4. Többszörös szekvencia−összerendezés és filogenetikai analízis

Többszörös szekvencia−összerendezés

Célja

Lehetõvé tenni egy−egy géncsaládon vagy fehérjecsaládon belül az összefüggések felismerését és a családra jellemzõ, konzerválódott, biológiailag jelentõséggel bíró mintázatok feltárását. Növeli a jel/zaj viszonyt a páronkénti összerendezéshez képest.

A többszörös szekvencia−összerendezés definíciója

- Kétdimenziós táblázat, melynek sorai a szekvenciákat tartalmazzák, oszlopai a pozíciókat képviselik. Példa:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
I	Y	D	G	G	A	V	−	E	A	L
II	Y	D	G	G	I	L	V	E	A	L
III	F	E	G	G	I	L	V	E	A	L
IV	F	D	−	G	I	L	V	Q	A	V
V	Y	E	G	G	A	V	V	Q	A	L
Konsz.	**y**	**d**	**G**	**G**	**A/I**	**V/L**	**V**	**e**	**A**	**l**

- **Abszolút pozíció:** Egy aminosav/nukleotid sorszáma az adott szekvencián belül. Mindig változatlan.
- **Relatív pozíció:** Annak az oszlopnak a sorszáma, amelyben az adott aminosav/nukleotid az összerendezésen belül szerepel. Az összerendezés változtatásával változik.
- **Konszenzusszekvencia:** Rendszerint az összerendezés aljára írt pszeudoszekvencia, amely egy sorban, szimbólumok segítségével összegzi az egyes pozíciók variabilitását. A pszeudoszekvencia matematikai objektumokból (pl. helyettesítési mátrixok) is állhat.

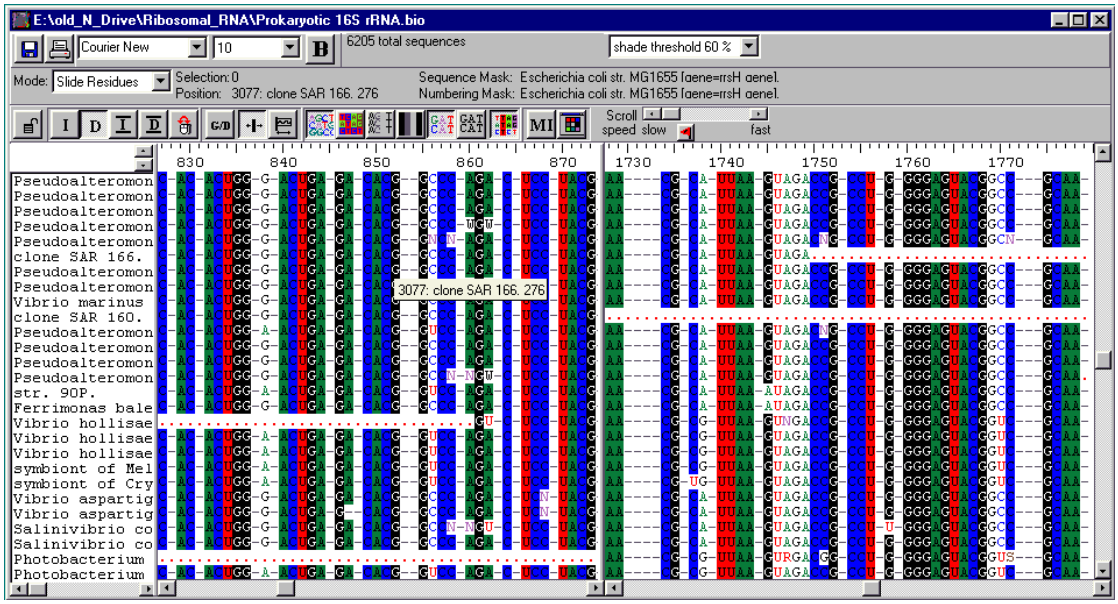
A többszörös szekvencia−összerendezés módszerei

Szimultán módszerek

A páronkénti szekvencia−összerendezõ algoritmusok (Needleman−Wunsch, Smith−Waterman, stb.) kiterjeszthetõk több szekvenciára is. Ez azonban többdimenziós mátrixokat igényel, a számítási idõ a szekvenciák számával exponenciálisan nõ. Ezért csak kisszámú, rövid szekvencia esetében alkalmazhatóak.

Manuális módszerek

- Összerendezés−szerkesztõ programok segítségével. Pl. BioEdit, GeneDoc (Windowshoz), SeaView (Unix/Linux, Windows), stb. Egy kép a BioEdit−rõl:



- A színezés nagy segítséget jelent a konzerválódott motívumok felismerésénél
- Bár felvethetõ, hogy a manuális módszer szubjektív, mindig szükség van az automatikus módszerrel nyert összerendezések utólagos manuális igazítására.

Progresszív módszerek

- Heurisztikus módszerek, melyek ésszerû idõ alatt próbálnak jó (ha nem is optimális) összerendezést találni
- Pl.:
  - ♦ Összes lehetséges szekvenciapár összerendezése
  - ♦ Mindegyik szekvencia összerendezése egy kiválasztott szekvenciával
  - ♦ A szekvenciák összerendezése tetszõleges sorrendben
  - ♦ A szekvenciák összerendezése egy filogenetikai törzsfa elágazásainak sorrendje szerint (ez a leggyakoribb)
- Elterjedt programok: Clustal, MultAlin

CLUSTAL

- Thompson, Higgins Gibson 1994
- Több verzió, legújabb a Clustal W, mely online is elérhetõ. Grafikus felhasználó felület hozzá: Clustal X
- A bemenõ szekvenciákon páronkénti összerendezést végez az összes lehetséges módon
- A páronkénti összerendezésekbõl távolságokat számít a szekvenciapárok között
- A távolságok alapján filogenetikai törzsfát készít, ez a **vezérfa**.
- A többszörös összerendezést a törzsfa elágazásainak sorrendje szerint végzi: elõbb a legközelebbi szekvenciákat rendezi össze, majd hozzárendezi az egyre távolabbiakat
- Legújabb verzió különlegességei:
  - ♦ Súlyozás: az összerendezés pontszámának kiszámításakor a közeli szekvenciákat kisebb súllyal veszi figyelembe (a közel azonos szekvenciák ne nyomják el a távolabbiakat)
  - ♦ Az aminosav−helyettesítési mátrixokat változtatja az éppen összerendezett szekvenciák távolsága szerint
  - ♦ Oldallánc− és pozícióspecifikus gap penalty−k

Összerendezés CLUSTAL−lal:

```
CLUSTAL W (1.60) multiple sequence alignment


hum−U1A         ------MAVPETRPNHTIYINNLNEKIKKDELKKSLYAIFSQFGQILDILVSRSLKMRGQ
mse−U1A         MATIATMPVPETRANHTIYINNLNEKIKKDELKKSLYAIFSQFGQILDILVSRIMKMRGQ
xla−U1A         ------MSIQEVRPNNTIYINNLNEKIKKDELKKSLYAIFSQFGQILDILVSRNLKMRGQ
dme−U1A         ---------MEMLPNQTIYINNLNEKIKKEELKKSLYAIFSQFGQILDIVALKTLKMRGQ
                          *   * ************.******************.  . .*****

hum−U1A         AFVIFKEVSSATNALRSMQGFPFYDKPMRIQYAKTDSDIIAKMKGTFVERDRKR−EKRKP
mse−U1A         AFVIFKEVTSATNALRSMQGFPFYDKPMRIQYAKTDSDIIAKMKGTYVERDRKR−EKRKP
xla−U1A         AFVIFKETSSATNALRSMQGFPFYDKPMRIQYSKTDSDIIAKMKGTFVERDRKRQEKRKV
dme−U1A         AFVIFKEIGSASNALRTMQGFPFYDKPMQIAYSKSDSDIVAKIKGTFKERPKKVKPPKPA
                *******  **.****.***********.* *.*.****.**.***. **       .

hum−U1A         KSQETPATKKAVQGGGATPVVGAVQGPVPGMPPMTQAPRIMHHMPGQPPYMPPPGMIPPP
mse−U1A         KSQETPAAKKAVQGGAAAPVVGAVQ−PVPGMPPMPQAPRIMHHMPGQPPYMPPPGMIPPP
xla−U1A         KVPEVQGVKNAMPGAALLPGVPGQMAAMQDMPGMTQAPRMMH−MAGQAPYMHHPGMMPPP
dme−U1A         PGTDEKKDKKKK−----------------------------------------------P
                   .    *                                                   *

hum−U1A         GLAPGQIPPGAMPPQQLMPGQMPPAQPLSENPPNHILFLTNLPEETNELMLSMLFNQFPG
mse−U1A         GLAPGQIPPGAMPPQQLMPGQMPPAQPLSENPPNHILFLTNLPEETNELMLSMLFNQFPG
xla−U1A         GMAPGQMPPGGMPHGQLMPGQMAPMQPISENPPNHILFLTNLPEETNELMLSMLFNQFPG
dme−U1A         SSAENSNP−----------------NAQTEQPPNQILFLTNLPEETNEMMLSMLFNQFPG
                  * .  *                 .  .*.***.************.***********

hum−U1A         FKEVRLVPGRHDIAFVEFDNEVQAGAARDALQGFKITQNNAMKISFAKK
mse−U1A         FKEVRLVPGRHDIAFVEFDNEVQAGAARDALQGFKITQNNAMKISFAKK
xla−U1A         FKEVRLVPGRHDIAFVEFDNEVQAGAARESLQGFKITQSNSMKISFAKK
dme−U1A         FKEVRLVPNRHDIAFVEFTTELQSNAAKEALQGFKITPTHAMKITFAKK
                ********.********* .*.*..**...*******  ..***.****
```

MultAlin

- Corpet 1988, online elérhetõ.
- Clustal−megközelítés problémája: a kezdeti, páronkénti, tökéletlen összerendezésekbõl származtatja a vezérfát. A valódi fát a végsõ összerendezésbõl lehetne megkapni.
- MultAlin: Rekurzív eljárás: a kapott többszörös összerendezésbõl újraszámolja a vezérfát, ennek alapján újabb összerendezést készít, ezt addig ismétli, míg már nem javul tovább a pontszám.
- Hátrány: kezdeti hibák a rekurzió során továbbadódnak

Összerendezés MultAlin−nal:

```
Multalin version 5.4.1
Copyright I.N.R.A. France 1989, 1991, 1994, 1996
Published research using this software should cite
Multiple sequence alignment with hierarchical clustering
F. CORPET, 1988, Nucl. Acids Res., 16 (22), 10881−10890
Symbol comparison table: blosum62
Gap weight: 12
Gap length weight: 2
Consensus levels: high=90% low=50%
Consensus symbols:
 ! is anyone of IV
 $ is anyone of LM
 % is anyone of FY
 # is anyone of NDQEBZ

 MSF:    289    Check:   0         ..
 Name: hum−U1A        Len:   289  Check: 9991  Weight:  0.67
 Name: mse−U1A        Len:   289  Check: 9174  Weight:  0.67
 Name: xla−U1A        Len:   289  Check: 7674  Weight:  0.67
 Name: dme−U1A        Len:   289  Check: 3145  Weight:  1.99
 Name: Consensus      Len:   289  Check:  255  Weight:  0.00

//
```

```
                1                                                    50
  hum-U1A       ......MAVP ETRPNHTIYI NNLNEKIKKD ELKKSLYAIF SQFGQILDIL
  mse-U1A       MATIATMPVP ETRANHTIYI NNLNEKIKKD ELKKSLYAIF SQFGQILDIL
  xla-U1A       ......MSIQ EVRPNNTIYI NNLNEKIKKD ELKKSLYAIF SQFGQILDIL
  dme-U1A       .........M EMLPNQTIYI NNLNEKIKKE ELKKSLYAIF SQFGQILDIV
Consensus       ......m... E.rpN.TIYI NNLNEKIKK# ELKKSLYAIF SQFGQILDIl

                51                                                   100
  hum-U1A       VSRSLKMRGQ AFVIFKEVSS ATNALRSMQG FPFYDKPMRI QYAKTDSDII
  mse-U1A       VSRIMKMRGQ AFVIFKEVTS ATNALRSMQG FPFYDKPMRI QYAKTDSDII
  xla-U1A       VSRNLKMRGQ AFVIFKETSS ATNALRSMQG FPFYDKPMRI QYSKTDSDII
  dme-U1A       ALKTLKMRGQ AFVIFKEIGS ASNALRTMQG FPFYDKPMQI AYSKSDSDIV
Consensus       vsr.$KMRGQ AFVIFKE..S AtNALRsMQG FPFYDKPMrI qYsKtDSDI!

                101                                                  150
  hum-U1A       AKMKGTFVER DRKR.EKRKP KSQETPATKK AVQGGGATPV VGAVQGPVPG
  mse-U1A       AKMKGTYVER DRKR.EKRKP KSQETPAAKK AVQGGAAAPV VGAVQ.PVPG
  xla-U1A       AKMKGTFVER DRKRQEKRKV KVPEVQGVKN AMPGAALLPG VPGQMAAMQD
  dme-U1A       AKIKGTFKER PKK..... ......... ......... .........
Consensus       AKmKGT%vER drKr.ekrk. k..e....k. a..g....p. v.........

                151                                                  200
  hum-U1A       MPPMTQAPRI MHHMPGQPPY MPPPGMIPPP GLAPGQIPPG AMPPQQLMPG
  mse-U1A       MPPMPQAPRI MHHMPGQPPY MPPPGMIPPP GLAPGQIPPG AMPPQQLMPG
  xla-U1A       MPGMTQAPRM MH.MAGQAPY MHHPGMMPPP GMAPGQMPPG GMPHGQLMPG
  dme-U1A       ......... ......... ......VKPP KPAPGTDEKK DKKKKPSSAE
Consensus       mp.m.qapr. mh.m.gq.py m..pgm.pPP g.APGq.ppg .mp..qlmpg

                201                                                  250
  hum-U1A       QMPPAQPLSE NPPNHILFLT NLPEETNELM LSMLFNQFPG FKEVRLVPGR
  mse-U1A       QMPPAQPLSE NPPNHILFLT NLPEETNELM LSMLFNQFPG FKEVRLVPGR
  xla-U1A       QMAPMQPISE NPPNHILFLT NLPEETNELM LSMLFNQFPG FKEVRLVPGR
  dme-U1A       NSNP.NAQTE QPPNQILFLT NLPEETNEMM LSMLFNQFPG FKEVRLVPNR
Consensus       #m.P.#p.sE #PPNhILFLT NLPEETNE$M LSMLFNQFPG FKEVRLVPgR

                251                                  289
  hum-U1A       HDIAFVEFDN EVQAGAARDA LQGFKITQNN AMKISFAKK
  mse-U1A       HDIAFVEFDN EVQAGAARDA LQGFKITQNN AMKISFAKK
  xla-U1A       HDIAFVEFDN EVQAGAARES LQGFKITQSN SMKISFAKK
  dme-U1A       HDIAFVEFTT ELQSNAAKEA LQGFKITPTH AMKITFAKK
Consensus       HDIAFVEFdn EvQagAAr#a LQGFKITq.n aMKIsFAKK
```

A dme−U1A szekvenciát (Drosophila fehérje) a Clustal és a MultAlin másképpen rendezi hozzá a többihez: Clustal két hézagot nyit, MultAlin csak egy hosszút. A MultAlinnál így több az azonosság, de a következtetés az, h. ez egy bizonytalan régió.

Melyik a legjobb módszer? Ezt nem lehet tudni, a helyes az, ha több módszert is alkalmazunk és az eredményekbõl konszenzust képezünk, manuálisan is igazítva

---

Többszörös összerendezések adatbázisai

A Weben többféle összerendezés−adatbázis is elérhetõ. Két példa:

Pfam

- Automatikusan származtatott összerendezéseket tartalmaz. Pl.:

```
A1AA_HUMAN  SLKYP....AI..MTER.KAA..AILALL.WVV.AL.VVSVGP.LLG...WKEPV..PPDE....RF
A1AA_RAT    SLKYP....AI..MTER.KAA..AILALL.WAV.AL.VVSVGP.LLG...WKEPV..PPDE....RF
A1AB_CANFA  SLQYP....TL..VTRR.KAI..LALLGV.WVL.ST.VISIGP.LLG...WKEPA..PNDD....KE
A1AB_HUMAN  SLQYP....TL..VTRR.KAI..LALLSV.WVL.ST.VISIGP.LLG...WKEPA..PNDD....KE
A1AB_MESAU  SLQYP....TL..VTRR.KAI..LALLSV.WVL.ST.VISIGP.LLG...WKEPA..PNDD....KE
A1AB_RAT    SLQYP....TL..VTRR.KAI..LALLSV.WVL.ST.VISIGP.LLG...WKEPA..PNDD....KE
A1AC_BOVIN  PLRYP....TI..VTQK.RGL..MALLCV.WAL.SL.VISIGP.LFG...WRQPA..PEDE.....T
A1AC_HUMAN  PLRYP....TI..VTQR.RGL..MALLCV.WAL.SL.VISIGP.LFG...WRQPA..PEDE.....T
A1AC_RAT    PLRYP....TI..VTQR.RGV..RALLCV.WVL.SL.VISIGP.LFG...WRQPA..PEDE.....T
OAR_DROME   PINYA....QK..RTVG.RVL..LLISGV.WLL.SL.LISSPP.LIG...W.NDW..PDEFT..SAT
D1DR_CARAU  PFRYE....RK..MTPR.VAF..VMISGA.WTL.SV.LISFIPVQLK...WHKAQ..PIGFL..EVN
D1DR_FUGRU  PFRYE....RK..MTPK.VAC..LMISVA.WTL.SV.LISFIPVQLN...WHKAQ..TASYVELNGT
DADR_DIDMA  PFRYE....RK..MTPK.AAF..ILISVA.WTL.SV.LISFIPVQLN...WHKARPLSSPDG..NVS
...
```

```
A1AA_HUMAN  ..............CGI..TE.....................EAG...YA.....VF......SS
A1AA_RAT    ..............CGI..TE.....................EVG...YA.....IF......SS
A1AB_CANFA  ..............CGV..TE.....................EPF...YA.....LF......SS
A1AB_HUMAN  ..............CGV..TE.....................EPF...YA.....LF......SS
A1AB_MESAU  ..............CGV..TE.....................EPF...YA.....LF......SS
A1AB_RAT    ..............CGV..TE.....................EPF...YA.....LF......SS
A1AC_BOVIN  ..............ICQI..NE.....................EPG...YV.....LF......SA
A1AC_HUMAN  ..............ICQI..NE.....................EPG...YV.....LF......SA
A1AC_RAT    ..............ICQI..NE.....................EPG...YV.....LF......SA
OAR_DROME   ..............PCEL..TS.....................QRG...YV.....IY......SS
D1DR_CARAU  ..............ASRR...DLPTDNC........DSSL....NRT...YA.....IS......SS
D1DR_FUGRU  ..............YAGD..LPPDNCD..........SSL....NRT...YA.....IS......SS
DADR_DIDMA  ..............SQDE...TMDNCD..........SSL....SRT...YA.....IS......SS
...

A1AA_HUMAN  V.CSFY...LPMAVIV.VMY.CRV.YVV....ARS..TTRSLEA.GVKR
A1AA_RAT    V.CSFY...LPMAVIV.VMY.CRV.YVV....ARS..TTRSL.....EA
A1AB_CANFA  L.GSFY...IPLAVIL.VMY.CRV.YIV....AKR..TTKNL.....EA
A1AB_HUMAN  L.GSFY...IPLAVIL.VMY.CRV.YIV....AKR..TTKNL.....EA
A1AB_MESAU  L.GSFY...IPLAVIL.VMY.CRV.YIV....AKR..TTKNL.....EA
A1AB_RAT    L.GSFY...IPLAVIL.VMY.CRV.YIV....AKR..TTKNL.....EA
A1AC_BOVIN  L.GSFY...VPLTIIL.VMY.CRV.YVV....AKR..ESRGLKS.GLKT
A1AC_HUMAN  L.GSFY...LPLAIIL.VMY.CRV.YVV....AKR..ESRGLKS.GLKT
A1AC_RAT    L.GSFY...VPLAIIL.VMY.CRV.YVV....AKR..ESRGLKS.GLKT
OAR_DROME   L.GSFF...IPLAIMT.IVY.IEI.FVA....TRR..RLRERA..RANK
D1DR_CARAU  L.ISFY...IPVAIMI.VTY.TQI.YRI....AQK..QIRRIS..ALER
D1DR_FUGRU  L.ISFY...IPVAIMI.VTY.TRI.YRI....AQK..QIRRIS..ALER
DADR_DIDMA  L.ISFY...IPVAIMI.VTY.TRI.YRI....AQK..QIRRIS..ALER
```

597 darab G−fehérjéhez kapcsolt receptor szekvenciájának összerendezése (részlet!) a Pfam adatbázisból

- Sok szekvencia, erõs divergenciákkal, ezért sok az inzerció/deléció. Az automatikus módszer ezeket nehezen kezeli, ezért túlságosan sok hézagot szúr be −−> sok az elszigetelt, "elárvult" aminosavrész. Ez szétforgácsolja és elnyújtja az összerendezést, biológiai jelentése nincs.

PRINTS

- Manuálisan származtatott összerendezéseket tartalmaz. Pl.:

```
OAR_DROME   PINYAQ...KRTVGRVLLLISGVWLLSLLISSP.PLIG.WND.....WPDEFTSATP...
D2DR_RAT    PMLYNTR..YSSKRRVTVMIAIVWVLSFTISCP.LLFG.LNN...T.DQNE.........
D3DR_RAT    PVHYQHGTGQSSCRRVALMITAVWVLAFAVSCP.LLFG.FNT...TGDPSI.........
DADR_RAT    PFQYER...KMTPKAAFILISVAWTLSVLISFI.PVQLSWHKAK.PTWPLDGNFTSLEDT
DBDR_RAT    PFRYER...KMTQRVALVMVGLAWTLSILISFI.PVQLNWHRDKAGSQGQEGLLSNGTPW
A1AB_RAT    SLQYPT...LVTRRKAILALLSVWVLSTVISIG.PLLG.WKE......PAPNDDKE....
B1AR_RAT    PFRYQS...LLTRARARALVCTVWAISALVSFL.PILMHWW.....RAESD.EARRCYND
B2AR_HUMAN  PFKYQS...LLTKNKARVIILMVWIVSGLTSFL.PIQMHWY.....RATHQ.EAINCYAN
B2AR_RAT    PFKYQS...LLTKNKARVVILMVWIVSGLTSFL.PIQMHWY.....RATHK.QAIDCYAK
B3AR_RAT    PLRYGT...LVTKRRARAAVVLVWIVSATVSFA.PIMSQWW.....RVGADAEAQECHSN
5HTA_RAT    PIDYVN...KRTPRRAAALISLTWLIGFLISIP.PMLG.WRTPEDRSDPDA.........
5HTD_RAT    ALEYSK...RRTAGHAAAMIAAVWAISICISIP.PLF..WRQ..ATAHEEMSD.......
5HT2_RAT    PIHHSR...FNSRTKAFLKIIAVWTISVGISMPIPVFG.LQDDSKVFKEGS.........
...

OAR_DROME   ............CELTSQRG.................YVIYSSLGSPFIPLAIMTIVYIEIFVATRRRLRERARANK
D2DR_RAT    ............CIIANPA..................FVVYSSIVSFYVPFIVTLLVYIKIYIVLRKRRKR......
D3DR_RAT    ............CSISNPD..................FVIYSSVVSFYVPFGVTVLVYARIYIVLRQRQRK......
DADR_RAT    ED........DNCDTRLSRT.................YAISSSLISFYIPVAIMIVTYTSIYRIAQKQIRR......
DBDR_RAT    EEGWELEGRTENCDSSLNRT.................YAISSSLISFYIPVAIMIVTYTRIYRIAQVQIRR......
A1AB_RAT    ............CGVTEEPF..................CALFCSLGSFYIPLAVILVMYCRVYIVAKRTTKN......
B1AR_RAT    PK..........CCDFVTNRA.................YAIASSVVSFYVPLCIMAFVYLRVFREAQKQVKK......
B2AR_HUMAN  ET..........CCDFFTNQA.................YAIASSIVSFYVPLVIMVFVYSRVFQEAKRQLQK......
B2AR_RAT    ET..........CCDFFTNQA.................YAIASSIVSFYVPLVVMVFVYSRVFQVAKRQLQK......
B3AR_RAT    PR..........CCSFASNMP.................YALLSSSVSFYLPLLVMLFVYARVFVVAKRQRRF......
5HTA_RAT    ............CTISKDHG.................YTIYSTFGAFYIPLLLMLVLYGRIFRAARFRIRK......
5HTD_RAT    ............CLVNTSQIS.................YTIYSTCGAFYIPSILLIILYGRIYVAARSRILN......
5HT2_RAT    ............CLLADDN..................FVLIGSFVAFFIPLTIMVITYFLTIKSLQKEATL......
...
```

687 G−fehérjéhez kapcsolt receptor szekvenciájának összerendezése (részlet!) a PRINTS adatbázisból

- Itt a hézagok beszúrásával óvatosak voltak, így felismerhetõ két konzerválódott domén. Ez is mutatja, mennyire fontos a manuális szerkesztés.

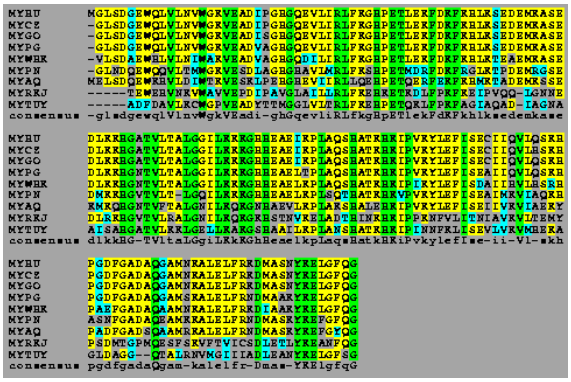
PSI−BLAST

- Position Specific Iterated Basic Local Alignment Search Tool: a BLAST összerendezés−keresõ program kiterjesztése:
  - ♦ Elõbb egy hagyományos BLAST kereséssel kigyûjti egy szekvencia homológjait, ezekbõl többszörös összerendezést készít
  - ♦ A többszörös összerendezésbõl elkészít egy szekvenciaprofilt
  - ♦ A szekvenciaprofillal újból keresést végez a szekvenciaadatbázison
  - ♦ Ezáltal távoli homológokat is megtalál
- Hátrány: ha bemászott egy nem homológ szekvencia (fals pozitív találat) a profilba, akkor a keresés további nem homológ szekvenciákat talál

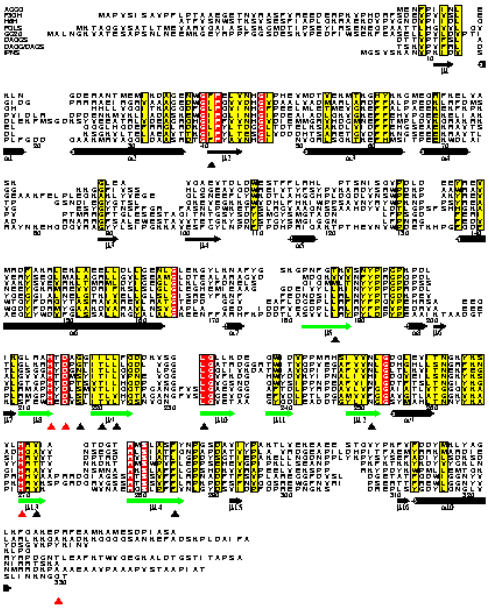
---

Prezentációs programok

- Olvashatóbbá, esztétikusabbá teszik a többszörös összerendezést, sokszor segítenek fontos összefüggések felismerésében, számolnak konszenzust, stb. Némelyik webszerveren keresztül is elérhetõ.
- BoxShade (Unix, Macintosh)



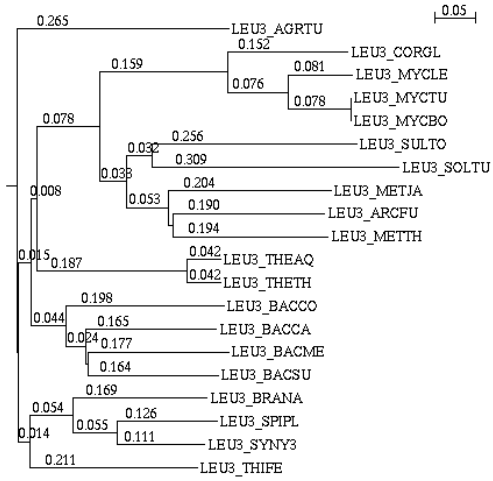
- ALSCRIPT (Unix, PC)



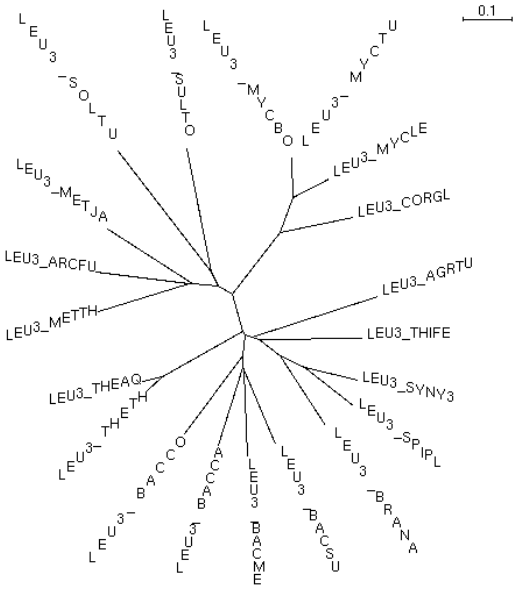
Filogenetikai analízis

Általános megjegyzések

- A filogenetikai analízis célja az evolúciós viszonyok, az evolúciós történet feltárása. Eredménye általában egy **filogenetikai fa** (törzsfa). A fák két fõ fajtája:



Gyökeres fa



Gyökértelen fa

- Az inkorrekt eredmények generálásának nagy a veszélye, mert a múltbeli eseményekrõl nincs közvetlen információnk, csak következtetni tudunk rájuk
- Ezért mindig fenntartásokkal kell kezelni az analízis eredményét. Nem elég egyszerûen lefuttatni egy filogenetikai programot, és kész.
- A filogenetikai törzsfákat építõ módszerek egy bizonyos evolúciós modell érvényességét feltételezik.
- Legfontosabb feltételezés, h. a törzsfejlõdés fával ábrázolható. Ez nem mindig igaz, pl. hibrid fajok, ill. (mikro)organizmusok között elõforduló DNS−átvitel (laterális géntranszfer) esetén.
- További feltételezés: a szekvenciák mind homológok
- Inkább DNS−szekvenciák alapján végzik, fehérjeszekvencia alapján ritkábban, nem is eléggé kidolgozott.

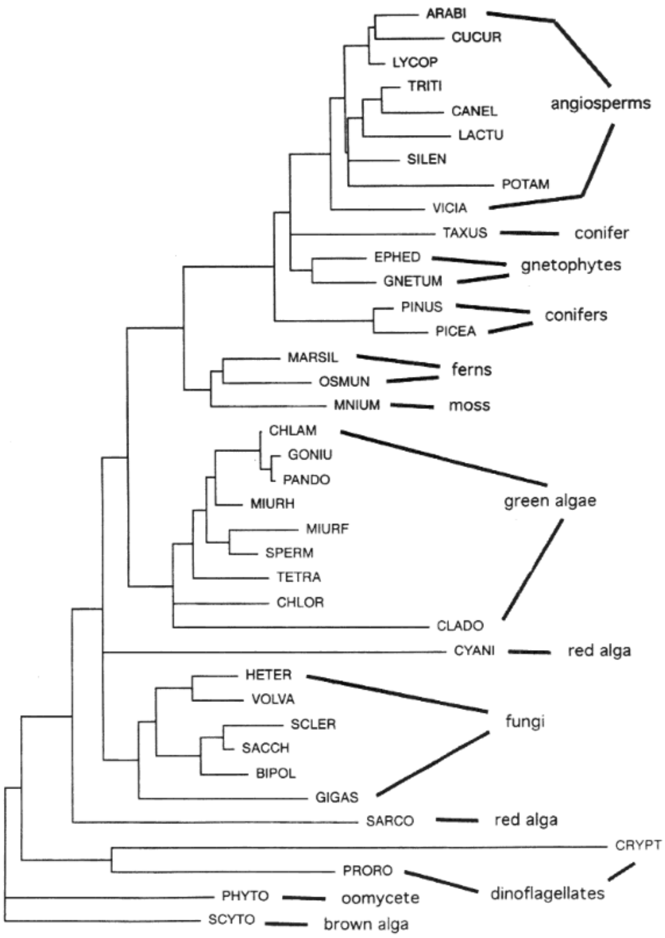
A filogenetikai analízis lépései

Négy lépés:

1. Összerendezés
2. A helyettesítési (tkp. evolúciós) modell meghatározása
3. Faépítés
4. A fa kiértékelése

1. Összerendezés

- Az ismertetett módszerekkel
- Clustal, Multalin, stb. kezdetleges filogenetikai analízist végez a vezérfa felépítésekor, de ezek nem megbízhatóak!



Clustal vezérfa az 5.8S rDNS szekvenciák alapján. Furcsaságok: vörösmoszatok több törzsben szétszórva szerepelnek, tûlevelûek (conifers) is, a páfrányok (ferns) és mohák viszont összekerültek.

- A vezérfát és a biológiai háttértudást felhasználva az összerendezést alkalmassá kell tenni az alaposabb filogenetikai analízisre --> **"összerendezés−sebészet"**. Pl.:

	116 |	122−144	155 |
1 ARABI	GCGCCC	---CAAGCCTTCT-GGCCG----	AGGGCACGTCT
2 LYCOP	C	---GAAGCCATTT-GGCCG----	A..........
3 triti	C	---GAGGCCACTC-GGCCG----	A.......C..
4 LACTU	C	---GAAGCCATCC-GGCTG----	A.......C..
5 SILEN	C	---GAAGC--TTC-GGCTG----	A..........
6 vicia	C	---GATGCCATTA-GGTTG----	A..........
7 CANEL	C	---GAGGCCACTA-GGCTG----	A....T..C..
8 potam	C	---TAAGCTTCCG-GGCCG----	A.....A.C..
9 ephed	C	---GAAGCC--TC-CGCCA----	A..........
10 gnetu	.-...C	TCCG-AGCC--TA-GGCCG----	A..........
11 PINUS	C	---GAGGCC--TC-GGTCG----	A..........

12 PICEA	C	---GAGNCC--TC-GGTCG----	A.........
13 TAXUS	.GC..G	---GAG-C---TC-GGCCG----	A.........
14 marsi	.-...C	---GAGGC---TC-GTCCG----	A.........
15 osmun	.-...C	---GCGGC---TC-GTCCA----	A.....T.C..
16 mnium	C	---GAGGC---TC-GTCCG----	A.....TT..C
17 CHLAM	....TC	---GAGGC--TTC-GGCCA----	A.A...T....
18 SPERM	....TC	---GAGGC--TTC-GGCCG----	A.A...T..T.
19 TETRA	....TC	---GAGGC--CTC-GGCCA----	A.A.....C..
20 CHLOR	....TC	---GAGAC--CTC-GGTCA----	A.A...T....
21 CLADO	....TC	---AAGTC--TAC-GGACT----	T.A...T....
22 HETER	C	TTT--GGT-ATT----CCGA---	A..-....C..
23 VOLVA	....TC	TTT--GGCCATT----CCGA---	A.A...T.C..
24 SCLER	C	CTT--GGT-ATT----CCGG---	G.....T.C..
25 sacch	C	CTT--GGT-ATT----CCAG---	G.....T.C..
26 BIPOL	C	TTT--GGT-ATT----CCAA---	A.....T.C..
27 GLOMU	....TC	CCT--GGT-ATT----CCGG---	G.A.T.T.C..
28 CYANI	....TT	TC--AGGAGAATTTTATTTTCCT	G.A........
29 SARCO	....TC	GC---GGTAA-TC--------CT	GCA.--T....
30 PHYTO	..A.TT	CCG--GGTTAGTC---CTG----	G.A.T.T.C..
31 SCYTO	...-TT	CCG--GGTTAGTC---CTG----	G.A...T.CT.
32 crypt	.---CT	CC---AGC--TGA---CT-----	------T...A
33 PRORO	....TT	TCG--GGATATCC---CTG----	AA....T.C..

Kiinduló összerendezés, a szekvenciák az organizmusok rendszertani hovatartozása szerinti sorrendben vannak. A középsõ részben lévõ szekvenciák három csoportra különülnek el (zöld növények, gombák, egysejtûek), s kevéssé hasonlítanak, összerendezésük önkényes. (A pontok itt az oszlop tetején jelölttel azonos bázist jelentenek.)

Ebbõl állítjuk elõ "sebészkedés" révén a filogenetikai analízisre elõkészített összerendezést:

	116	122-144	155	[122'-141']
1 ARABI	GCGCCC	???CAAGCCTTCT?GGCCG????	AGGGCACGTCT	??????????????????
2 LYCOP	GCGCCC	???GAAGCCATTT?GGCCG????	AGGGCACGTCT	??????????????????
3 triti	GCGCCC	???GAGGCCACTC?GGCCG????	AGGGCACGCCT	??????????????????
4 LACTU	GCGCCC	???GAAGCCATCC?GGCTG????	AGGGCACGCCT	??????????????????
5 SILEN	GCGCCC	???GAAGC?-TTC?GGCTG????	AGGGCACGTCT	??????????????????
6 vicia	GCGCCC	???GATGCCATTA?GGTTG????	AGGGCACGTCT	??????????????????
7 CANEL	GCGCCC	???GAGGCCACTA?GGCTG????	AGGGCTCGCCT	??????????????????
8 potam	GCGCCC	???TAAGCTTCCG?GGCCG????	AGGGCAAGCCT	??????????????????
9 ephed	G-GCCC	???GAAGCC?-TC?CGCCA????	AGGGCACGTCT	??????????????????
10 gnetu	G-GCCC	TCCG?AGCC?-TA?GGCCG????	AGGGCACGTCT	??????????????????
11 PINUS	GCGCCC	???GAGGCC?-TC?GGTCG????	AGGGCACGTCT	??????????????????
12 PICEA	GCGCCC	???GAG?CC?-TC?GGTCG????	AGGGCACGTCT	??????????????????
13 TAXUS	GGCCCG	???GAG-C?-?TC?GGCCG????	AGGGCACGTCT	??????????????????
14 marsi	G-GCCC	???GAGGC??-TC?GTCCG????	AGGGCACGTCT	??????????????????
15 osmun	G-GCCC	???GCGGC??-TC?GTCCA????	AGGGCATGCCT	??????????????????
16 mnium	GCGCCC	???GAGGC??-TC?GTCCG????	AGGGCATTTCC	??????????????????
17 CHLAM	GCGCTC	???GAGGC?-TTC?GGCCA????	AGAGCATGTCT	??????????????????
18 SPERM	GCGCTC	???GAGGC?-TTC?GGCCG????	AGAGCATGTTT	??????????????????
19 TETRA	GCGCTC	???GAGGC?-CTC?GGCCA????	AGAGCACGCCT	??????????????????
20 CHLOR	GCGCTC	???GAGAC?-CTC?GGTCA????	AGAGCATGTCT	??????????????????
21 CLADO	GCGCTC	???AAGTC?-TAC?GGACT????	TGAGCATGTCT	??????????????????
22 HETER	GCGCCC	??????????????????????	AGG-CACGCCT	TTT??GGT-ATT????CCGA
23 VOLVA	GCGCTC	??????????????????????	AGAGCATGCCT	TTT??GGCCATT????CCGA
24 SCLER	GCGCCC	??????????????????????	GGGGCATGCCT	CTT??GGT-ATT????CCGG
25 sacch	GCGCCC	??????????????????????	GGGGCATGCCT	CTT??GGT-ATT????CCAG
26 BIPOL	GCGCCC	??????????????????????	AGGGCATGCCT	TTT??GGT-ATT????CCAA
27 GLOMU	GCGCTC	??????????????????????	GGAGTATGCCT	CCT??GGT-ATT????CCGG
28 CYANI	GCGCTT	??????????????????????	GGAGCACGTCT	??????????????????
29 SARCO	GCGCTC	??????????????????????	GCAG-?TGTCT	??????????????????
30 PHYTO	GCACCT	??????????????????????	GGAGTATGCCT	??????????????????
31 SCYTO	GCG-TT	??????????????????????	GGAGCATGCTT	??????????????????
32 crypt	G??-CT	??????????????????????	??????TGTCA	??????????????????
33 PRORO	GCGCTT	??????????????????????	AAGGCATGCCT	??????????????????

A filogenetikai elemzéshez elõkészített összerendezés.

Ez az összerendezés a következõ módosítások által keletkezett a kiinduló összerendezésbõl:

1. A gombáknál és az egysejtûeknél a középsõ szekvenciadarabot kivágtuk és ?-ekkel (jelentése: hiányzó nukleotid) helyettesítettük, mert a zöld növényekhez való rendezésük önkényes volt.
2. A gombák kivágott szakaszát a szekvencia végére illesztettük, mert egymás közti variációik fontosak. (A szekvencia épsége a filogenetikai elemzésnél nem szükséges.) Az egysejtûek kivágott szakaszát elhagytuk, nem hordoz filogenetikailag értékes információt.

3. Az egy nukleotidnál hosszabb hézagokat egy hézagra és megfelelõ számú ?−re cseréltük (a hosszabb hézagot nem helyes egymástól független deléciók egymásutánjának tekinteni).

A helyettesítési (evolúciós) modell meghatározása

Három eleme (paramétere) van:

- Bázisgyakoriságok
- A bázisok egymás közti cseréjének gyakorisága
- A szekvencián belüli pozíciók mutációgyakoriságának heterogenitása

Ezek meghatározásának két módja:

- *Empirikus* módszer: korábbi elemzésekbõl meghatározott értékeket használunk fel, mint fix értékeket. Elõny: könnyû számíthatóság. Hátrány: az adott adathalmazra nem biztos, hogy jók a paraméterek.
- *Paraméteres* módszer: magából a vizsgált adathalmazból vezetjük le a paramétereket. Elõny: pontosabb lehet. Hátrány: félrevezethet, ha az adatkészlet nem megfelelõ.

A bázisok egymás közti cseréjének gyakorisága

- Lehet elõre rögzített mátrix pl.:

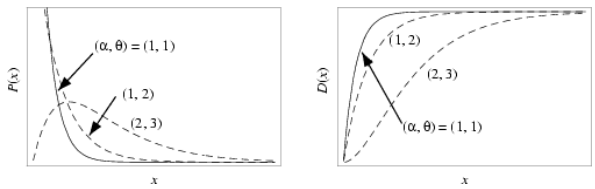
```
  A C G T
A − 2 1 2
C 2 − 2 1
G 1 2 − 2
T 2 1 2 −
```

  A pontszámok a csere költségét mutatják, purinbázis pirimidinbázisra való cseréje (transzverzió) ritkább, ezért költségesebb, mint a purin−purin és a pirimidin−pirimidin csere (tranzíció).

- Dolgozhatunk az összerendezésbõl számított helyettesítési mátrixokkal is
- Idõreverzibilis mátrixok: az oda− és visszacsere pontszáma azonos, akkor teljesül, ha nincs meghatározott idõbeli eltolódás a bázisfrekvenciákban (stacionárius modell). Létezik egy korrekciós módszer nemstacionárius esetre.

A szekvencián belüli pozíciók mutációgyakoriságának heterogenitása

- A szekvencián belül a mutációk gyakorisága erõsen variábilis. Pl. fehérjét kódoló szakasznál a kodonok harmadik bázisai sokkal variábilisabbak, mint az elsõ kettõ. A fehérjék konzerválódott régióit kódoló részek is kevésbé variábilisek.
- Modellek a mutációgyakoriság heterogenitásának leírására:
    1. Nemparaméteres módszer: az egyes pozíciókat (pl. szekvenciaszakaszokat) kategóriákba sorolja a megfigyelt mutációs gyakoriság alapján
    2. Invariánsok módszere: A pozíciók egy bizonyos hányadát invariánsnak tekinti, a többit azonos valószínûséggel változónak
    3. Gamma eloszlás módszere (legkorszerûbb): feltételezi, h. a mutációs gyakoriságok eloszlása a gamma valószínûségeloszlás szerinti, ennek az alakját egy paraméter jellemzi, melyet meg kell becsülni. Lehet folytonos vagy diszkrét.



Gamma eloszlás. Bal oldalon a valószínûségeloszlás sûrûségfüggvénye, jobb oldalon az eloszlásfüggvény, mindkettõ a két paraméter (alfa, théta) több lehetséges értéke mellett.

Melyik helyettesítési modellt válasszuk? A kevés paraméterrel dolgozó modellek jobban alkalmazhatóbbak, megbízhatóbbak, a túlságosan leegyszerûsített modellek viszont hibás eredményt adhatnak. Fontos a tranzíciós és a transzverzió megkülönböztetése és a mutációgyakoriság heterogenitásának figyelembe vétele. Gondosan kell kiválasztani az adott adatokhoz legjobban illeszkedõ evolúciós modellt.

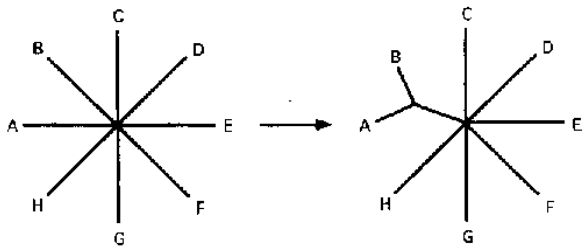
A fa felépítése

A faépítõ módszerek kétféle osztályozása:

1. ♦ **Algoritmus alapú**: egy algoritmus levezet egy bizonyos fát
   ♦ **Kritérium alapú**: az összes lehetséges fát generálja, ezeket értékeli valamilyen optimalizációs kritérium alapján.
2. ♦ **Távolság alapú:** Páronkénti távolságokat számít a szekvenciák között, majd ezekkel a távolságokkal dolgozik tovább, fákat levezetve belõlük. A távolságszámításnál mindig információvesztés van.
   ♦ **Karakter alapú:** (Karakter = pozíció az összerendezésben.) Olyan fákat származtat le, amelyek mindegyik pozícióra optimalizálják az adatmintázatok eloszlását.

Távolság alapú módszerek

- A szekvenciák közötti távolság egy határértéket ér el, ahogy a távolság nõ. Ha egy pozícióban már történt mutáció, a további mutációk már nem teszik távolibbá. A távolság alapú módszerek korrigálnak erre az effektusra.
- **Neighbor Joining (NJ)**: Egy csillag alakú fából kiindulva a legközelebbi szomszédokat összekapcsolja, helyettesíti õket az átlagukkal, majd ezt ismételgeti a teljes fa kialakulásáig.



- **Minimum Evolution (ME)**: A legrövidebb olyan fát találja meg, amely összeegyeztethetõ a szekvenciák közötti távolságokkal. (A faágak hossza evolúciós távolságnak felel meg, így két szekvencia távolsága a fa szerint számítható a megfelelõ ágak hosszának összeadásával.) Hasonló elven mûködik a Fitch−Margoliash (FM) módszer.

Karakter alapú módszerek

- **Maximum Parsimony (MP)**: "legnagyobb takarékosság" módszere: Olyan fát épít, ami a lehetõ legkevesebb mutációs eseménnyel magyarázza meg a meglévõ szekvenciák létrejöttét. Számos azonos pontszámú fát szolgáltat, ezek közös részét vehetjük mint megbízhatót. Nagy távolságú szekvenciák esetében hátránya, h. azonos bázis esetén azt tételezi fel, h. nem történt mutáció, holott valószínûbb a visszacserélõdés.
- **Maximum Likelihood (ML)**: "legnagyobb valószínûség" módszere: Komplikált módszer. Minden pozícióra kiszámítja, hogy adott fa és helyettesítési modell mellett mi a valószínûsége annak, hogy a megfigyelt variációs mintázat jöjjön létre az adott pozícióban. Az egyes pozíciókra kapott valószínûségek összeszorzásával adódik a teljes fa valószínûsége. Ezt sok fára a legjobbat kiválasztja. Ezt többféle helyettesítési modell mellett is elvégezhetjük, ezek közül is kiválasztva a legjobbat. Igen számításigényes, de ez a legmegbízhatóbb.

A fák kiértékelése

Kétféle módszer:

- Randomizált adatokra kapott eredményekkel való összehasonlítás
- A kapott fa alátámasztottságának tesztelése ún. "resampling" statisztikai módszerekkel (bootstrapping, jackknife). Lényegük: a meglévõ adatokból véletlenszerûen mintákat veszünk, ezekre végezzük el a számítást, majd statisztikát készítünk. (Nem részletezzük.)

Filogenetikai szoftver

- **PHYLIP** (Philogeny Inference Package): Ingyenes csomag, kb. 30 programmal, melyek a filogenetikai analízis mindenféle részkérdésére megoldást kínálnak. Parancssorral vezérelhetõ (nem grafikus).
- **PAUP** (Phylogenetic Analysis Using Parsimony): Megvásárolható program (kb. $100). Grafikus felhasználói felületû, menüvezérelt (de csak Macintoshon, más rendszereken parancssorvezérelt). A legkülönfélébb eljárások széles választékát kínálja.