, azonban a gépek nem oldják meg helyettünk a feladatokat, csak **segítséget** nyújtanak hozzá.

Gépekre az **adatok mennyisége** és a **számítások összetettsége** miatt van szükségünk.

A bemeneti adatok megfelelő **előkészítés**e és a kimenet **biológiai** értelmének, **jelentőség**ének meghatározása a **kutató feladata**!

A bioinformatikai elemzés nem ér véget az adott program futásának befejezésével, hanem valójában akkor kezdődik el a lényegi része:

- *Mit hihetek el a kapott adatokból és mit nem? (Biológiai tudás / algoritmuosk korlátainak ismerete / józan ész!)*
- *Milyen újabb elemzésekkel tudom megerősíteni/megcáfolni az első vizsgálatok alapján kapott képet?*
- *Ha két módszer ellentmondó eredményeket ad, melyiknek higgyek? Esetleg egyiknek sem?*
- *Milyen, biológialag értékes és használható új információhoz jutottam?*

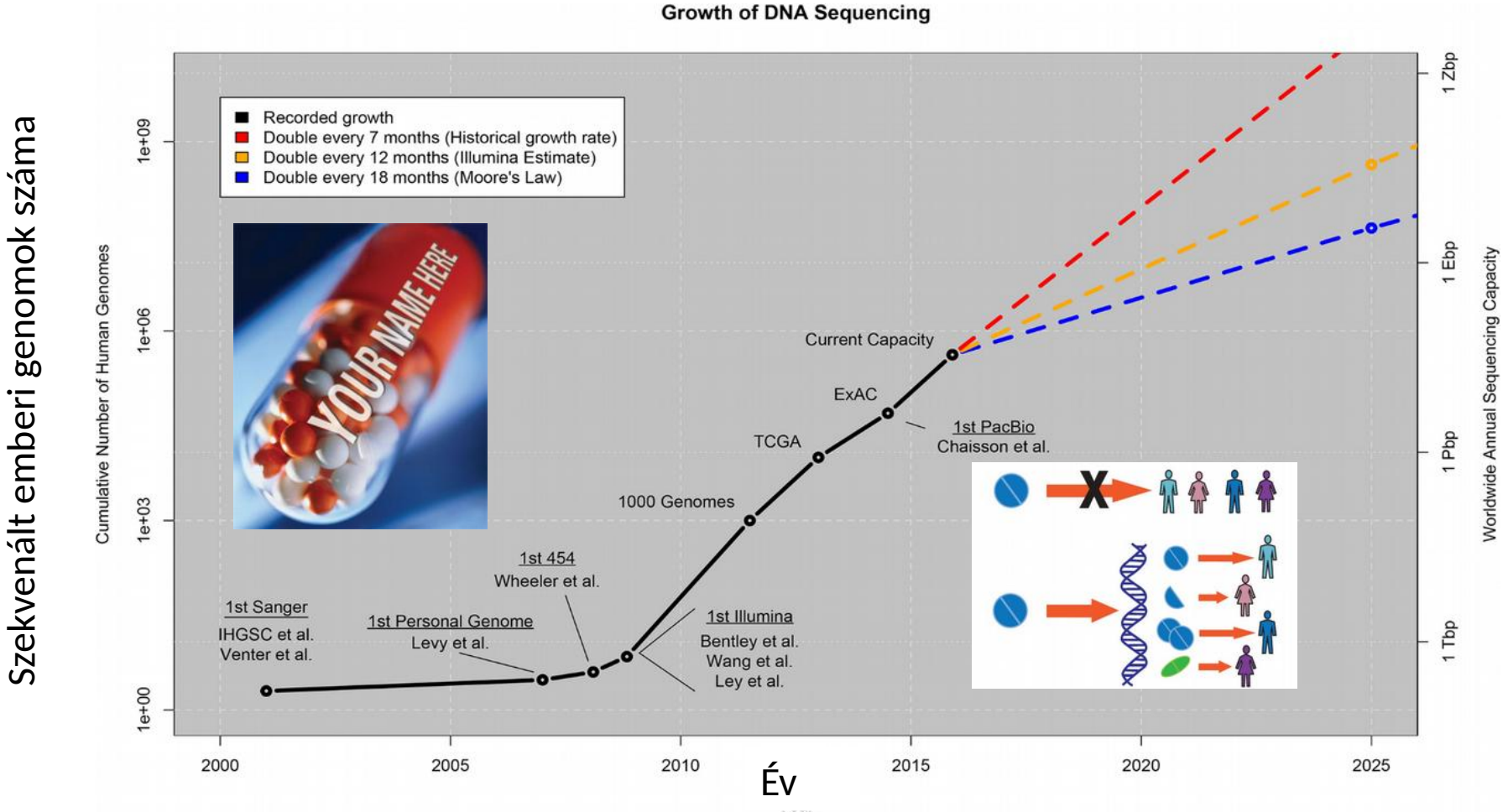
Tipikus bioinformatikai kérdések

(szubjektív lista → rokon területekre való utalásokkal)

- **Adott szekvenciához milyen funkció / biológiai jelentőség tartozik?**
  - Hány és milyen fehérjét kódol adott genom/genomi szakasz? (**génpredikció, genomannotáció**)
  - Adott fehérje milyen szerkezettel/aktivitással rendelkezhet? (**szerkezet/funkció predikció**)
  - Milyen fizikai kötőpartnerei lehetnek?
  - Milyen más génekkel szabályozódhat együtt?
- **Két szekvencia / genom között mely különbségek felelősek egyes funkcionális eltérésekért?**
  - Milyen genetikai háttér milyen betegségekre hajlamosít? (**GWAS**)
  - Mivel érdemes kezelni adott betegséget? (**személyre szabott gyógyítás**)
  - Miért működik egy fehérje máshogyan, mint egy másik hasonló?
- **Két sejt génexpressziós/epigenetikai/splicing stb. mintázata között mely különbségeknek van biológiai jelentősége?**
  - Milyen funkcióval bírnak az együtt szabályozott fehérjék/DNS-szekvenciák? (**enrichment analysis**)
  - Mely változásoknak van a legnagyobb jelentősége? (**adatbányászat**)
    Mely változások a kiváltó okok és melyek a következmények?
- **Adott gén/fehérje működése hogyan befolyásolható?**
  - Milyen szabályozó mechanizmusok megléte valószínűsíthető adott genetikai környezetben? (**genomannotáció,** → **rendszerbiológia**)
  - Milyen módon befolyásolható a fehérjeműködés a szerkezet ismeretében? (→**gyógyszertervezés, biotechnológia**)
- **Hogyan tervezzek adott szerkezettel/funkcióval bíró szekvenciát?**
  - Hogyan állítsak össze egy általam kívánt szabályozással bíró rendszert? (→ **szintetikus biológia**)
  - Hogyan tervezzek adott szerkezetű / funkciójú fehérjét?

Bioinformatikai kihívások a XXI. század elején

Genomszekvenálástól a személyre szabott terápiáig

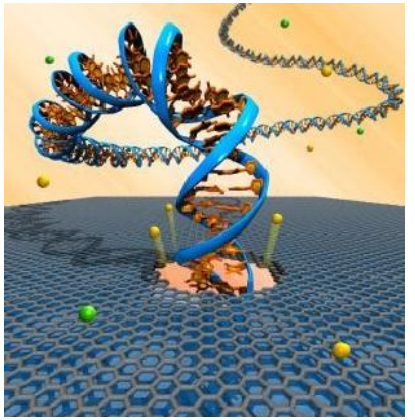


Bioinformatikai kihívások a XXI. század elején

Szekvenálási technikák fejlődése

Minion

MinION

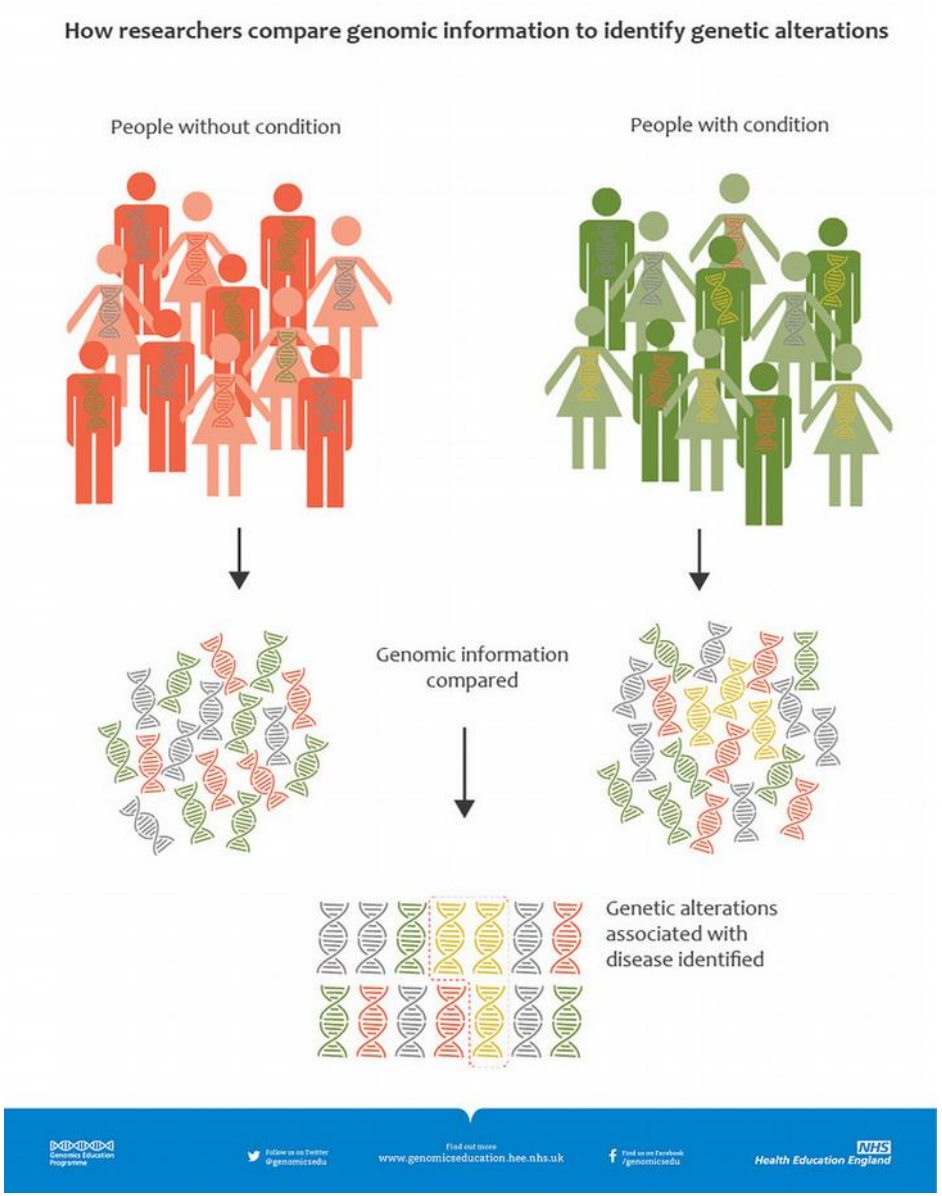


DNS-szekvenálás:

- Ár ↓, volumen ↑
- Egyedi molekulák (SMRT)
- Hordozhatóság

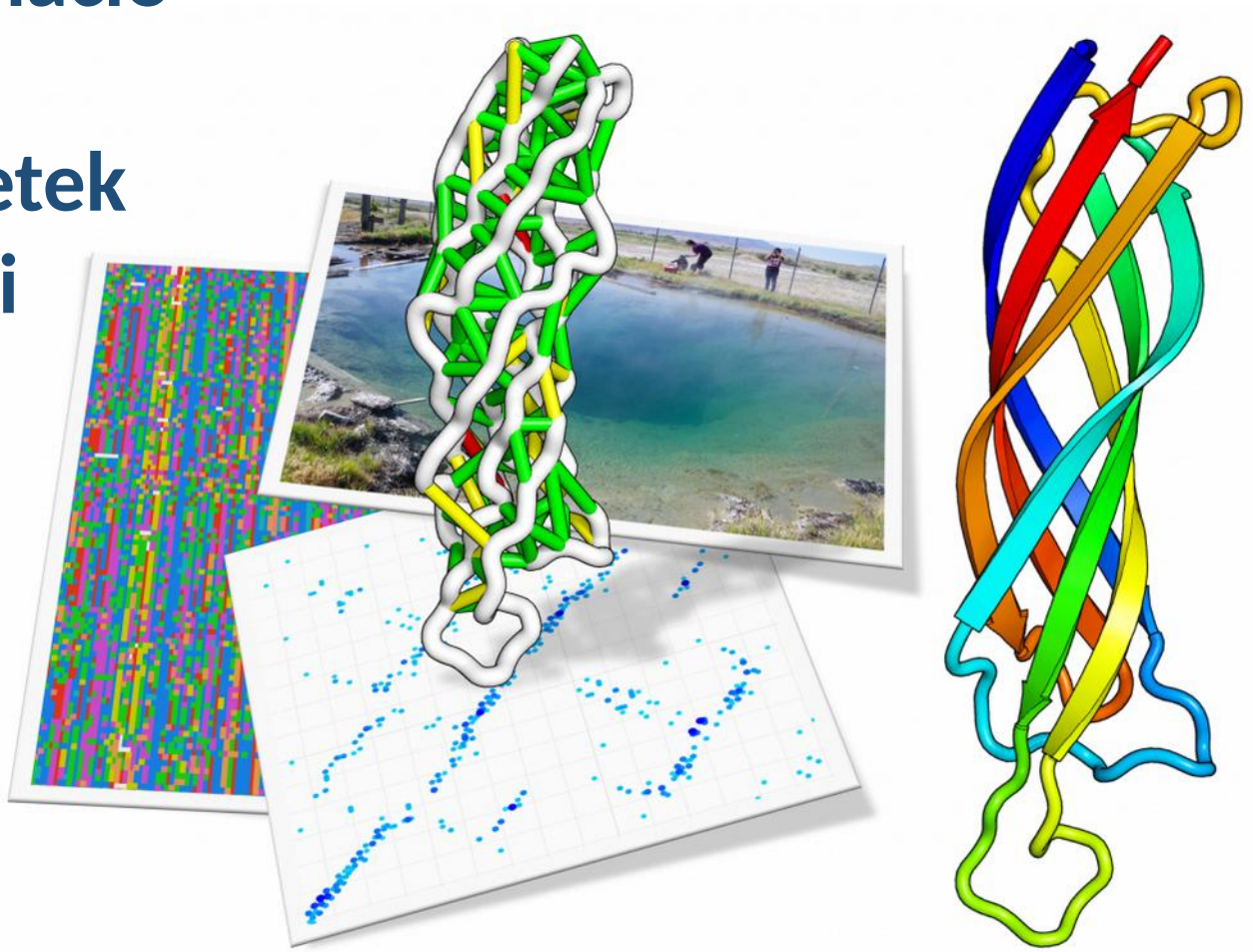
Bioinformatikai kihívások a XXI. század elején

Klinikailag releváns információ kinyerése: genomszintű asszociációs vizsgálatok



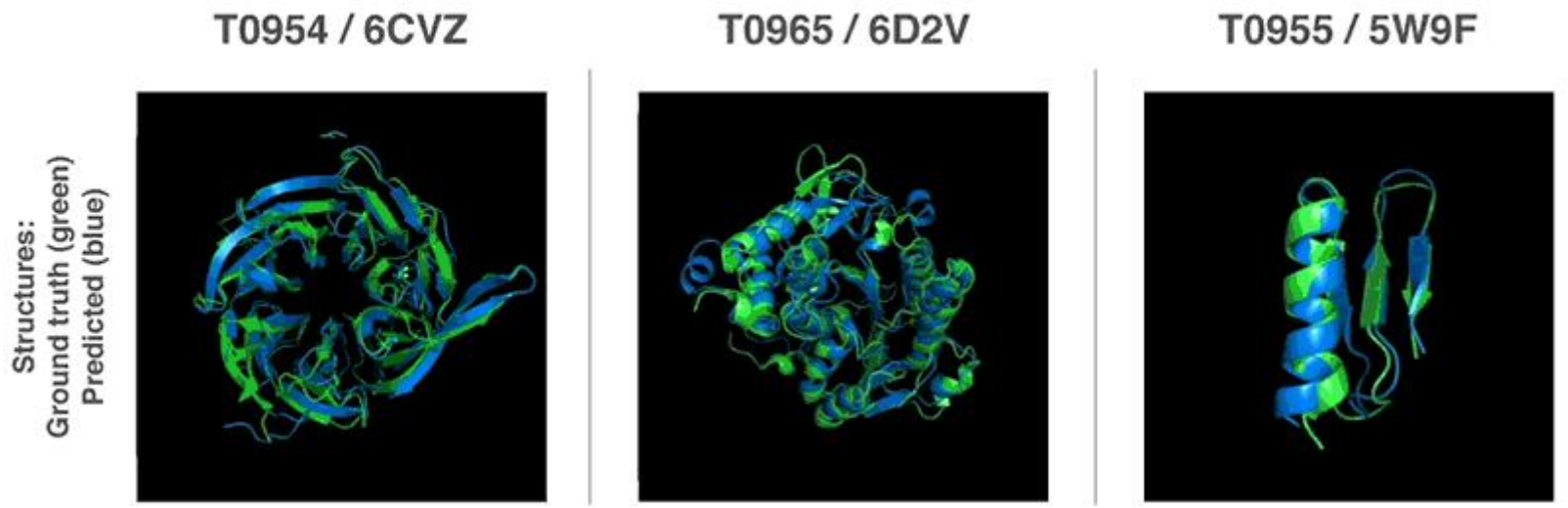
Bioinformatikai kihívások a XXI. század elején

1D -> 3D információ kinyerése: fehérjeszerkezetek metagenomikai adatokból

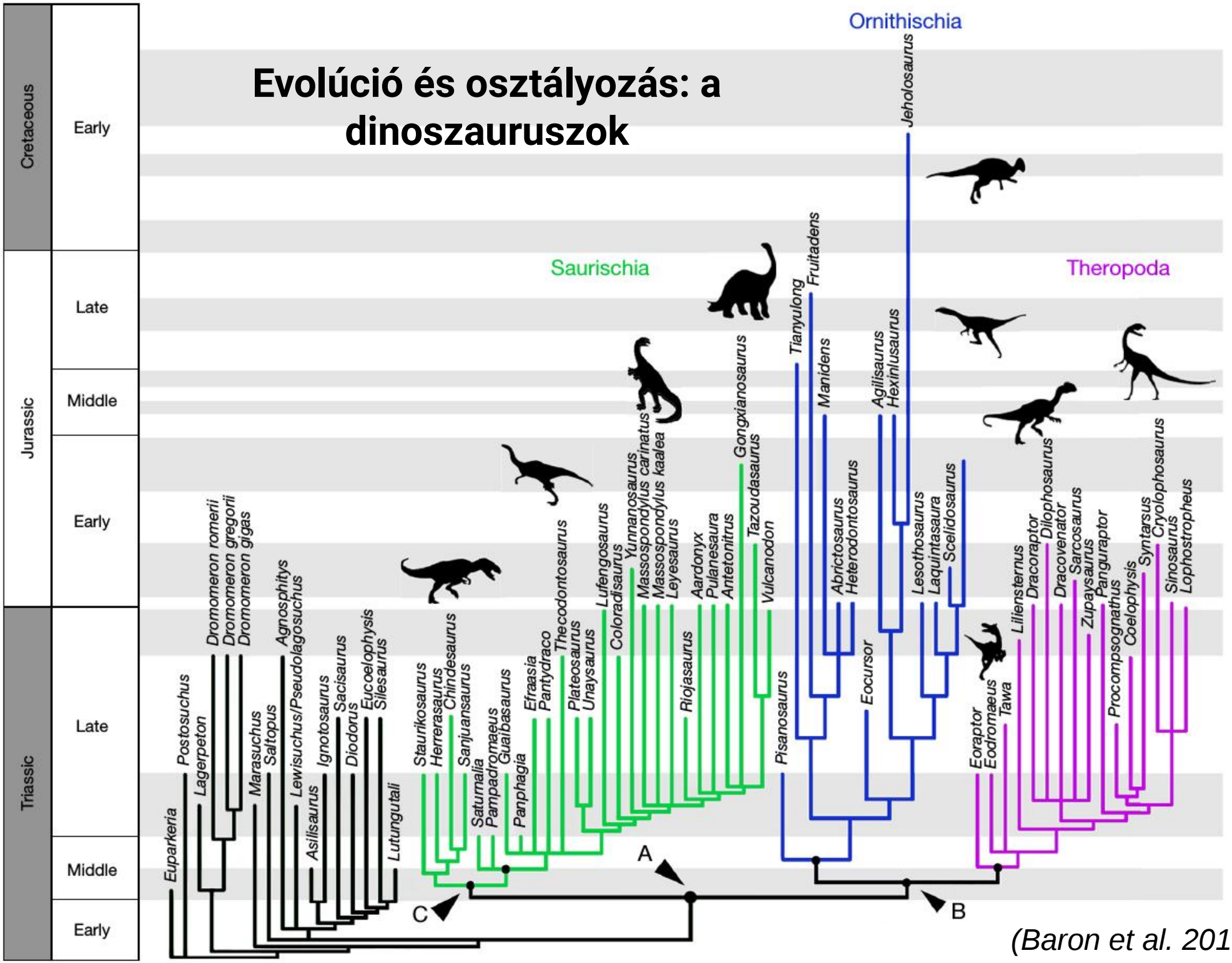


Bioinformatikai kihívások a XXI. század elején

1D -> 3D információ predikciója mesterséges intelligenciával



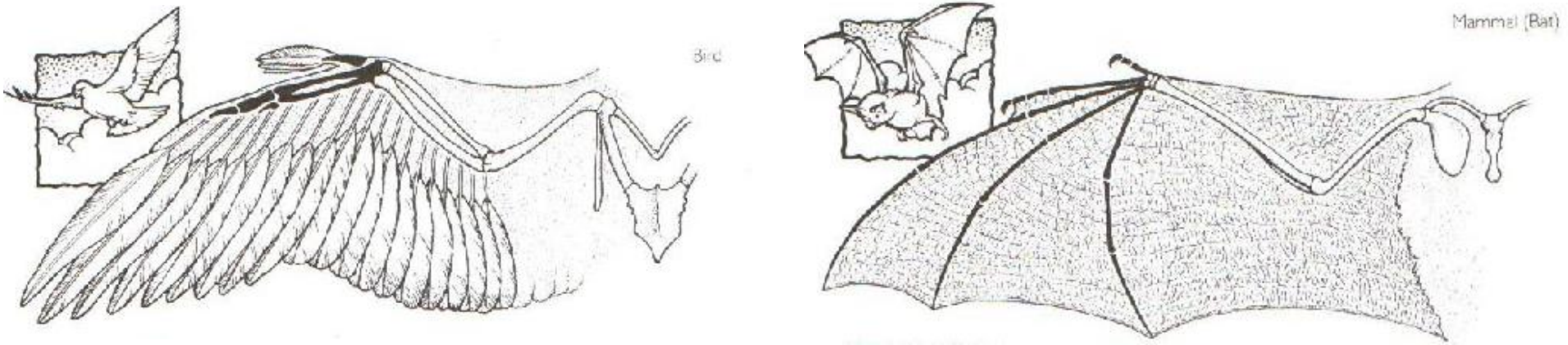
Evolúció és osztályozás: a dinoszauruszok



(Baron et al. 2017)

Homológia és analógia

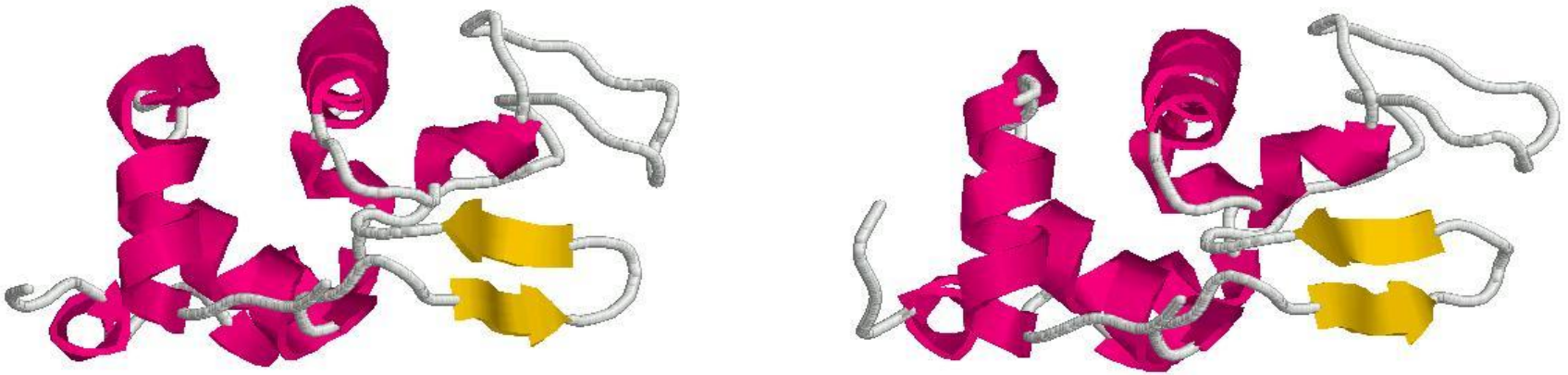
A **homológia** evolúciós rokonságot jelent. Két szerv, csont, sejt, gén vagy fehérje akkor homológ, ha közös őstől származnak. Fontos, hogy ez önmagában nem feltétlenül jelent egyébfajta, pl. funkcionális vagy alaki hasonlóságot, csak a vizsgált képletek történetére vonatkozik. A homológia megállapítása nem feltétlenül triviális feladat.
Az **analógia** ezzel szemben alaki vagy funkcionális hasonlóságot jelent, leszármazási történettől függetlenül.



A homológia és analógia komplex viszonyban lehetnek egymással, attól függően, hogy milyen szinten vizsgáljuk az adott biológiai objektumot.
Példa: a madarak és a denevérek szárnya

- a két szárny **mint teljes végtag** egymással **homológ** (a két élőlény utolsó közös ősének mellső végtagjából származtathatóak)
- egyúttal, mint **teljes szárnyak**, **analógok** is (repülésre használatosak)
- ugyanakkor a szárnyak egyes **részei nem homológok** (pl. a szárnyak vége: toll vs. 3. ujj)
- és az egyes **homológ csontok nem analóg** helyen vannak a szárnyakon belül
- ennek oka, hogy a legutolsó közös ős mellső végtagja még nem volt szárny, a két csoportban ezek egymástól **függetlenül** fejlődtek szárnnyá (**konvergencia**)

Homológia és analógia a gének és fehérjék világában



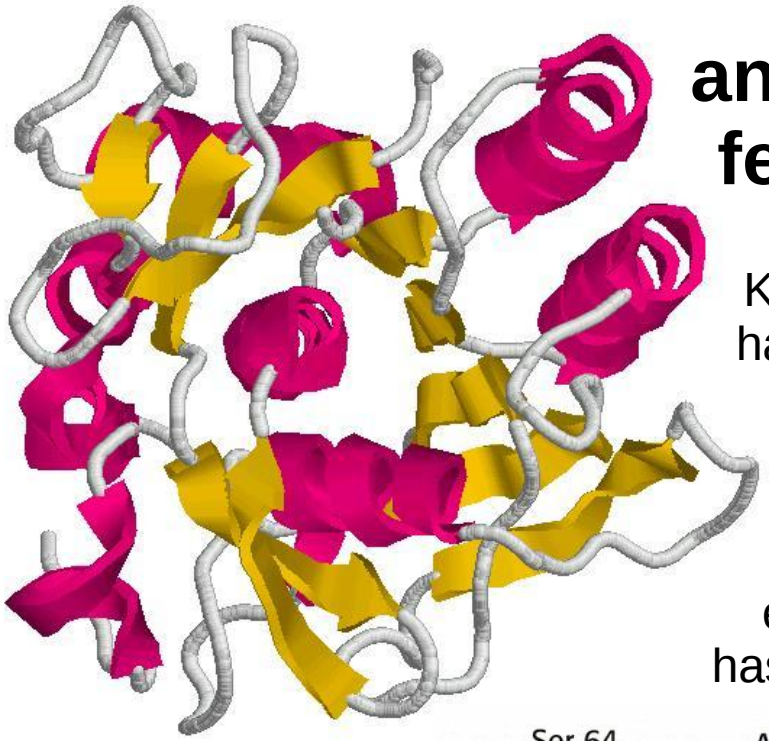
```
LALBA_HUMAN   1    MRFFVPLFLVGILFPAILAKQFTKCELSQLLK--DIDGYGGIALPELICTMFHTSGYDTQ  58
                   M+  + L LV +L   +  K F +CEL++ LK   +DGY GI+L   +C     SGY+T+
LYSC_HUMAN    1    MKALIVLGLV-LLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTR  59

LALBA_HUMAN   59   AIVEN--NESTEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDDDITDDIMCAKKIL-  115
                   A   N  + ST+YG+FQI+++ WC   + P + N C +SC   L D+I D + CAK+++
LYSC_HUMAN    60   ATNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVR  119

LALBA_HUMAN   116  DIKGIDYWLAHKALCTEK  133
                   D +GI  W+A +  C  +
LYSC_HUMAN    120  DPQGIRAWVAWRNRCQNR  137
```

A lizozim (balra) és α-laktalbumin (jobbra) kb. 40%-os **szekvenciaazonosság**ot mutatnak. Ezt a **homológia** jelének tekintjük, csakúgy, mint a nagyon **hasonló térszerkezetet**: a közös eredet a legegyszerűbb tudományos magyarázat.

A lizozim is **enzim:** baktériumok sejtfalát bontja a védekező mechanizmusok részeként.

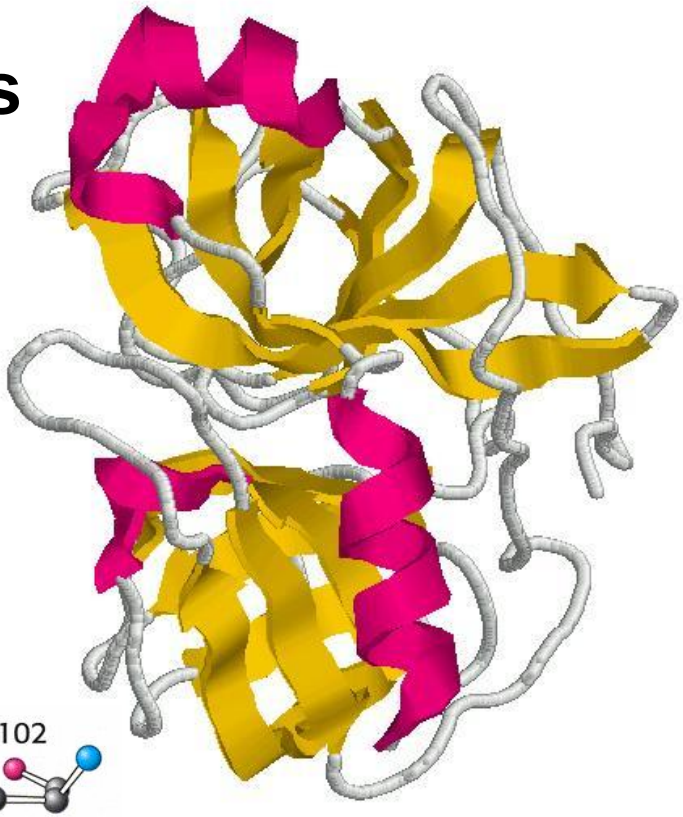
Az α-laktalbumin a galaktoziltranszferáz enzim egyik **regulációs alegysége**, mely lehetővé teszi, hogy a tejmirigyben glükózt is felismerjen az enzim, amely ezáltal képes laktózt előállítani. Az α-laktalbumin a tejbe is átkerülő fehérje.

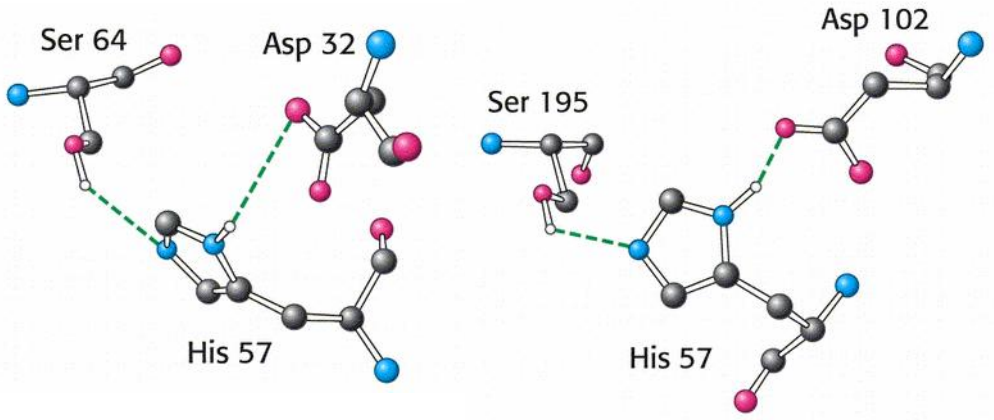
Homológia és analógia a gének és fehérjék világában



Különböző 3D szerkezet, de hasonló lokális elrendeződés az aktív centrumban

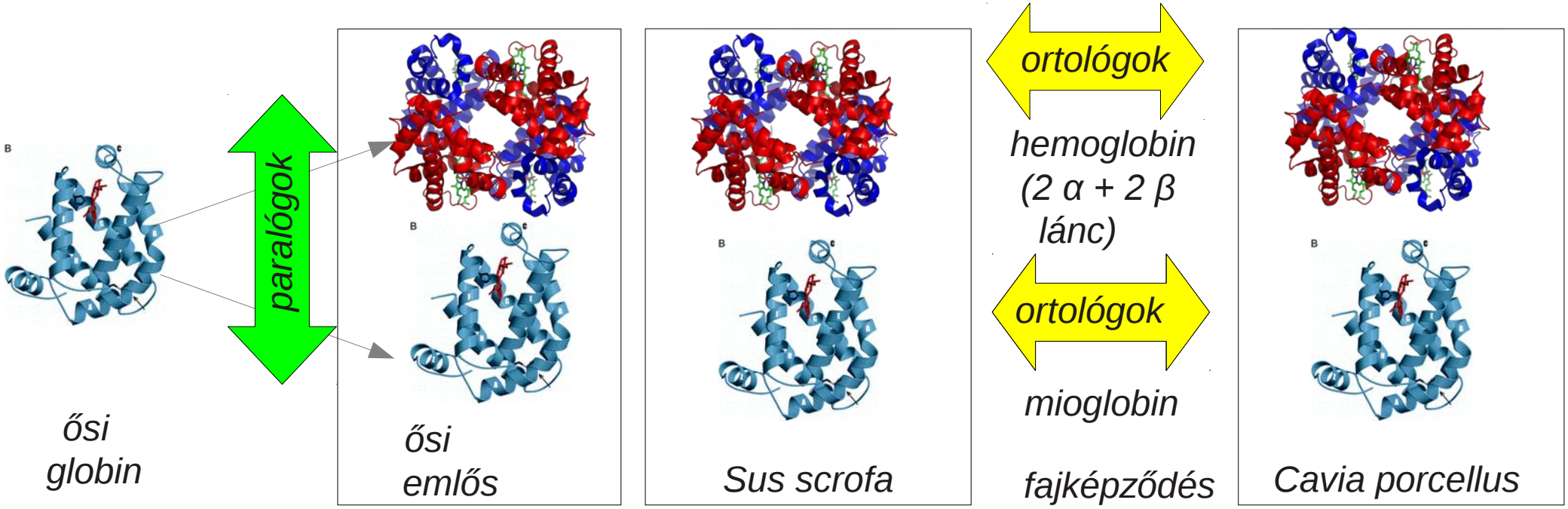
eltérő evolúciós eredet, de hasonló funkció (konvergencia)





A szubtilizin (balra) és a kimotripszin (jobbra) **szerin protázok,** melyek jellegzetes katalitikus triáddal rendelkeznek: Ser, His és Asp. A két fehérje azonban evolúciósan nem rokon, erre egyértelműen utal a különböző térszerkezetük, és hogy a triád aminosavai a szekvenciákban más sorrendben helyezkednek el.
Valójában több, mint 50(!), egymással rokonságban nem álló szerinproteáz-családot ismerünk.

A homológia alesetei: ortológia és paralógia



Az **ortológ** gének/fehérjék története alapvetően a fajok történetét tükrözi. Ezzel szemben a **paralógok génduplikációval** jönnek létre. A példában a disznó (*Sus scrofa*) fehérjék és azok tengerimalac (*Cavia porcellus*) megfelelői ortológok, míg a

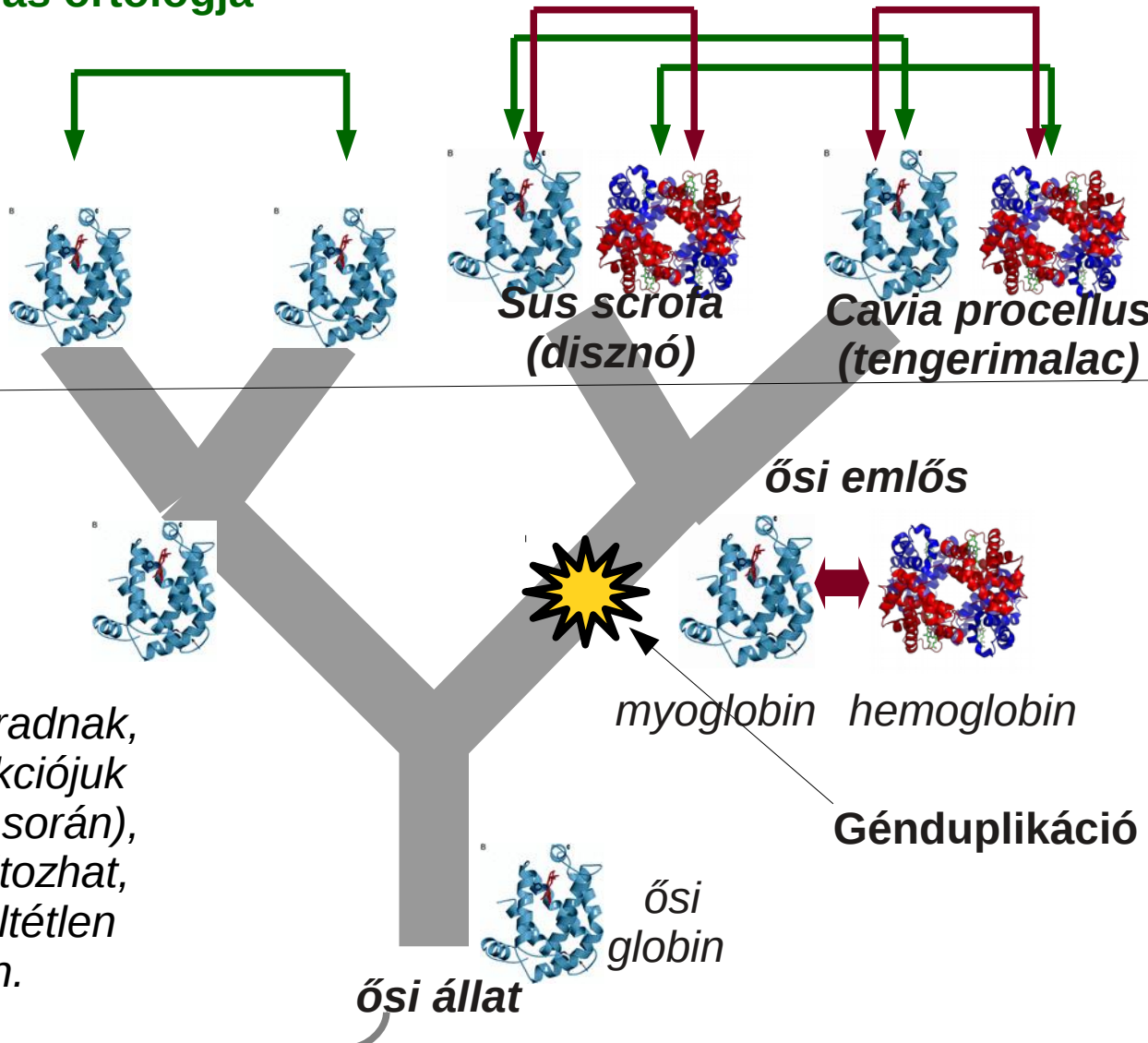
mioglobin - α-hemoglobin,
mioglobin - β -hemoglobin és a
α-hemoglobin – β-hemoglobin

párok paralógok, valamelyest eltérő funkcióval

Ortológia és paralógia

- **minden globin egymás között homológ**
- **minden mioglobin egymás ortológja**
- **minden α/β hemoglobin egymás ortológja**
- **a hemoglobinok és a mioglobinok paralógok**
- **a hemoglobin 2 paralóg alegységet tartalmaz (α és β)**

Ma élő élőlények



Idő/evolúció

Az ortológ gének a genomokban egymásnak megfelelő pozícióban maradnak, és sokszor hasonló a funkciójuk ("megmarad" az evolúció során), míg a paralógoké megváltozhat, de ezek egyikére sincs feltétlen kényszer általánosságban.

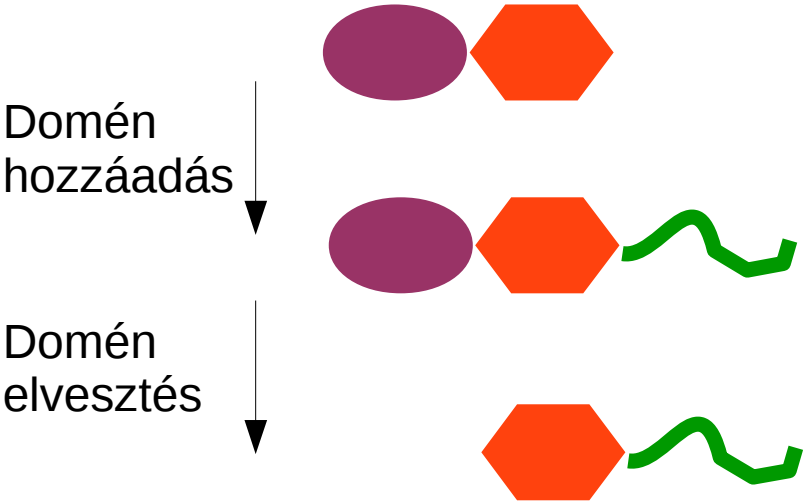
Génduplikáció nélküli leszármazási vonal

Változások a szekvenciában → változások a funkcióban / térszerkezetben

Változás:
- mutáció kémiailag hasonló/eltérő aminosavra

1) AATREEFGHIKNMVDERWA-DT
2) AATRDEFGHIKNMVDERWA-DT
3) AATREEFGHIANMVDERWA-DT
4) AATREEFGHIKNMVDE-WASDT

- deléció vagy inszeció



- nagyobb (akár önálló funkciójú) szakaszok beillesztése/törlése

A mutáció biológiai hatása:

1) nem okoz számottevő változást (=neutrális)
2) funkcióváltozást okoz: ez lehet nyerés vagy vesztés (pl. kötőhely, katalitikus centrum elvész / megjelenik / megváltozik a hatékonysága)
3) megváltozik a fehérje térszerkezete / belső dinamikája / stabilitása → indirekt módon kihat a funkcióra

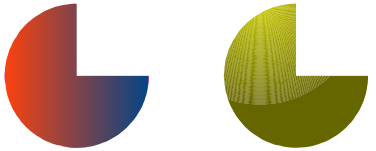
Adott aminosav másikra való cseréjének hatása mindig függ a konkrét fehérjétől és pozíciótól! Ezért a 'kémiai hasonlóság' pontos mibenléte kontextusfüggő!

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**

Lokális

Általában doméneket/motívumokat vizsgálunk

Doméneknél kisebb egységekre értjük általában

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)

Szekvenciák összehasonlítása

Kérdések
- Mennyire hasonlít két szekvencia? → Mekkora a valószínűsége, hogy evolúciósan rokonok?
- Melyek két szekvenciában a hasonló/azonos részek, hol vannak és mik az eltérések?

Alkalmazások
- Páronkénti illesztés (mennyire hasonlít két szekvencia?)
- Többszörös illesztés (mi a közös/eltérés egy fehérjecsalád tagjai között?)
- Keresés adatbázisban (van-e a keresett szekvenciához hasonló az adatbázisban?)

Adattípusok
- Nukleinsav- vagy fehérjeszekvenciák (4 vagy 20 betűs ABC)

Paraméterek
- Hogyan értékeljük a hasonlóságokat és különbségeket? → pontozómátrix
  Attól is függ, hogy mennyire távoli hasonlóságot fogadok el (érzékenység vs. specifitás),
  Távolabbi rokon szekvenciák esetében a hasonlóság mértéke kisebb lehet, de ilyet megengedve nagyobb eséllyel kap vétetlenszerű egyezés is magas pontszámot
- Hogyan kezeljük az inszerciókat/deléciókat → “gap penalty” (résbüntetés)

Algoritmusok
- Egzakt megoldást adó algoritmusok: Needleman-Wunsch, Smith-Waterman
  Adott paraméterek mellett **megtalálják az optimális illesztés**(eke)**t**.
- Gyorsítások (heurisztikák): BLAST, NGS illesztők

Szekvenciák összehasonlítása: globális illesztés

A Needleman-Wunsch algoritmus

- Példa: nukleinsav-szekvenciák (`A,C,G,U`)

- A két szekvencia: `AUGCCAUUG` és `AGCCUCGCU`

- Pontozás (egy nagyon egyszerű eset):

$$M(a,b)=\begin{cases}1, \text{ ha } a = b\\ 0, \text{ ha } a \neq b\end{cases} \qquad G = -1 \text{ (gap)}$$

A mátrix feltöltése:

$$H(i,j)=\max\begin{cases}H(i-1,j-1)+M(a_i, b_j)\\ H(i-1,j)+G\\ H(i,j-1)+G\end{cases}$$

Azaz “átlósan lépve” az M függvényt használjuk, vízszintesen és függőlegesen lépve pedig rést (gap) vezetünk be.

A mátrix első sora és oszlopa feltöltve:

		A	**U**	**G**	**C**	**C**	**A**	**U**	**U**	**G**
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1									
G	-2									
C	-3									
C	-4									
U	-5									
C	-6									
G	-7									
C	-8									
U	-9									

Szekvenciák összehasonlítása: globális illesztés

A Needleman-Wunsch algoritmus

- Példa: nukleinsav-szekvenciák (`A,C,G,U`)
- A két szekvencia: `AUGCCAUUG` és `AGCCUCGCU`

- Pontozás (egy nagyon egyszerű eset):

$$M(a,b)=\begin{cases}1, \text{ ha } a = b\\ 0, \text{ ha } a \neq b\end{cases} \qquad G = -1 \text{ (gap)}$$

A mátrix feltöltése:

$$H(i,j)=\max\begin{cases}H(i-1,j-1)+M(a_i, b_j)\\ H(i-1,j)+G\\ H(i,j-1)+G\end{cases}$$

A mátrix teljesen feltöltve:

A **globális illesztés pontértéke** a **jobb alsó** cellában lévő érték

		A	**U**	**G**	**C**	**C**	**A**	**U**	**U**	**G**
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7
G	-2	0	1	1	0	-1	-2	-3	-4	-5
C	-3	-1	0	1	2	1	0	-1	-2	-3
C	-4	-2	-1	0	2	3	2	1	0	-1
U	-5	-3	0	-1	0	2	3	3	2	1
C	-6	-1	-1	0	0	1	2	3	3	2
G	-7	-2	-1	0	0	0	1	2	3	4
C	-8	-3	-2	-1	1	1	0	1	2	3
U	-9	-4	-2	-2	0	0	1	1	2	**2**

Szekvenciák összehasonlítása: globális illesztés

A Needleman-Wunsch algoritmus

- Példa: nukleinsav-szekvenciák (A,C,G,U)
- A két szekvencia: AUGCCAUUG és AGCCUCGCU

- Pontozás (egy nagyon egyszerű eset):

$$M(a,b)=\begin{cases}1, \text{ ha } a = b\\ 0, \text{ ha } a \neq b\end{cases} \qquad G = -1 \text{ (gap)}$$

A mátrix feltöltése:

$$H(i,j)=\max\begin{cases}H(i-1,j-1)+M(a_i, b_j)\\ H(i-1,j)+G\\ H(i,j-1)+G\end{cases}$$

		A	U	G	C	C	A	U	U	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	**1**	0	-1	-2	-3	-4	-5	-6	-7
G	-2	0	**1**	**1**	0	-1	-2	-3	-4	-5
C	-3	-1	0	1	**2**	1	0	-1	-2	-3
C	-4	-2	-1	0	2	**3**	2	1	0	-1
U	-5	-3	0	-1	0	2	**3**	**3**	2	1
C	-6	-1	-1	0	0	1	2	3	**3**	2
G	-7	-2	-1	0	0	0	1	2	3	**4**
C	-8	-3	-2	-1	1	1	0	1	2	**3**
U	-9	-4	-2	-2	0	0	1	1	2	**2**

A feltöltött mátrixban **visszakövetjük** a maximális értékeket, egészen a bal felső celláig.

Az útvonal megadja az illesztést:

AUGCCAUUG--

A-GCC-UCGCU

Szekvenciák összehasonlítása: lokális illesztés

A Smith-Waterman algoritmus

- Példa: nukleinsav-szekvenciák (`A,C,G,U`)
- A két szekvencia: `AUGCCAUUG` és `AGCCUCGCU`

- Pontozás (egy nagyon egyszerű eset):

$$M(a,b)=\begin{cases}1, \text{ ha } a = b\\ 0, \text{ ha } a \neq b\end{cases} \qquad G = -1 \text{ (gap)}$$

A mátrix feltöltése:

$$H(i,j)=\max\begin{cases}H(i-1,j-1)+M(a_i, b_j)\\ H(i-1,j)+G\\ H(i,j-1)+G\\ \mathbf{0}\end{cases}$$

Új elem a globális illesztéshez képest a 0!

A mátrix első sora és oszlopa feltöltve:

		A	**U**	**G**	**C**	**C**	**A**	**U**	**U**	**G**
	0	0	0	0	0	0	0	0	0	0
A	0									
G	0									
C	0									
C	0									
U	0									
C	0									
G	0									
C	0									
U	0									

Szekvenciák összehasonlítása: lokális illesztés

A Smith-Waterman algoritmus

- Példa: nukleinsav-szekvenciák (`A,C,G,U`)
- A két szekvencia: `AUGCCAUUG` és `AGCCUCGCU`

- Pontozás (egy nagyon egyszerű eset):

$$M(a,b)=\begin{cases}1, \text{ ha } a = b\\ 0, \text{ ha } a \neq b\end{cases} \qquad G = -1 \text{ (gap)}$$

A mátrix feltöltése:

$$H(i,j)=\max\begin{cases}H(i-1,j-1)+M(a_i, b_j)\\ H(i-1,j)+G\\ H(i,j-1)+G\\ \mathbf{0}\end{cases}$$

A mátrix teljesen feltöltve:

A **globális illesztés pontértéke** a mátrixban előforduló **maximális** érték

(további illesztéseket lehet találni az ebből induló útvonalon kívüli második, harmadik stb. legnagyobb értékekből)

		A	**U**	**G**	**C**	**C**	**A**	**U**	**U**	**G**
	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	0	0	0	0	0	0
G	0	0	1	1	0	0	0	0	0	0
C	0	0	0	1	2	1	0	0	0	0
C	0	0	0	0	2	3	2	0	0	0
U	0	0	0	0	1	2	3	3	1	0
C	0	0	0	0	0	2	2	3	3	1
G	0	0	0	0	0	1	2	2	3	**4**
C	0	0	0	0	0	0	1	2	2	3
U	0	0	0	0	0	0	0	1	2	2

Szekvenciák összehasonlítása: lokális illesztés

A Smith-Waterman algoritmus

- Példa: nukleinsav-szekvenciák (`A,C,G,U`)
- A két szekvencia: `AUGCCAUUG` és `AGCCUCGCU`

- Pontozás (egy nagyon egyszerű eset):

$$M(a,b)=\begin{cases}1, \text{ ha } a = b\\ 0, \text{ ha } a \neq b\end{cases} \qquad G = -1 \text{ (gap)}$$

A mátrix feltöltése:

$$H(i,j)=\max\begin{cases}H(i-1,j-1)+M(a_i, b_j)\\ H(i-1,j)+G\\ H(i,j-1)+G\\ \mathbf{0}\end{cases}$$

A feltöltött mátrixban visszakövetjük maximális értékeket az első nulláig.

Az útvonal megadja az illesztést:

AUGCCAUUG
A–GCC–UCG

		A	**U**	**G**	**C**	**C**	**A**	**U**	**U**	**G**
	0	0	0	0	0	0	0	0	0	0
A	0	**1**	0	0	0	0	0	0	0	0
G	0	0	**1**	**1**	0	0	0	0	0	0
C	0	0	0	1	**2**	1	0	0	0	0
C	0	0	0	0	2	**3**	2	0	0	0
U	0	0	0	0	1	2	**3**	**3**	1	0
C	0	0	0	0	0	2	2	3	**3**	1
G	0	0	0	0	0	1	2	2	3	**4**
C	0	0	0	0	0	0	1	2	2	3
U	0	0	0	0	0	0	0	1	2	2

Érdeklődőknek továbbinduláshoz javasolható a wikipedia, informatív és korrekt a vonatkozó szócikk.

Pontozómátrixok

- Két nukleotid / aminosav hasonlóságát adják meg, 4x4 ill. 20x20 -as mátrixok
- Alkalmazástól / származtatástól függően többféle mátrix létezik, **fehérjék esetében** a legfontosabbak:
    - **PAM (point accepted mutation) sorozat**:

    PAM1: az aminosavak 1/100 részének változását várjuk. A nagyobb számmal jellemzett PAM mátrixok (pl. PAM250) a PAM1 önmagával többször megszorozott (hatványozott) változatai. Azaz nagyobb számú PAM mátrix = nagyobb evolúciós távolság, tehát távolabbi hasonlóság detektálására is alkalmas mátrix

    - **BLOSUM sorozat:**

    Adott szekvenciaazonosságot mutató fehérjerégiók illesztéseiből készült. Pl. a BLOSUM80 a 80%-os azonossággal rendelkező régiók alapján. Itt tehát a nagyobb számú BLOSUM mátrix = nagyobb hasonlóság, közelebbi evolúciós hasonlóság detektálására alkalmas.

Érdeklődők további infót itt találnak.

A két sorozat nagyjából így felel meg egymásnak:

PAM	BLOSUM
PAM250	BLOSUM45
PAM160	BLOSUM62
PAM120	BLOSUM80

Affin résbüntetés

- A példánkban is alkalmazott résbüntetés **minden egyes beszúrásnál** adott értéket von le a pontszámból (azaz egy 10 aminosavas beillesztés tízszerese egy egy aminosavasnak)
- Ez nem különböztet meg sok rövid és néhány nagyobb inszerciót, holott a sok rövid biológiailag sokkal kevésbé valószínű (több evolúciós eseményt feltételez)
- **Biológiailag reálisabb** a a résbüntetést kettébontani:
    - **résnyitás**i büntetés (gap opening penalty)
    - **réskiterjesztés**i büntetés (gap extension penalty)
    - az ún. affin résbüntetés függvénye:

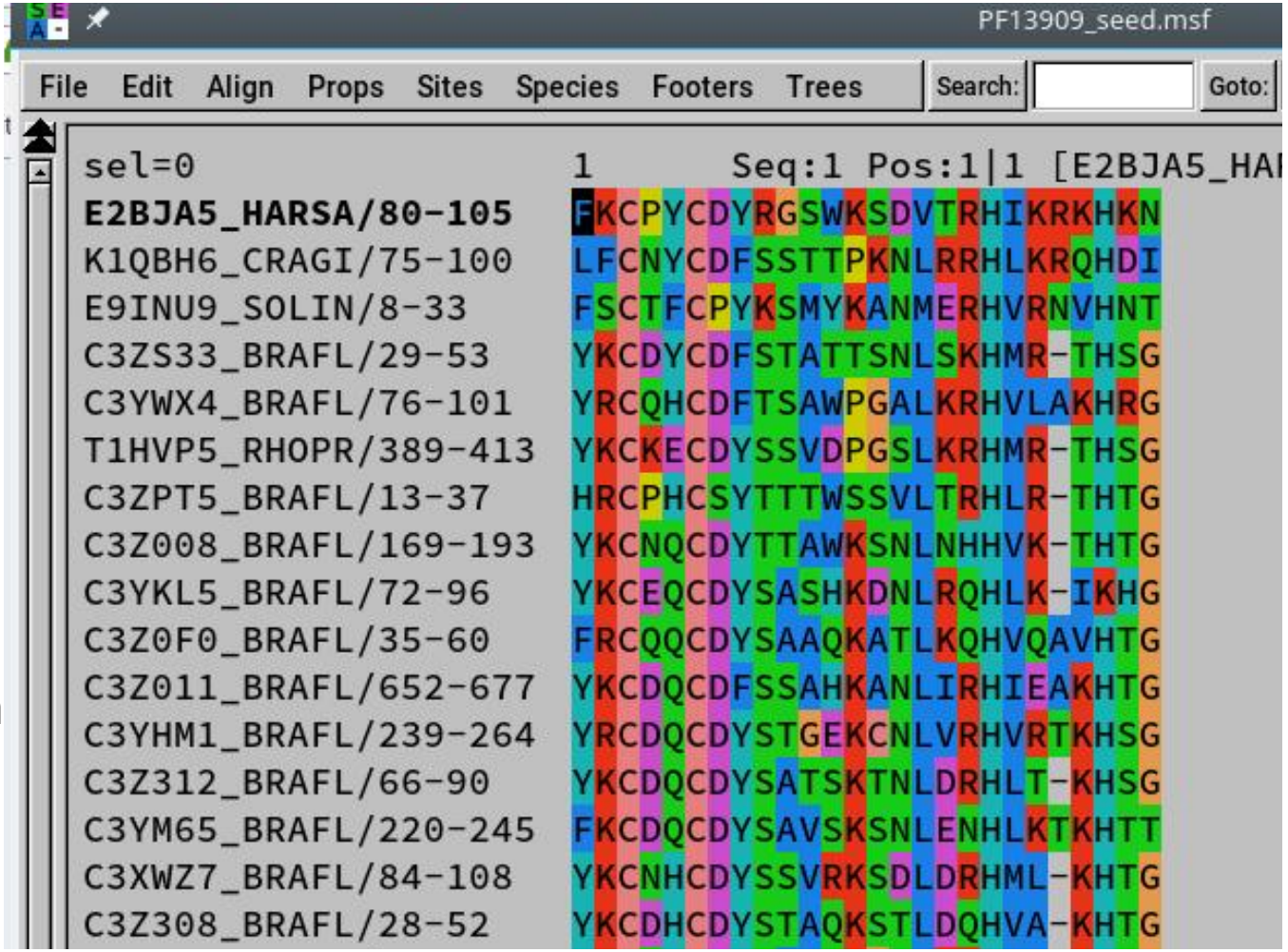
$$G(k)=-O-(k-1)E$$

Ahol k a rés hossza, O a résnyitási büntetés, E a kiterjesztési büntetés, és E < O, azaz a egy rés létrejöttét jobban büntetjük, mint egy meglévő hosszabbítását (több, de hosszabb résnek kedvez)

- Léteznek egyéb büntetési sémák is, ahol nem lineáris a réskiterjesztés, mint az affin esetben.

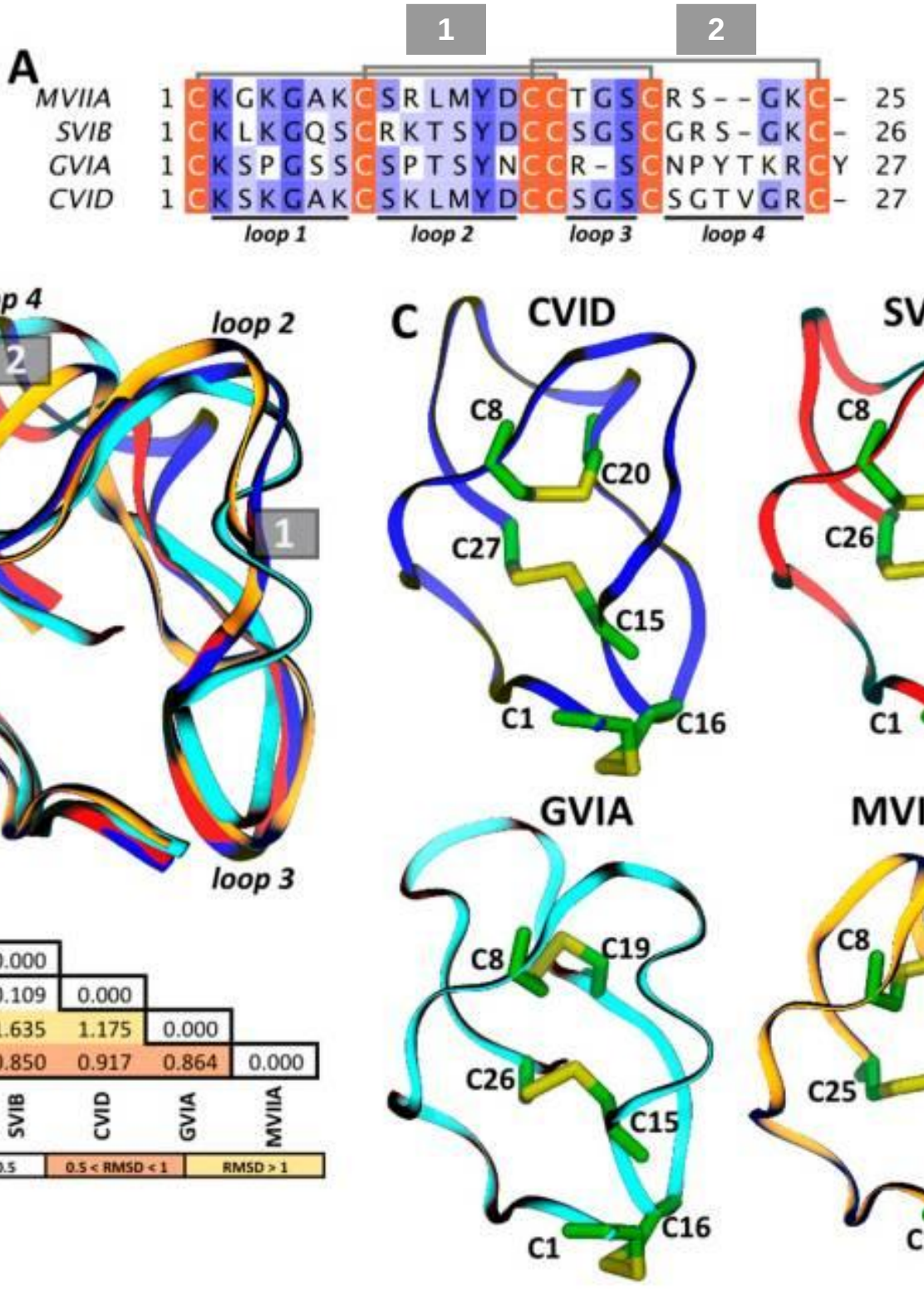
Többszörös szekvenciaillesztés

- Kettőnél több szekvencia illesztése → **informatívabb**, sokkal jobban látszanak az evolúciósan konzervált aminosavak, mint a páronkénti illesztésekben (általában “tömörebb” is)
- Általában **iteratív** módon lehet elvégezni:
  - Egy mindenki mindenki elleni illesztés segítségével meghatározzuk az egymáshoz leginkább hasonló szekvenciákat, majd az ebből eredő **sorrend**ben felépítjük az illesztést
  - Az első két szekvencia illesztéséből egy **profil**t készítünk, a további szekvenciákat ehhez illesztjük a Smith-Waterman algoritmussal analóg módon (profil-szekvencia és profil-profil illesztésekkel)



Szekvenciák összehasonlítása: példa

- **ω-conotoxinok** szekvenciájának és szerkezetének összehasonlítása
- A többszörös illesztés **kiemeli** az evolúciósan **konzervált** aminosavakat, a diszulfidhídkötésben részt vevő ciszteinek tipikusan ilyenek



Többszörös illesztés információtartalmának reprezentációi

Többszörös illesztés

```
AA-GCCTGCG
AATGGATCCG
ATTGC-TCCG
```

Szekvencia logo

grafikus megjelenítés

Szekvenciamintázat

az illesztés alapján készült konszenzus

pl. PROSITE

```
A-[AT]-x(0,1)-G-[GC]-x(0,1)-T-[CG]-C-G
```

Szekvenciaprofil

(N+1) x L – es gyakorisági mátrix

Pozícióspecifikus mátrix!

(A gyakorlatban log-odds változatát használják inkább)

	1	2	3	4	5	6	7	8	9	10
A	1	0.7	0	0	0	0.3	0	0	0	0
T	0	0.3	0.7	0	0	0	1	0	0	0
G	0	0	0	1	0.3	0	0	0.3	0	1
C	0	0	0	0	0.7	0.3	0	0.7	1	0
-	0	0	0.3	0	0	0.4	0	0	0	0

Többszörös illesztés információtartalmának reprezentációi

Többszörös illesztés

```
AA-GCCTGCG
AATGGATCCG
ATTGC-TCCG
```

Rejtett Markov modell

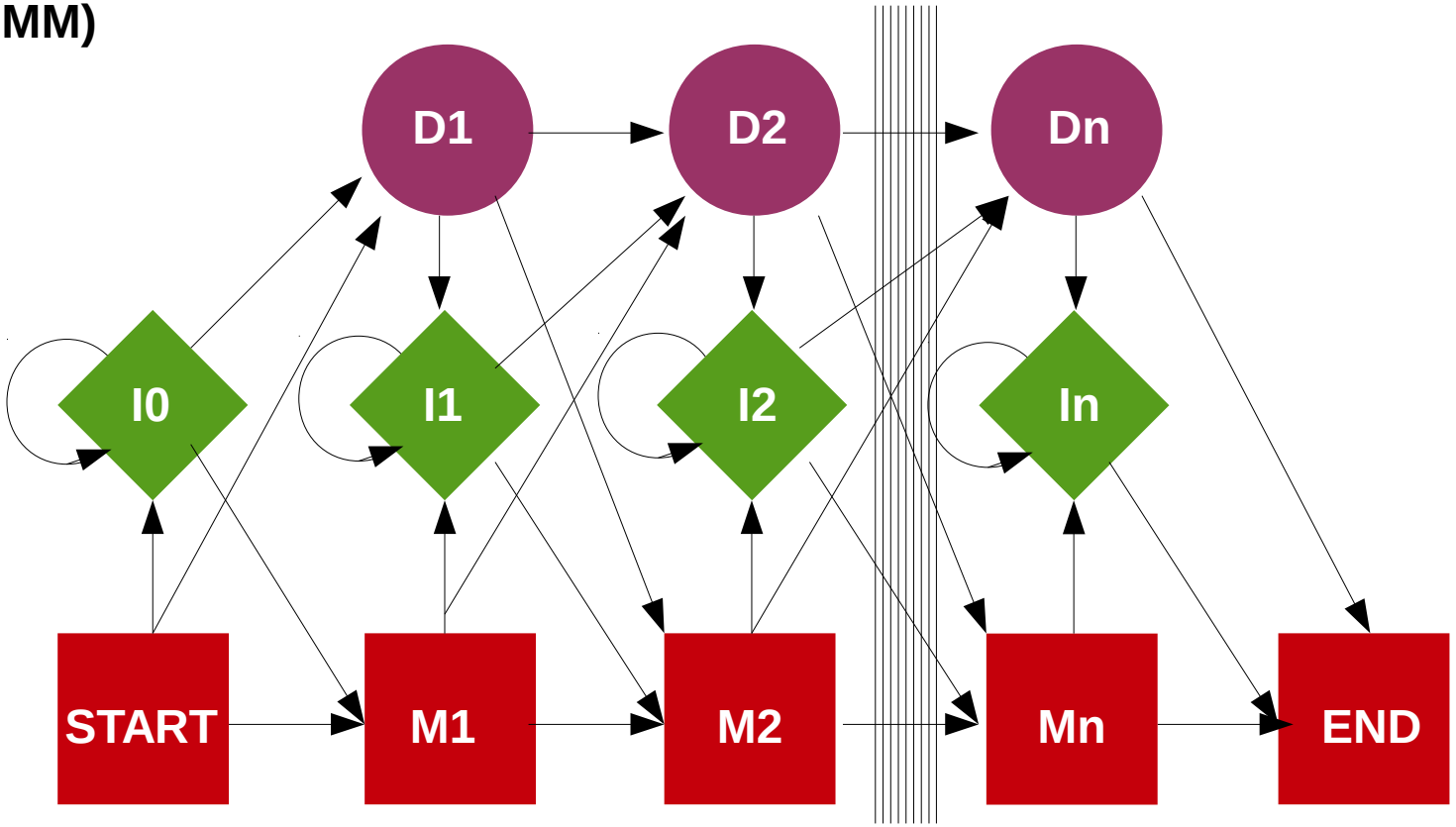
(Hidden Markov Model, HMM)

Átmeneti valószínűségeken alapuló reprezentáció

DELÉCIÓ

INSZERCIÓ

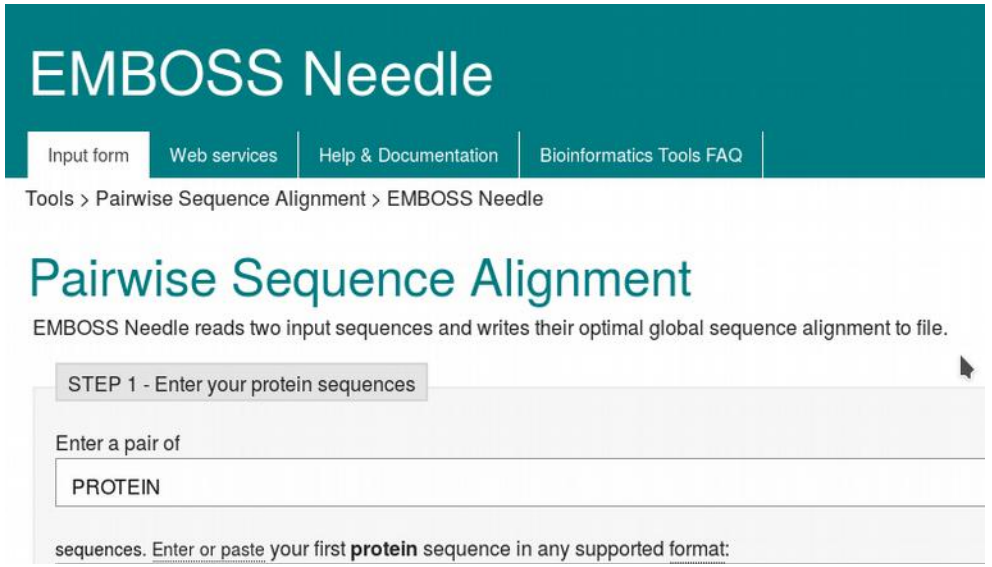
MEGFELELÉS (MATCH / MISMATCH)

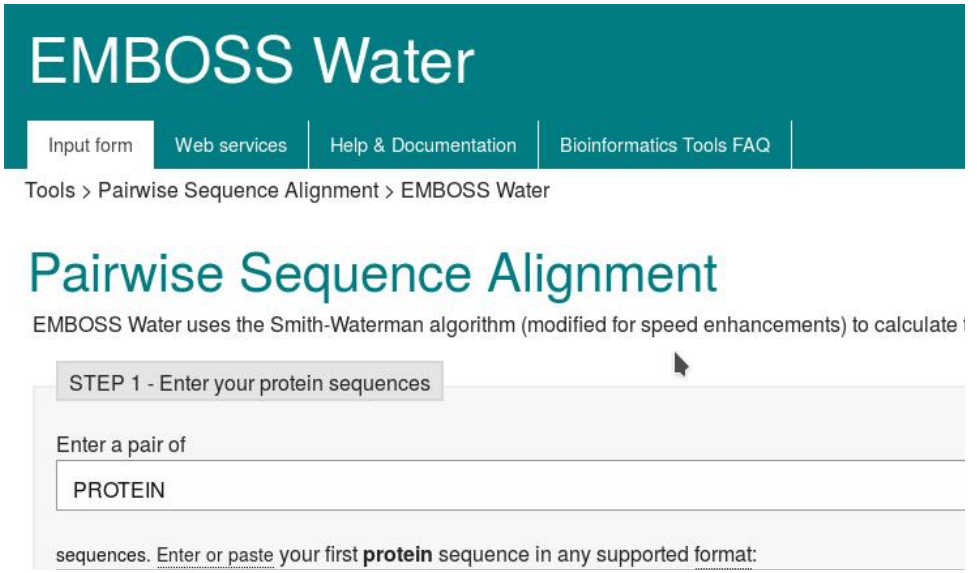


Szekvenciák összehasonlítása

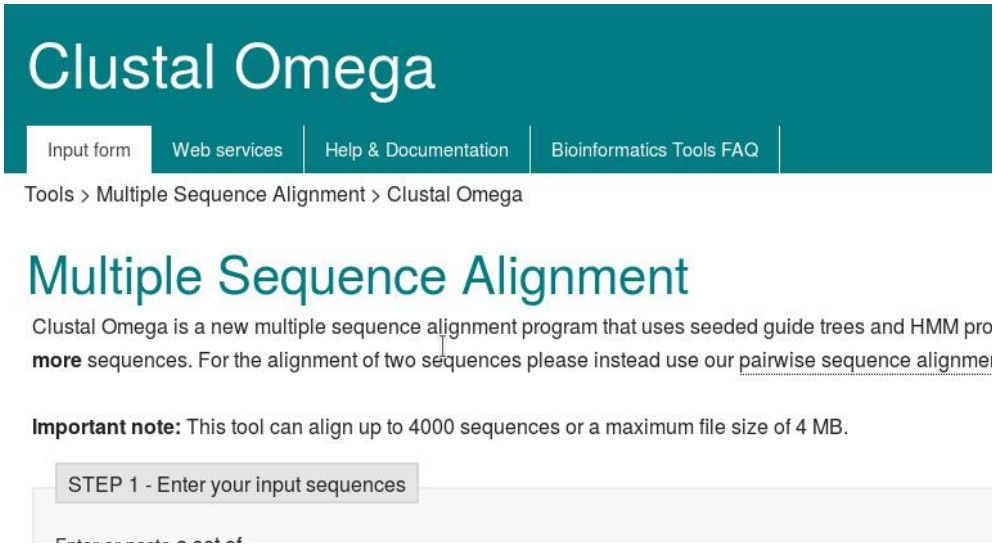
Programok, webszerverek

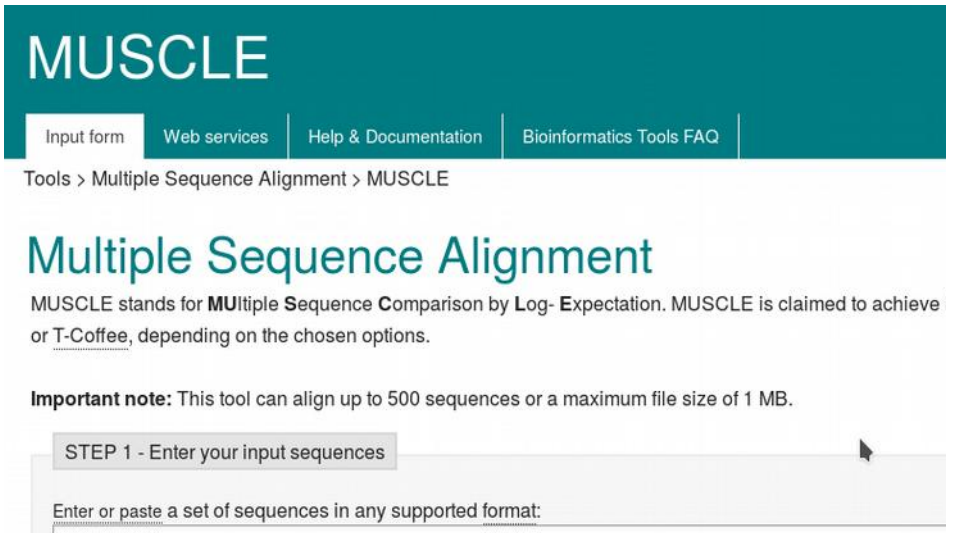
- Páronkénti illesztés





- Többszörös illesztés





BLAST: Basic Local Alignment Search Tool

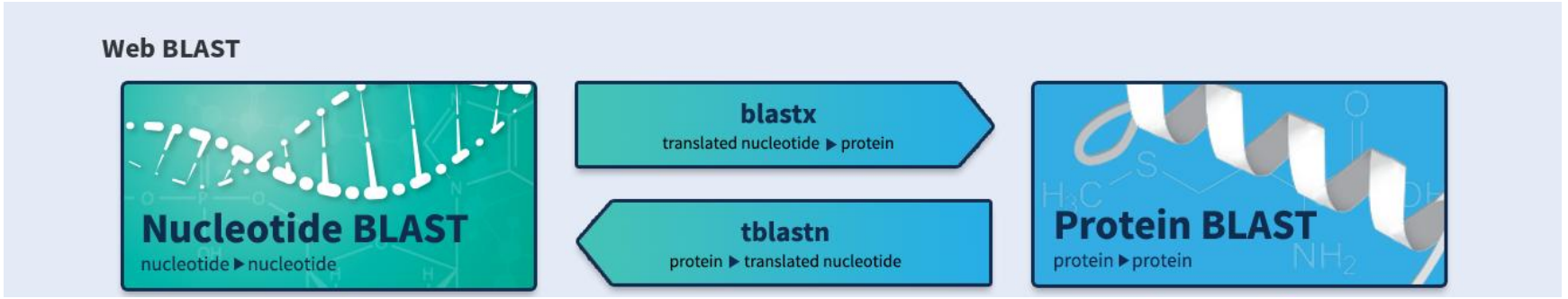
- A BLAST eljárás alkalmas arra, hogy egy adott szekvenciához **lokálisan hasonlóakat** azonosítson egy **nagyméretű adatbázis**ban.
- Hogy a hatalmas adatmennyiséget kezelni tudja, a BLAST **heurisztikus** egyszerűsítéssel él: adott hosszúságú (pl. 3 vagy 4) **szegmens**eket készít a kereső szekvenciából és ezek segítségével szűri az adatbázisbeli szekvenciákat ún. kezdeti illesztések (seed alignments) generálásával. A további lépésekben ezeket terjeszti ki, amíg a pontszám egy adott küszöb alá nem esik.
- A kimenetben minden találathoz az illesztéssel együtt megkapjuk:
    - Az E (**expectation**) értéket: adott adatbázison hány ilyen pontszámú találat várható
    - A P (**probability**) értéket: mekkora valószínűsége, hogy a kapott illesztés véletlenszerű

- A BLAST variánsai képesek **fehérje** és **nukleinsav**-adatbázisokban is keresni, sőt, keresztbe is (a szekvenciák “lefordításával”, ha kell)

- **PSI-BLAST**: position specific iterated BLAST: többkörös keresés, az első találatok alapján a fehérjecsaládra optimált, pozícióspecifikus pontozómátrixszal dolgozik → távoli evolúciós rokonságot is képes megtalálni. (Az aminosavak változásait **kontextusba helyezi**!)

BLAST: Basic Local Alignment Search Tool

https://blast.ncbi.nlm.nih.gov/



non-structural polyprotein 1a [Bat SARS-like coronavirus]

Sequence ID: ATO98204.1 **Length:** 4382 **Number of Matches:** 1

Range 1: 3241 to 3546 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
634 bits(1634)	0.0	Compositional matrix adjust.	294/306(96%)	302/306(98%)	0/306(0%)

```
Query  1     SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDLLIR  60
             SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDD VYCPRHVICT+EDMLNPNYEDLLIR
Sbjct  3241  SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDTVYCPRHVICTAEDMLNPNYEDLLIR  3300

Query  61    KSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVDTANPKTPKYKFVRIQPGQTFSVLACYNG  120
             KSNH+FLVQAGNVQLRVIGHSMQNC+L+LKVDT+NPKTPKYKFVRIQPGQTFSVLACYNG
Sbjct  3301  KSNHSFLVQAGNVQLRVIGHSMQNCLLRLKVDTSNPKTPKYKFVRIQPGQTFSVLACYNG  3360

Query  121   SPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEGN  180
             SPSGVYQCAMRPN TIKGSFLNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEG
Sbjct  3361  SPSGVYQCAMRPNHTIKGSFLNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEGK  3420

Query  181   FYGPFVDRQTAQAAGTDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYE  240
             FYGPFVDRQTAQAAGTDTTIT+NVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYE
Sbjct  3421  FYGPFVDRQTAQAAGTDTTITLNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYE  3480

Query  241   PLTQDHVDILGPLSAQTGIAVLDMCASLKELLQNGMNGRTILGSALLEDEFTPFDVVRQC  300
             PLTQDHVDILGPLSAQTGIAVLDMCA+LKELLQNGMNGRTILGS +LEDEFTPFDVVRQC
Sbjct  3481  PLTQDHVDILGPLSAQTGIAVLDMCAALKELLQNGMNGRTILGSTILEDEFTPFDVVRQC  3540

Query  301   SGVTFQ  306
             SGVTFQ
Sbjct  3541  SGVTFQ  3546
```