Bioinformatika és genomanalízis az orvostudományban

Többszörös szekvencia illesztés

Cserző Miklós

2020

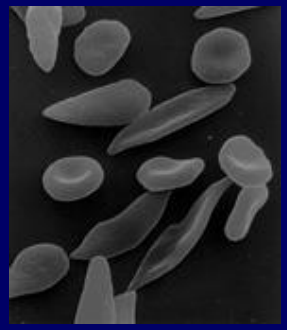
https://semmelweis.zoom.us/j/96102872458?pwd=Rk1PL2tqS21sdlUwc3B4eDFCZkNKQT09

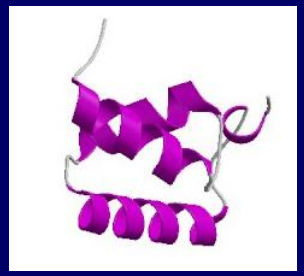
A mai előadás

- A többszörös illesztés biológiai jelentősége
- A probléma komplexitása
- Progresszív módszer
- Iteratív módszer
- Globális módszerek
  - Genetikus algoritmus
  - „szimulált dermedés”
- Kapcsolódó adatbázisok

Aminósavak helyettesítése

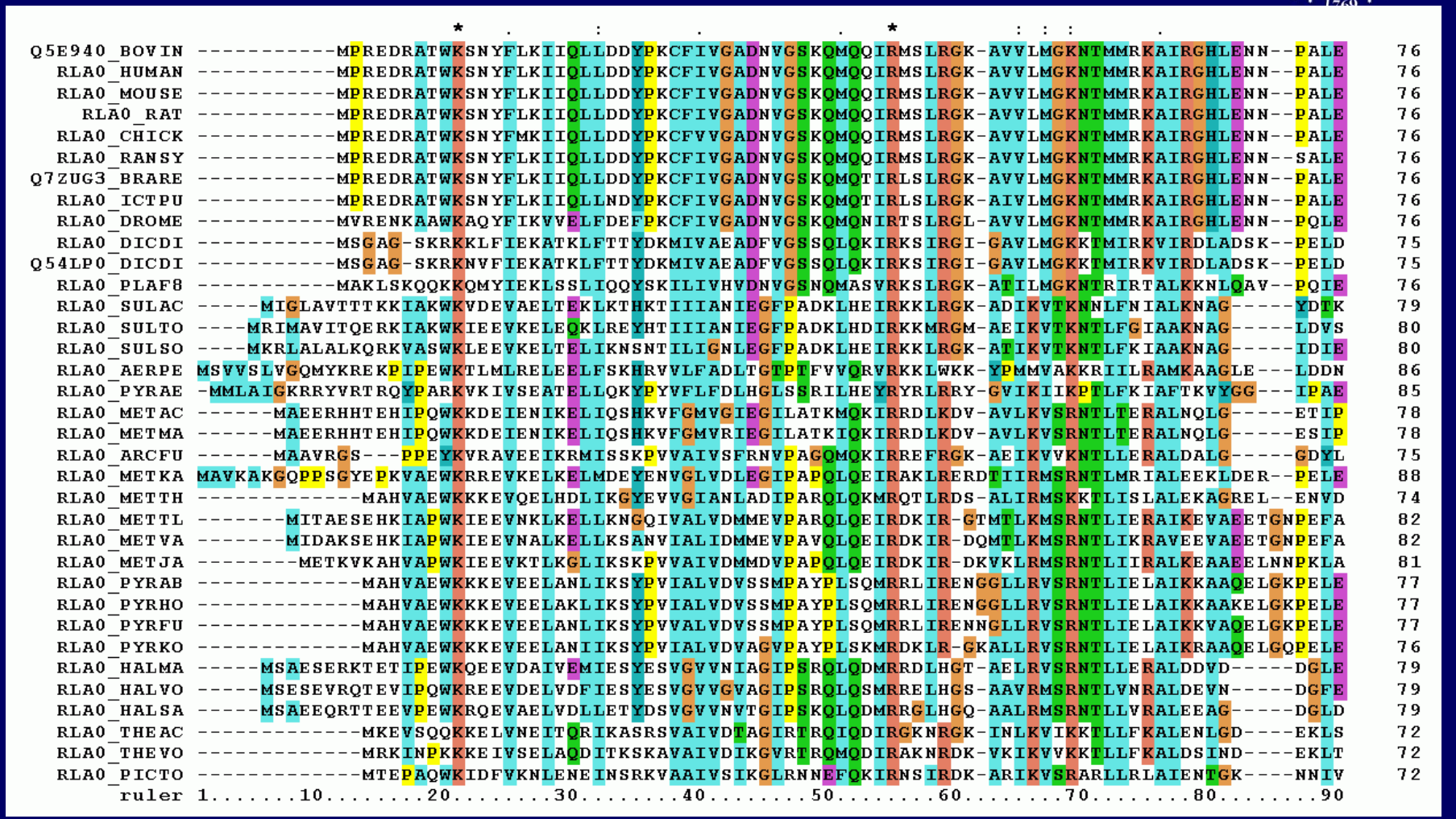
- Egy elrontott aminósav csere tönkre teszi a fehérjét
- Akár szekvenciálisan 85%-ban eltérő fehérjék szerkezete is lehet azonos
- Bizonyos aminósavak a szerkezet bizonyos pontjain bizonyos mértékben helyettesíthetik egymást
- Más pontokon más szabályok érvényesek





A többszörös illesztés jelentősége

- Az illesztés során láthatóvá válnak a szekvenciák konzervált és nem konzervált részei
- A szerkezet és funkció szempontjából mindkét rész fontos, csak másképp
- „Two homologous sequences whisper ... a full multiple alignment shouts out loud.” A. Lesk


Q5E940_BOVIN
RLA0_HUMAN
RLA0_MOUSE
RLA0_RAT
RLA0_CHICK
RLA0_RANSY
Q7ZUG3_BRARE
RLA0_ICTPU
RLA0_DROME
RLA0_DICDI
Q54LP0_DICDI
RLA0_PLAF8
RLA0_SULAC
RLA0_SULTO
RLA0_SULSO
RLA0_AERPE
RLA0_PYRAE
RLA0_METAC
RLA0_METMA
RLA0_ARCFU
RLA0_METKA
RLA0_METTH
RLA0_METTL
RLA0_METVA
RLA0_METJA
RLA0_PYRAB
RLA0_PYRHO
RLA0_PYRFU
RLA0_PYRKO
RLA0_HALMA
RLA0_HALVO
RLA0_HALSA
RLA0_THEAC
RLA0_THEVO
RLA0_PICTO
ruler
1.....10.....20.....30.....40.....50.....60.....70.....80.....90

A probléma komplexitása

- 3 szekvencia esetén illesztőfelület helyett illesztőtérfogat (1000^3 fehérjékre)
- *N* szekvencia esetén N-dimenziós absztakt tér Len^{Nseq} – ezen még lehet segíteni
- Dinamikus programozás:
  - Nem kell a teljes illesztőfelület – elég egy sor
  - Memóriatakarékos
  - Viszont többször kell átszámolni az illesztőfelületet – futási idő hosszabb

További bonyodalmak

- Hány esetet kell vizsgálni egy-egy elemi lépésben?
  - Minden oszlop betűk és betoldások kombinációja
  - Betű: 1, betoldás: 0
  - A lehetőségek száma: 2^n-1
- Heurisztikus megoldás:
  - Nem vizsgál meg minden lehetőséget
  - Nem garantált, hogy a legjobb megoldást találja meg
  - Eljut egy elég jó megoldáshoz – gyorsan

Progresszív (hierarchikus) módszer

- A bemenő szekvenciákat páronként illesztjük (N-W szerint, n^2-tel arányos)
- Filogenetikai fát építünk ez alapján
- Kiválasztjuk a két legközelebbi rokont
- Ehhez egyesével hozzávesszük a többi szekvenciát
- A hasonlóktól haladunk a távoli rokonok felé
- Az egyszer már illesztett részt nem piszkáljuk

Az eljárás menete

Bemenő adatok

- Csirke
- Egér
- Ember
- Kukac
- Kutya
- Majom

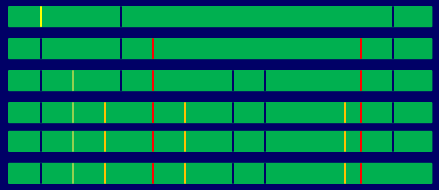
Illesztés

- Csirke – egér
- Csirke – ember
- Csirke – kukac
- .....

Rendezés

- Ember
- Majom
- Egér
- Kutya
- Csirke
- Kukac

Többszörös illesztés



A Clustal család

- Letölthető: http://www.clustal.org/
- Windows, Mac és Linux verzióban is
- Ugyanott dokumentáció, tutorial stb..
- Vagy web-en keresztül elérhető:
  http://www.ebi.ac.uk/Tools/msa/clustalw2/
  http://www.ch.embnet.org/software/ClustalW.html
- Nem kell helyi gépre feltenni, de megkötésekkel lehet csak használni

Clustal: Multiple Sequence Alignment

Multiple alignment of nucleic acid and protein sequences

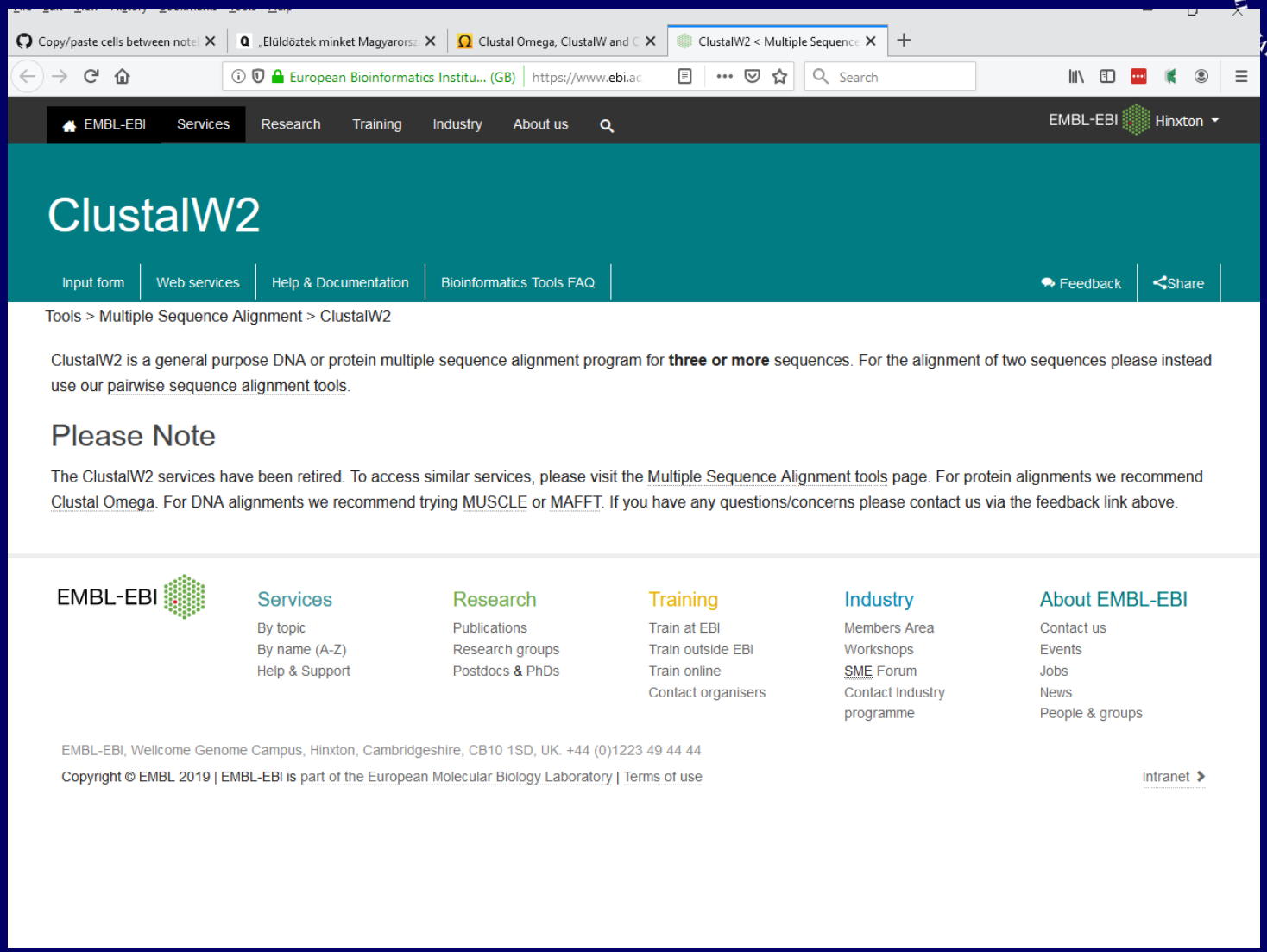
Clustal Omega

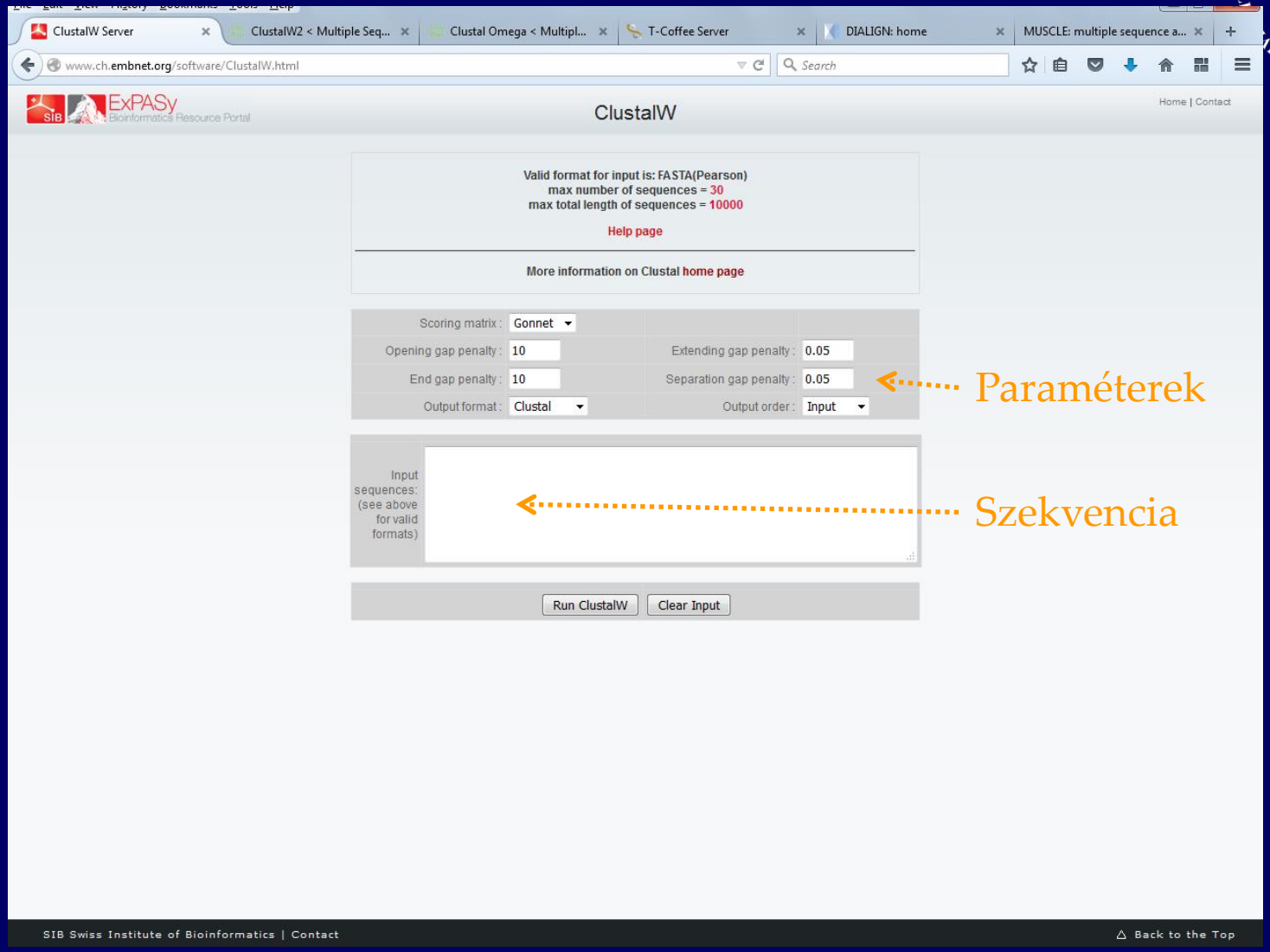
- Latest version of Clustal - fast and scalable (can align hundreds of thousands of sequences in hours), greater accuracy due to new HMM alignment engine
- Command line/web server only
- Can currently only align protein sequences

ClustalW/ClustalX

- "Classic Clustal"
- GUI (ClustalX), command line (ClustalW), web server versions available
- Can align proteins, DNA, RNA

Valid XHTML and Valid CSS | Viewable With Any Browser | Last modified on 08/04/2011 11:29:49


Copy/paste cells between note
„Elüldöztek minket Magyarorsz
Clustal Omega, ClustalW and C
ClustalW2 < Multiple Sequence
European Bioinformatics Institu... (GB)
https://www.ebi.ac
Search
EMBL-EBI
Services
Research
Training
Industry
About us
EMBL-EBI
Hinxton
ClustalW2
Input form
Web services
Help & Documentation
Bioinformatics Tools FAQ
Feedback
Share
Tools > Multiple Sequence Alignment > ClustalW2
ClustalW2 is a general purpose DNA or protein multiple sequence alignment program for three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.
Please Note
The ClustalW2 services have been retired. To access similar services, please visit the Multiple Sequence Alignment tools page. For protein alignments we recommend Clustal Omega. For DNA alignments we recommend trying MUSCLE or MAFFT. If you have any questions/concerns please contact us via the feedback link above.
EMBL-EBI
Services
By topic
By name (A-Z)
Help & Support
Research
Publications
Research groups
Postdocs & PhDs
Training
Train at EBI
Train outside EBI
Train online
Contact organisers
Industry
Members Area
Workshops
SME Forum
Contact Industry programme
About EMBL-EBI
Contact us
Events
Jobs
News
People & groups
EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. +44 (0)1223 49 44 44
Copyright © EMBL 2019 | EMBL-EBI is part of the European Molecular Biology Laboratory | Terms of use
Intranet


ClustalW Server
ClustalW2 < Multiple Seq...
Clustal Omega < Multipl...
T-Coffee Server
DIALIGN: home
MUSCLE: multiple sequence a...
www.ch.embnet.org/software/ClustalW.html
ExPASy
Bioinformatics Resource Portal
ClustalW
Home | Contact
Valid format for input is: FASTA(Pearson)
max number of sequences = 30
max total length of sequences = 10000
Help page
More information on Clustal home page
Scoring matrix : Gonnet
Opening gap penalty : 10
Extending gap penalty : 0.05
End gap penalty : 10
Separation gap penalty : 0.05
Output format : Clustal
Output order : Input
Input sequences: (see above for valid formats)
Run ClustalW
Clear Input
SIB Swiss Institute of Bioinformatics | Contact
Back to the Top
Paraméterek
Szekvencia

A Gonnet mátrix

Database: AAindex
Entry: GONG920101
LinkDB: GONG920101

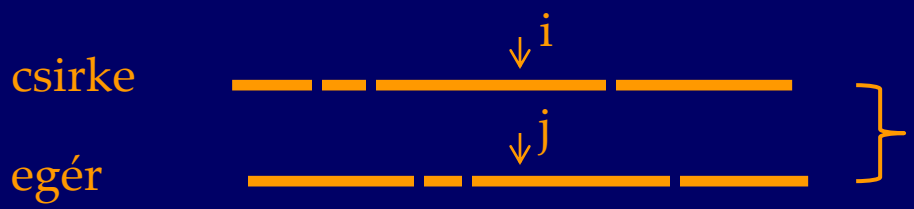
```
H GONG920101
D The mutation matrix for initially aligning (Gonnet et al., 1992)
R LIT:1813110 PMID:1604319
A Gonnet, G.H., Cohen, M.A. and Benner, S.A.
T Exhaustive matching of the entire protein sequence database
J Science 256, 1443-1445 (1992)
M rows = ARNDCQEGHILKMFPSTWYV, cols = ARNDCQEGHILKMFPSTWYV
     2.4
    -0.6     4.7
    -0.3     0.3     3.8
    -0.3    -0.3     2.2     4.7
     0.5    -2.2    -1.8    -3.2    11.5
    -0.2     1.5     0.7     0.9    -2.4     2.7
     0.0     0.4     0.9     2.7    -3.0     1.7     3.6
     0.5    -1.0     0.4     0.1    -2.0    -1.0    -0.8     6.6
    -0.8     0.6     1.2     0.4    -1.3     1.2     0.4    -1.4     6.0
    -0.8    -2.4    -2.8    -3.8    -1.1    -1.9    -2.7    -4.5    -2.2     4.0
    -1.2    -2.2    -3.0    -4.0    -1.5    -1.6    -2.8    -4.4    -1.9     2.8     4.0
    -0.4     2.7     0.8     0.5    -2.8     1.5     1.2    -1.1     0.6    -2.1    -2.1     3.2
    -0.7    -1.7    -2.2    -3.0    -0.9    -1.0    -2.0    -3.5    -1.3     2.5     2.8    -1.4     4.3
    -2.3    -3.2    -3.1    -4.5    -0.8    -2.6    -3.9    -5.2    -0.1     1.0     2.0    -3.3     1.6     7.0
     0.3    -0.9    -0.9    -0.7    -3.1    -0.2    -0.5    -1.6    -1.1    -2.6    -2.3    -0.6    -2.4    -3.8     7.6
     1.1    -0.2     0.9     0.5     0.1     0.2     0.2     0.4    -0.2    -1.8    -2.1     0.1    -1.4    -2.8     0.4     2.2
     0.6    -0.2     0.5     0.0    -0.5     0.0    -0.1    -1.1    -0.3    -0.6    -1.3     0.1    -0.6    -2.2     0.1     1.5     2.5
    -3.6    -1.6    -3.6    -5.2    -1.0    -2.7    -4.3    -4.0    -0.8    -1.8    -0.7    -3.5    -1.0     3.6    -5.0    -3.3    -3.5    14.2
    -2.2    -1.8    -1.4    -2.8    -0.5    -1.7    -2.7    -4.0     2.2    -0.7     0.0    -2.1    -0.2     5.1    -3.1    -1.9    -1.9     4.1     7.8
     0.1    -2.0    -2.2    -2.9     0.0    -1.5    -1.9    -3.3    -2.0     3.1     1.8    -1.7     1.6     0.1    -1.8    -1.0     0.0    -2.6    -1.1     3.4
//
```

A T-Coffee család

- Honlap: http://tcoffee.crg.cat/
- Letölthető, de csak Linux rendszerre
- A honlapon szerver szolgáltatás is elérhető
- Van lehetőség az összes lehetséges paraméter állítására
- Fő eltérés: képes más programok illesztéseit kombinálni
- Felhasználható szerkezeti információ is

Az algoritmus

- A bemenő szekvenciákat páronként illesztjük
- Ebből „könyvtárat” készítünk: ez az egymásnak megfelelő aminósavak listája
- Az így kapott lista-elemekhez súlyokat rendelünk
- Bővítjük a listát: egy harmadik szekvencián keresztül is összetartozik a két aminósav?
- Ha igen, megnöveljük az eredeti súlyfaktort
- Az illesztés a súlyfaktorok alapján készül

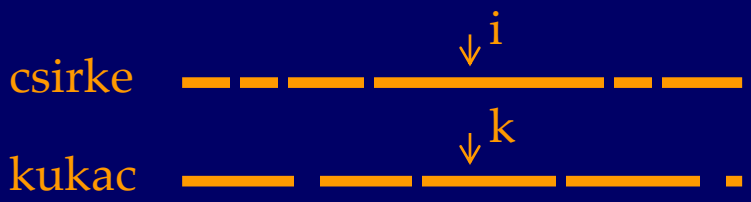


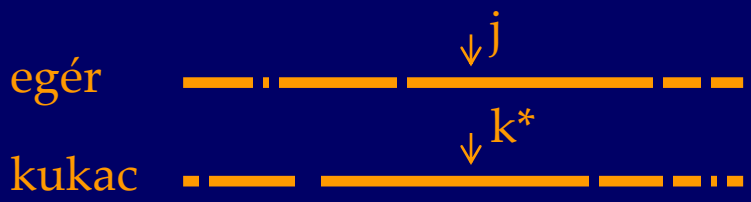
„P" konzervált A.A.

Könyvtári bejegyzés: csirke(i) – egér(j)

súlyozás: W(csirke(i),egér(j))=P

A könyvtár kiterjesztése:





W'(csirke(i),kukac(k))

W''(egér(j),kukac(k*))

k = k* ?

W'' a kisebb a kettő közül

W + W''

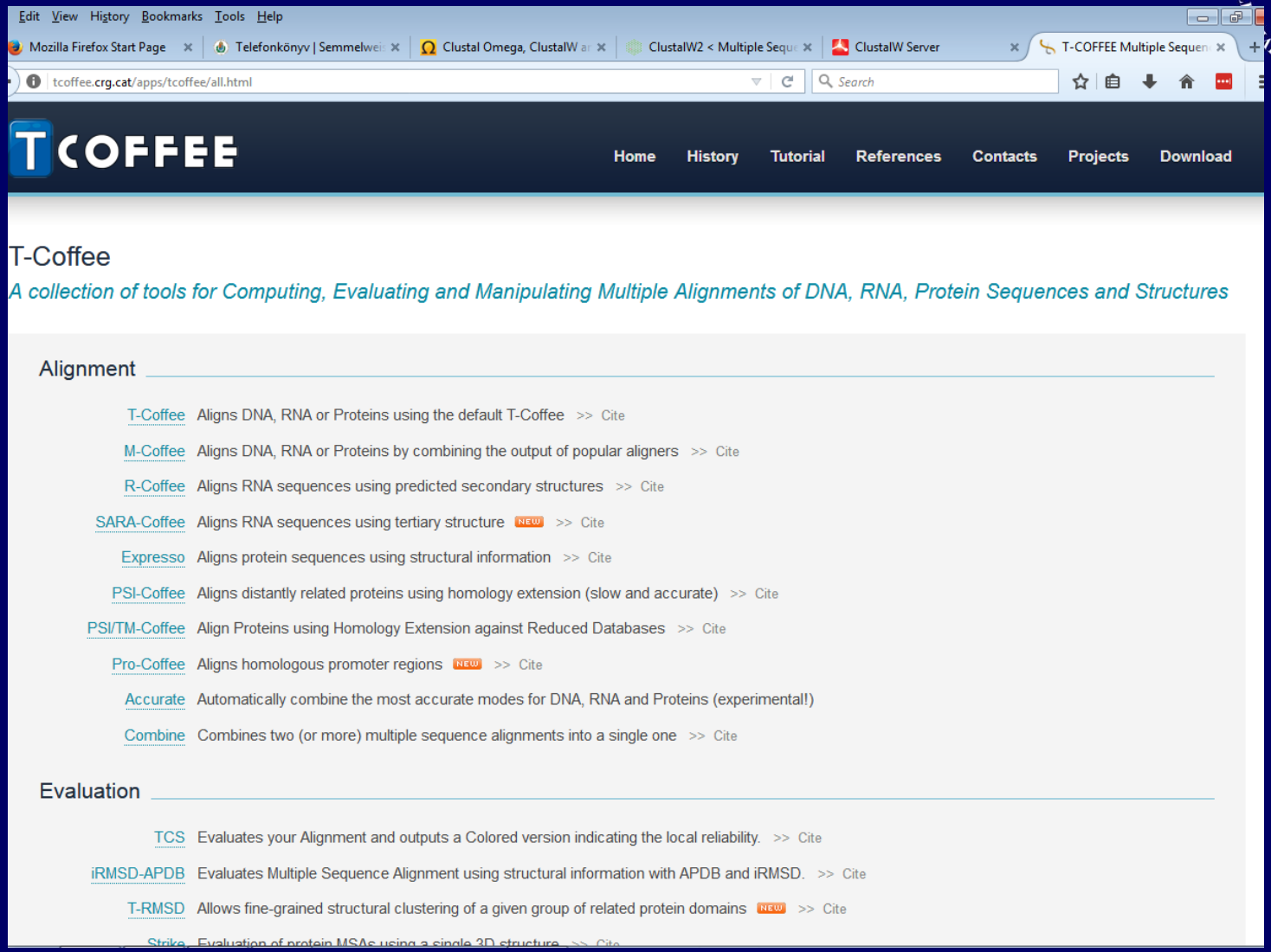
Stb..

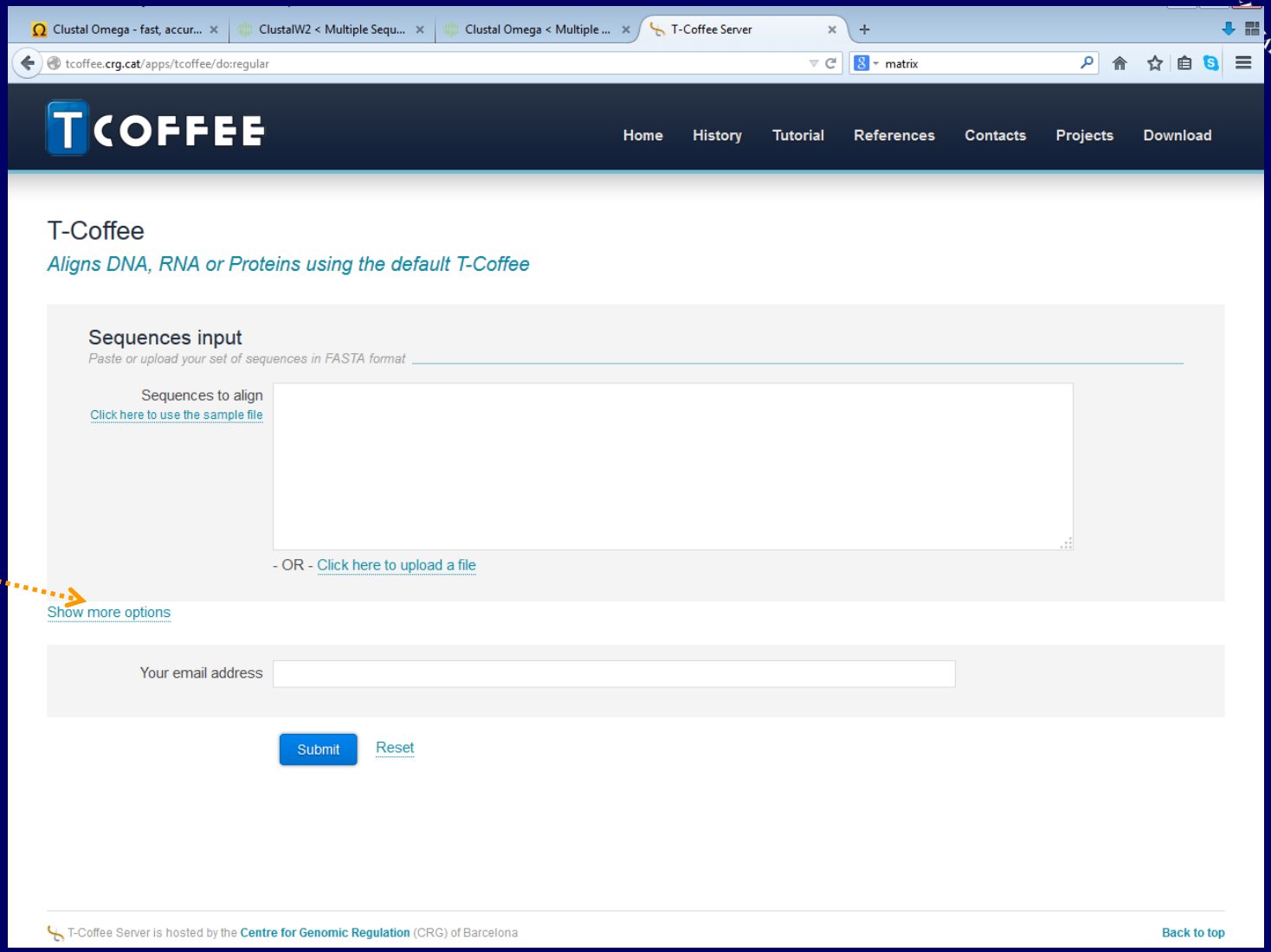
Progresszív illesztés

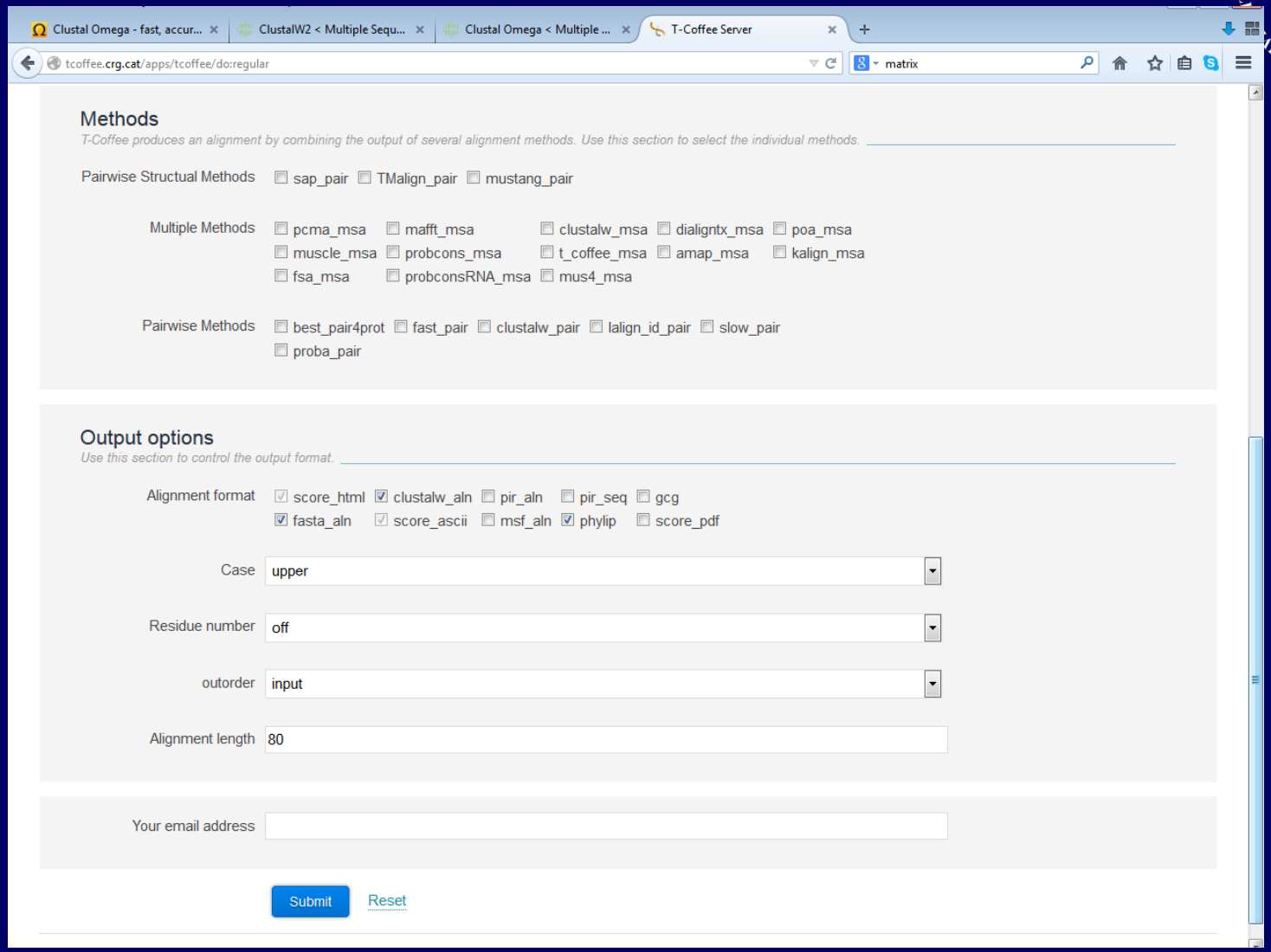
- A szekvenciákat páronként illesztjük
- Először a leginkább hasonlókat vesszük
- A távolabbi rokonok felé haladunk
- Az illesztőfelületet a súlyozó faktorok adják

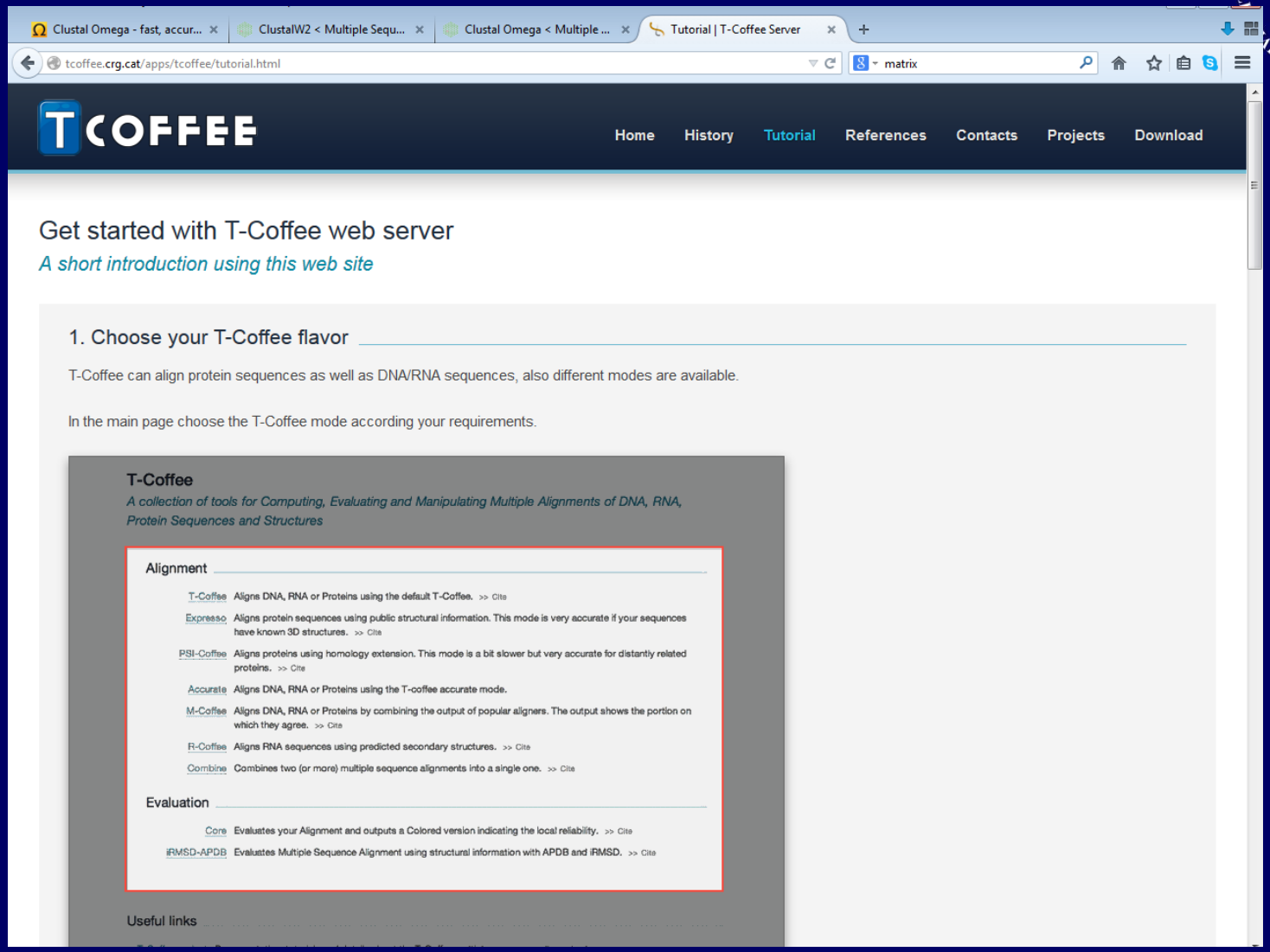
A módszer tulajdonságai

- Több szekvencia-illsztést is fel lehet használni (lokális és globális eredményt is)
- Az egyes eredmények nem feltétlenül vannak összhangban egymással
- Ilyenkor győzzön a jobb (a nagyobb súlyú)
- Pontosabb eredményt ad ClustalW-nál, és elég gyors is


Edit View History Bookmarks Tools Help
Mozilla Firefox Start Page
Telefonkönyv | Semmelweis
Clustal Omega, ClustalW
ClustalW2 < Multiple Seque
ClustalW Server
T-COFFEE Multiple Sequen
tcoffee.crg.cat/apps/tcoffee/all.html
Search
TCOFFEE
Home
History
Tutorial
References
Contacts
Projects
Download
T-Coffee
A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures
Alignment
T-Coffee Aligns DNA, RNA or Proteins using the default T-Coffee >> Cite
M-Coffee Aligns DNA, RNA or Proteins by combining the output of popular aligners >> Cite
R-Coffee Aligns RNA sequences using predicted secondary structures >> Cite
SARA-Coffee Aligns RNA sequences using tertiary structure NEW >> Cite
Expresso Aligns protein sequences using structural information >> Cite
PSI-Coffee Aligns distantly related proteins using homology extension (slow and accurate) >> Cite
PSI/TM-Coffee Align Proteins using Homology Extension against Reduced Databases >> Cite
Pro-Coffee Aligns homologous promoter regions NEW >> Cite
Accurate Automatically combine the most accurate modes for DNA, RNA and Proteins (experimental!)
Combine Combines two (or more) multiple sequence alignments into a single one >> Cite
Evaluation
TCS Evaluates your Alignment and outputs a Colored version indicating the local reliability. >> Cite
iRMSD-APDB Evaluates Multiple Sequence Alignment using structural information with APDB and iRMSD. >> Cite
T-RMSD Allows fine-grained structural clustering of a given group of related protein domains NEW >> Cite


Clustal Omega - fast, accur...
ClustalW2 < Multiple Sequ...
Clustal Omega < Multiple ...
T-Coffee Server
tcoffee.crg.cat/apps/tcoffee/do:regular
matrix
TCOFFEE
Home
History
Tutorial
References
Contacts
Projects
Download
T-Coffee
Aligns DNA, RNA or Proteins using the default T-Coffee
Sequences input
Paste or upload your set of sequences in FASTA format
Sequences to align
Click here to use the sample file
- OR - Click here to upload a file
Show more options
Your email address
Submit
Reset
T-Coffee Server is hosted by the Centre for Genomic Regulation (CRG) of Barcelona
Back to top


Clustal Omega - fast, accur...
ClustalW2 < Multiple Sequ...
Clustal Omega < Multiple ...
T-Coffee Server
tcoffee.crg.cat/apps/tcoffee/do:regular
matrix
Methods
T-Coffee produces an alignment by combining the output of several alignment methods. Use this section to select the individual methods.
Pairwise Structual Methods
sap_pair TMalign_pair mustang_pair
Multiple Methods
pcma_msa mafft_msa clustalw_msa dialigntx_msa poa_msa
muscle_msa probcons_msa t_coffee_msa amap_msa kalign_msa
fsa_msa probconsRNA_msa mus4_msa
Pairwise Methods
best_pair4prot fast_pair clustalw_pair lalign_id_pair slow_pair
proba_pair
Output options
Use this section to control the output format.
Alignment format
score_html clustalw_aln pir_aln pir_seq gcg
fasta_aln score_ascii msf_aln phylip score_pdf
Case
upper
Residue number
off
outorder
input
Alignment length
80
Your email address
Submit
Reset

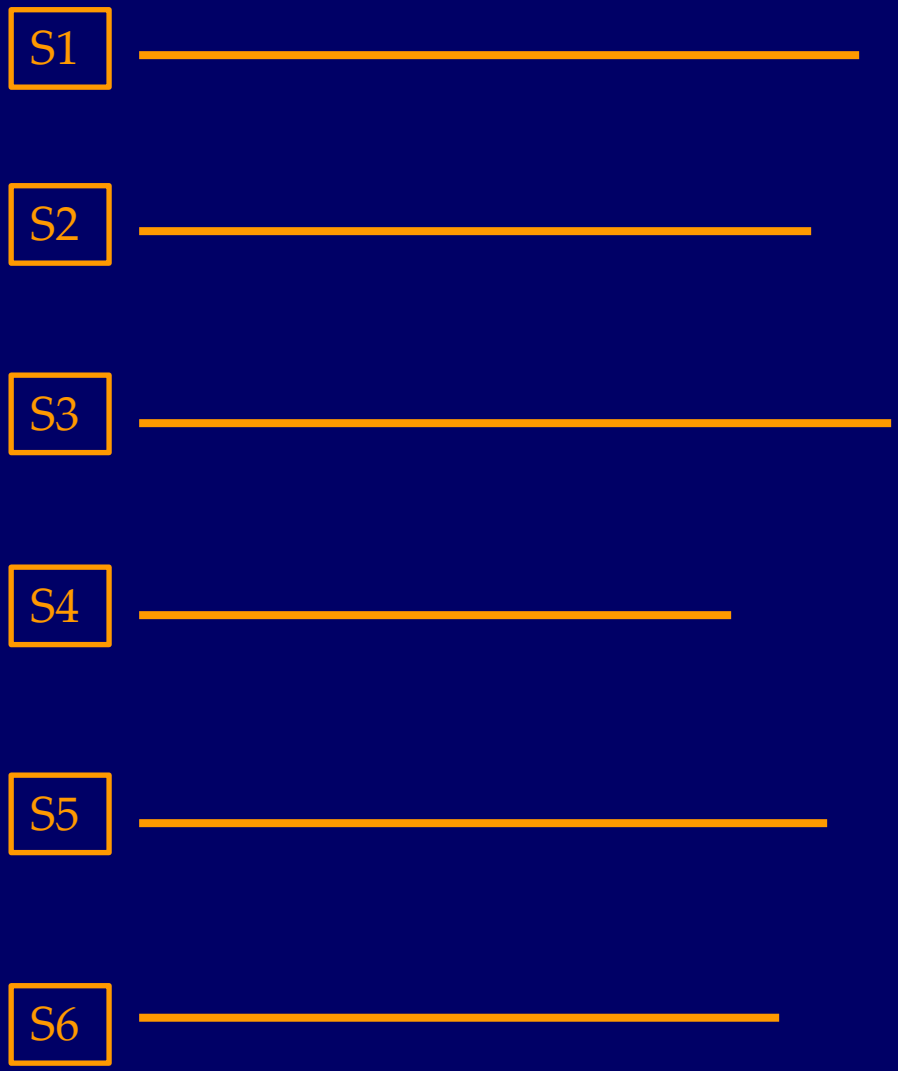

Clustal Omega - fast, accur...
ClustalW2 < Multiple Sequ...
Clustal Omega < Multiple ...
Tutorial | T-Coffee Server
tcoffee.crg.cat/apps/tcoffee/tutorial.html
matrix
TCOFFEE
Home
History
Tutorial
References
Contacts
Projects
Download
Get started with T-Coffee web server
A short introduction using this web site
1. Choose your T-Coffee flavor
T-Coffee can align protein sequences as well as DNA/RNA sequences, also different modes are available.
In the main page choose the T-Coffee mode according your requirements.
T-Coffee
A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures
Alignment
T-Coffee Aligns DNA, RNA or Proteins using the default T-Coffee. >> Cite
Expresso Aligns protein sequences using public structural information. This mode is very accurate if your sequences have known 3D structures. >> Cite
PSI-Coffee Aligns proteins using homology extension. This mode is a bit slower but very accurate for distantly related proteins. >> Cite
Accurate Aligns DNA, RNA or Proteins using the T-coffee accurate mode.
M-Coffee Aligns DNA, RNA or Proteins by combining the output of popular aligners. The output shows the portion on which they agree. >> Cite
R-Coffee Aligns RNA sequences using predicted secondary structures. >> Cite
Combine Combines two (or more) multiple sequence alignments into a single one. >> Cite
Evaluation
Core Evaluates your Alignment and outputs a Colored version indicating the local reliability. >> Cite
iRMSD-APDB Evaluates Multiple Sequence Alignment using structural information with APDB and iRMSD. >> Cite
Useful links

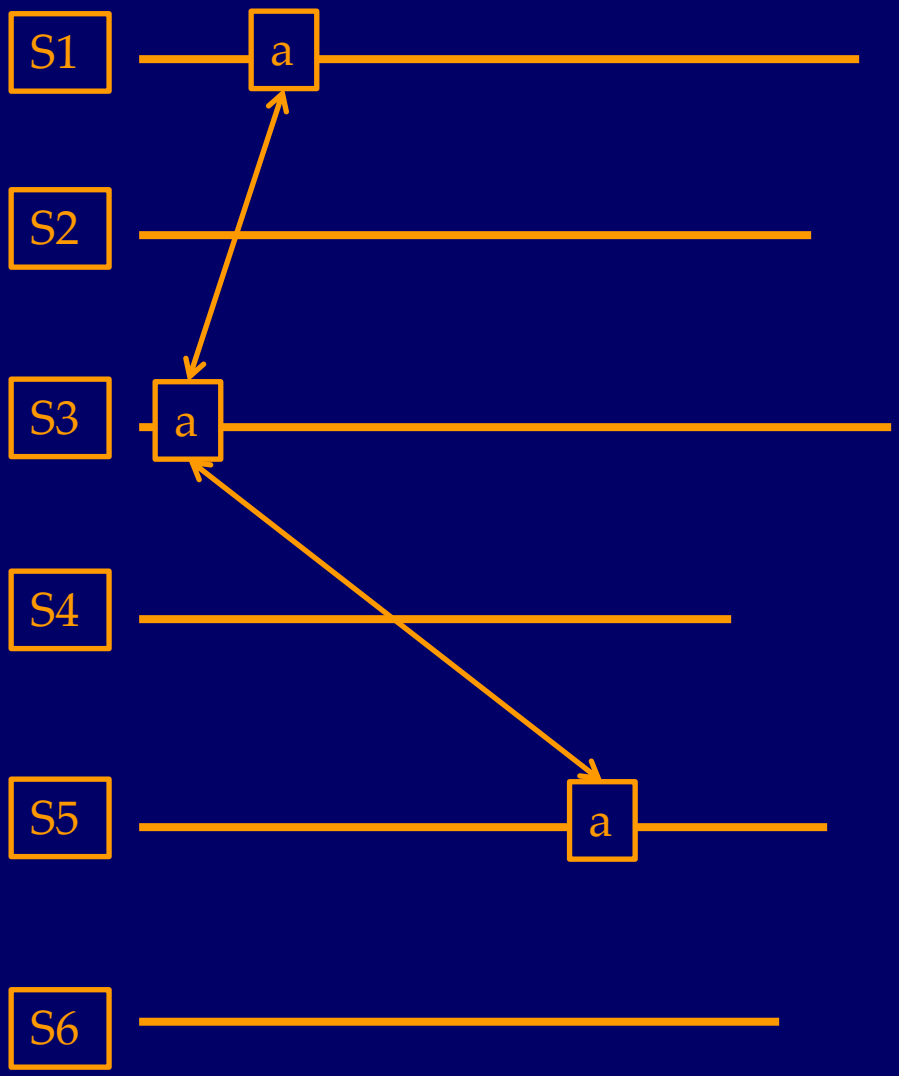
A DIALIGN rendszer

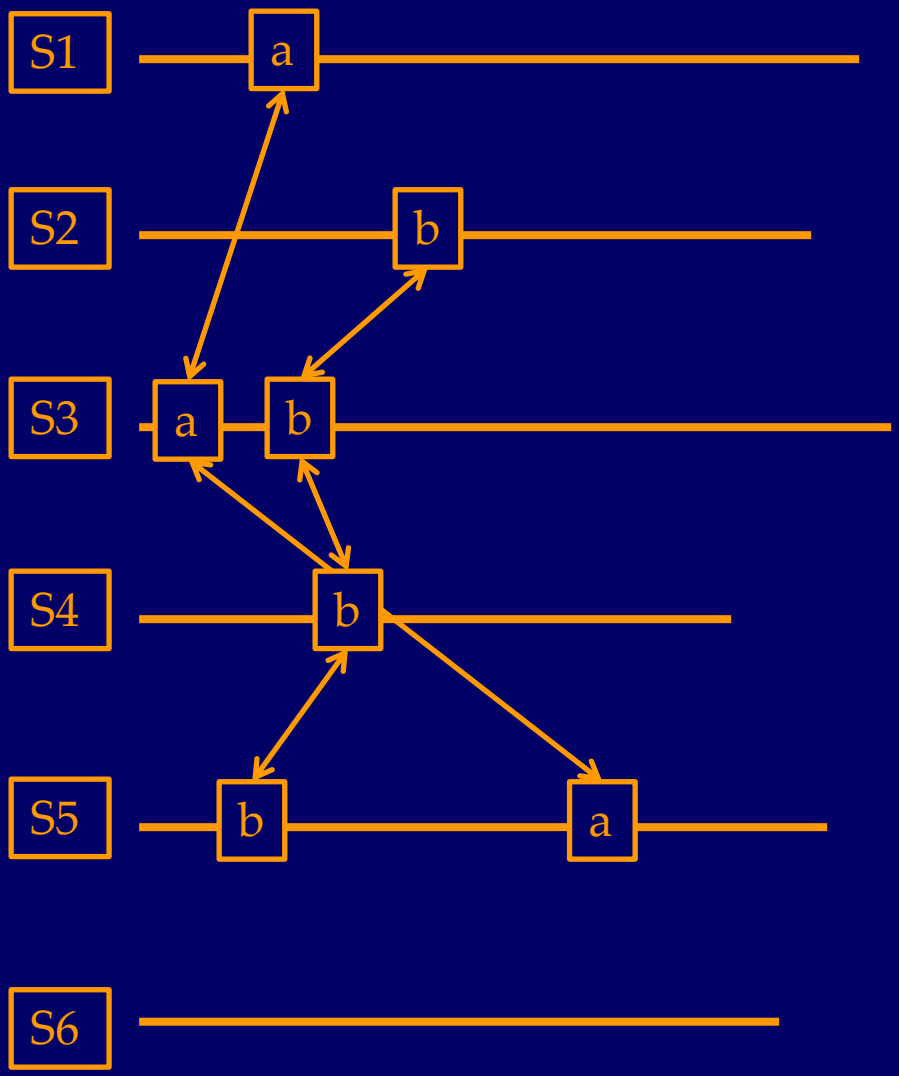
- Honlap: http://dialign.gobics.de/
- Letölthető Linux rendszerre
- Web-en keresztül is használható
- Ugyanott elérhető: interaktív illesztés megjelenítő
- Fehérje és DNS szekvenciát is kezel

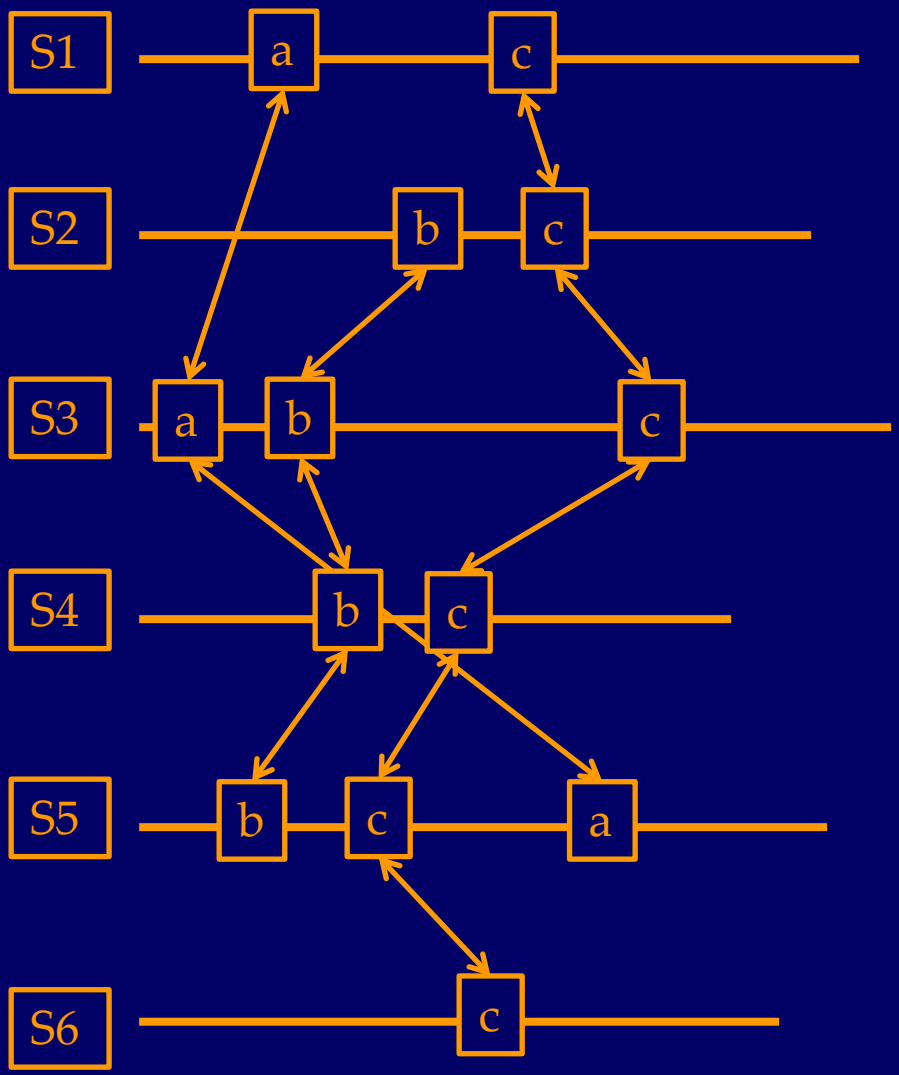
Hogyan működik

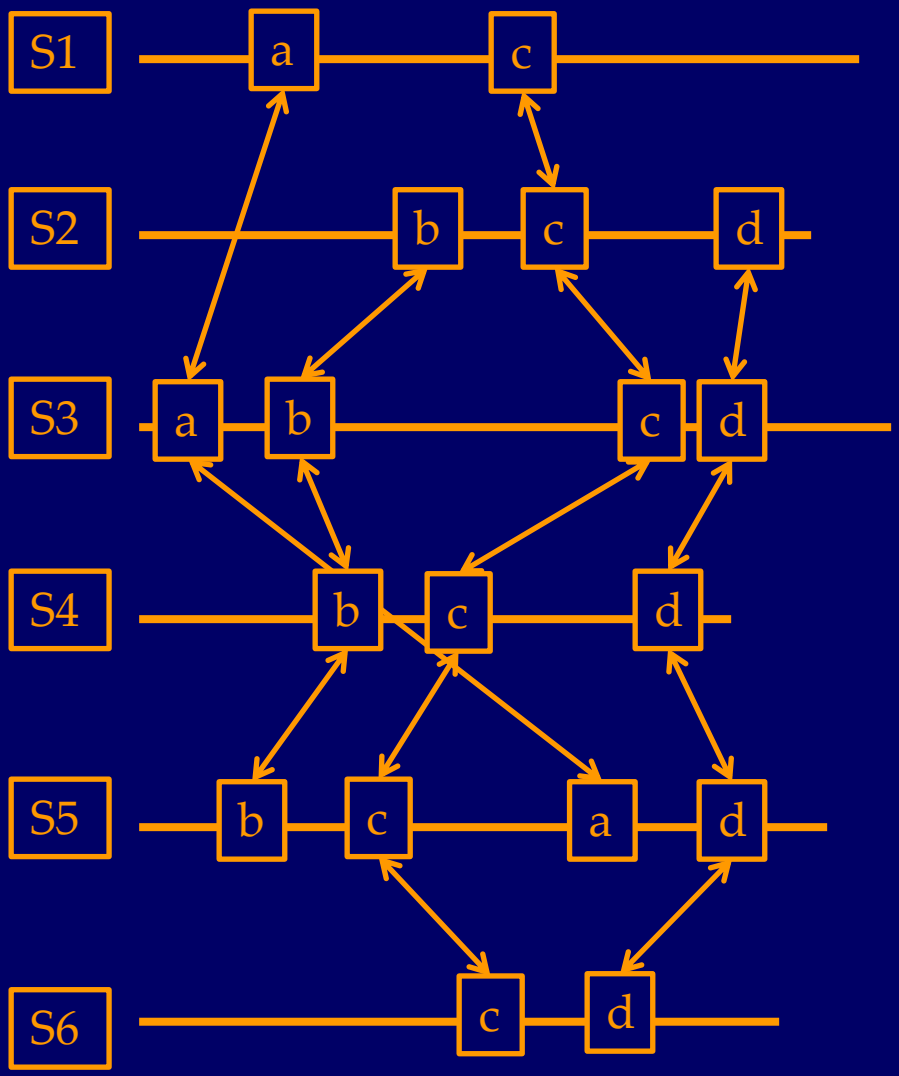
- Rövid, toldás nélküli szakaszokat keres
- Ezek adják majd az illsztés vázát
- A talált szakaszok nem feltétlenül képesek konzisztens illesztést adni
- Ki kell dobni az inkonzisztens szakaszokat
- A többi horgony-pontokat jelöl ki az illesztésben
- A horgony-pontokból kiindulva illesztünk

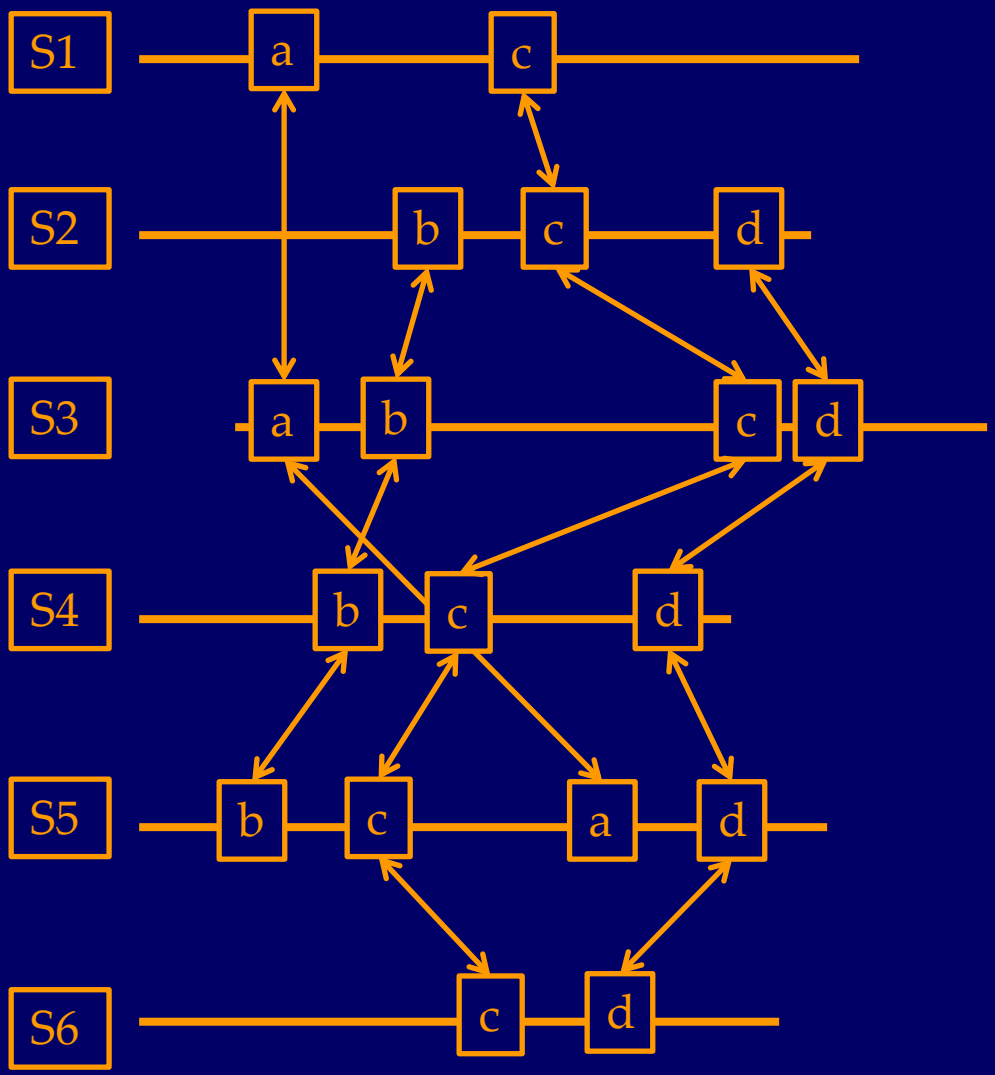

S1
S2
S3
S4
S5
S6

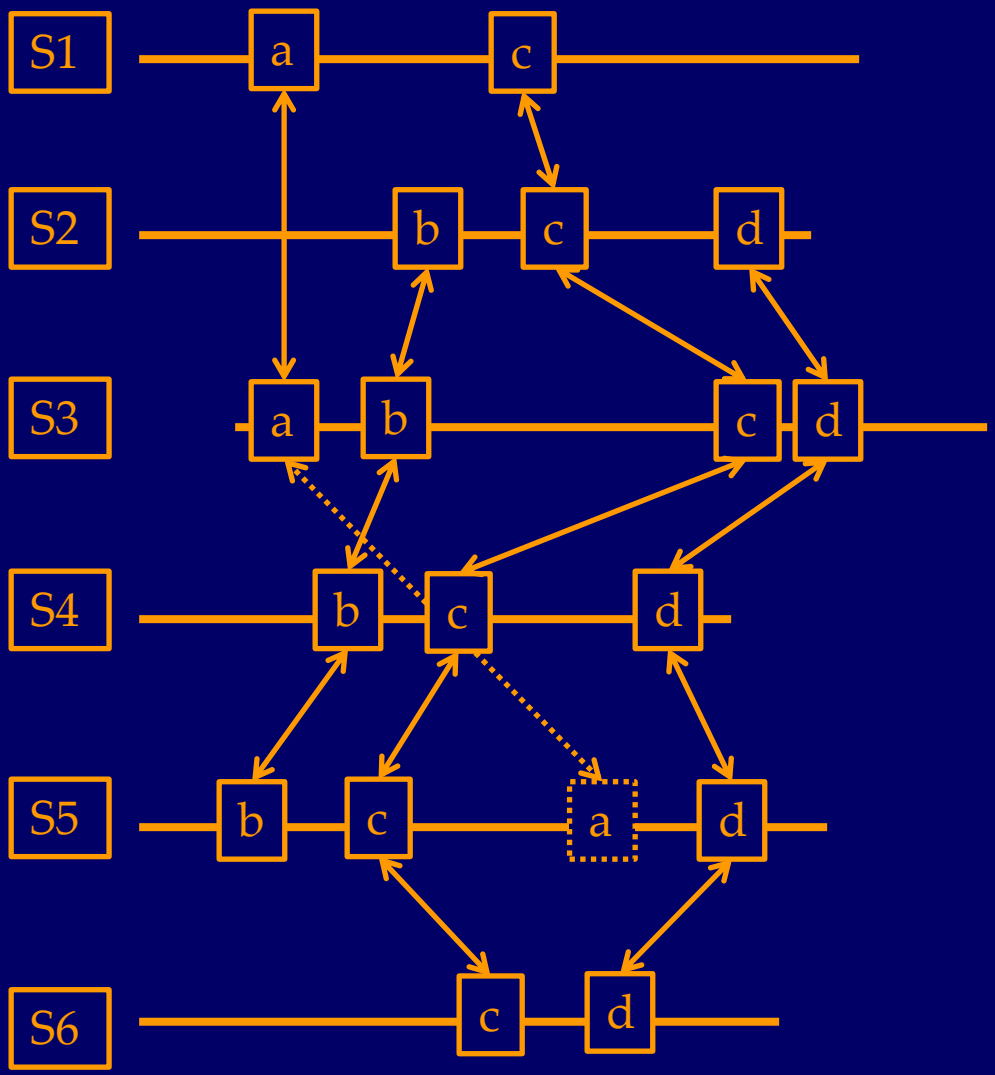

S1
a
S2
S3
a
S4
S5
a
S6

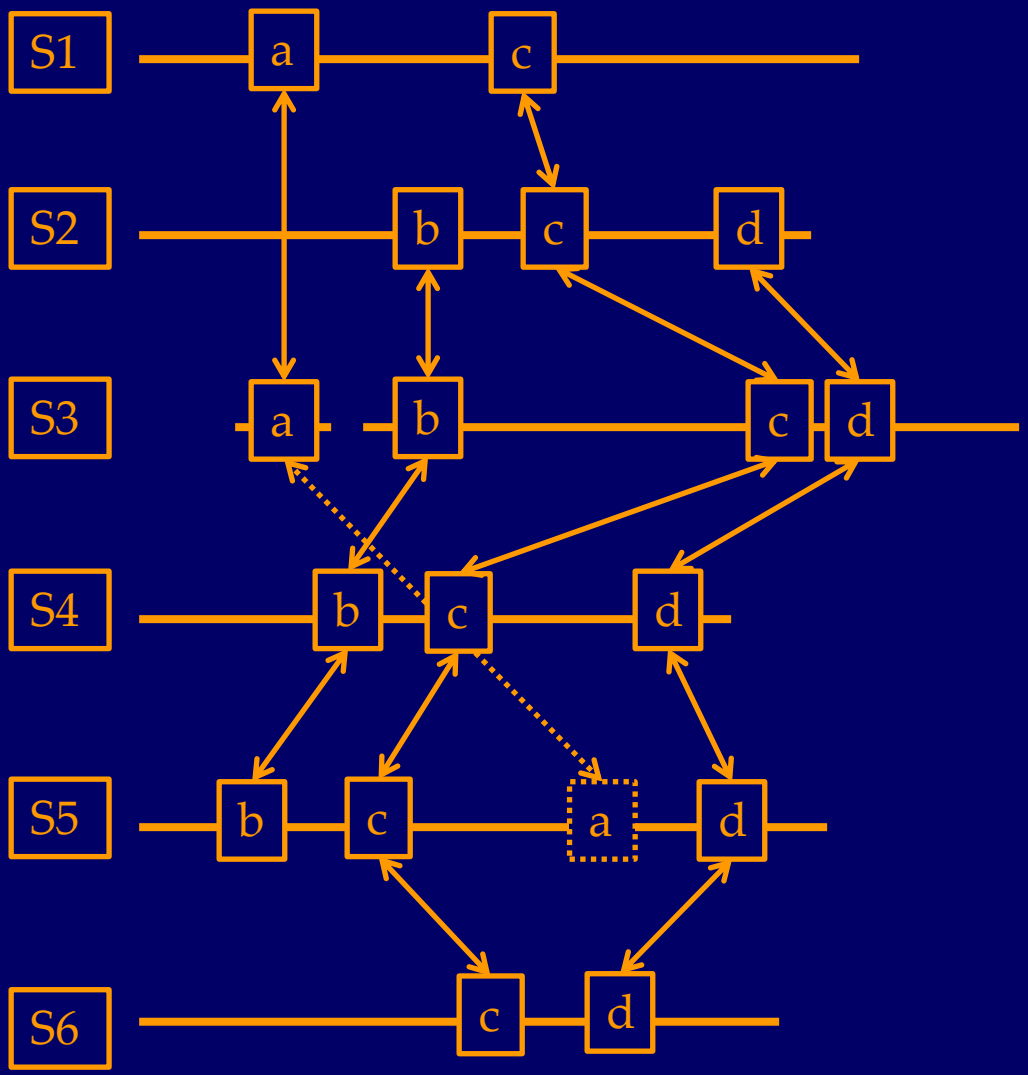

S1
a
S2
b
S3
a
b
S4
b
S5
b
a
S6

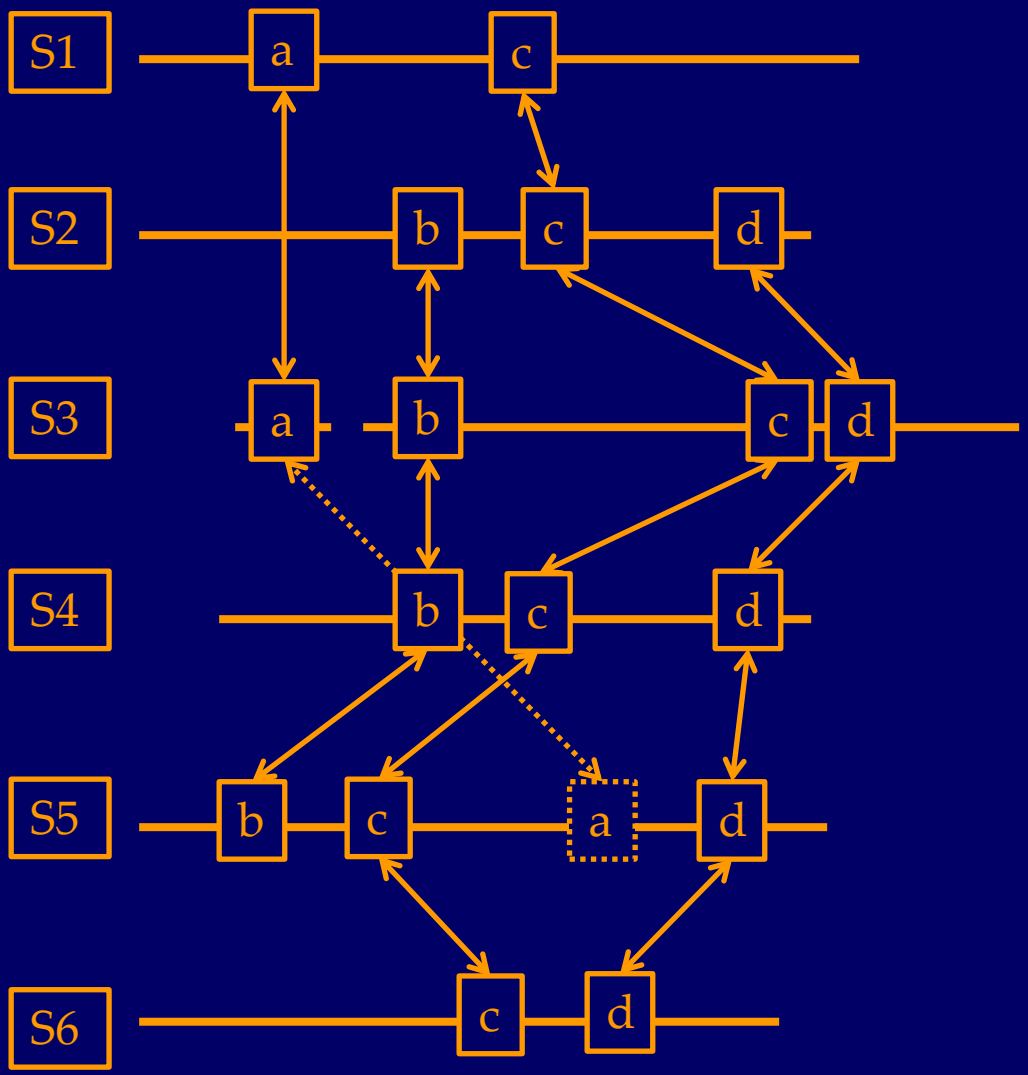

S1
a
c
S2
b
c
S3
a
b
c
S4
b
c
S5
b
c
a
S6
c

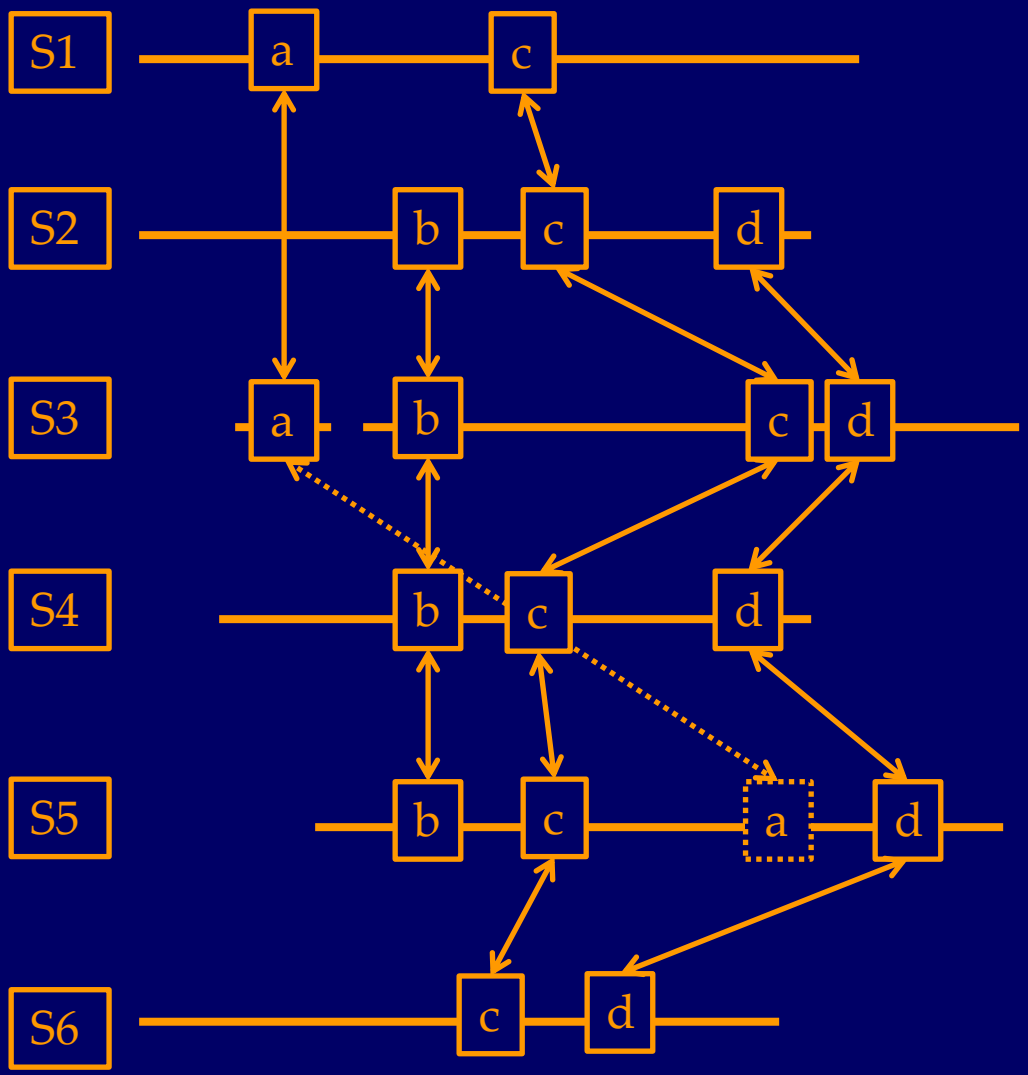

S1
a
c
S2
b
c
d
S3
a
b
c
d
S4
b
c
d
S5
b
c
a
d
S6
c
d

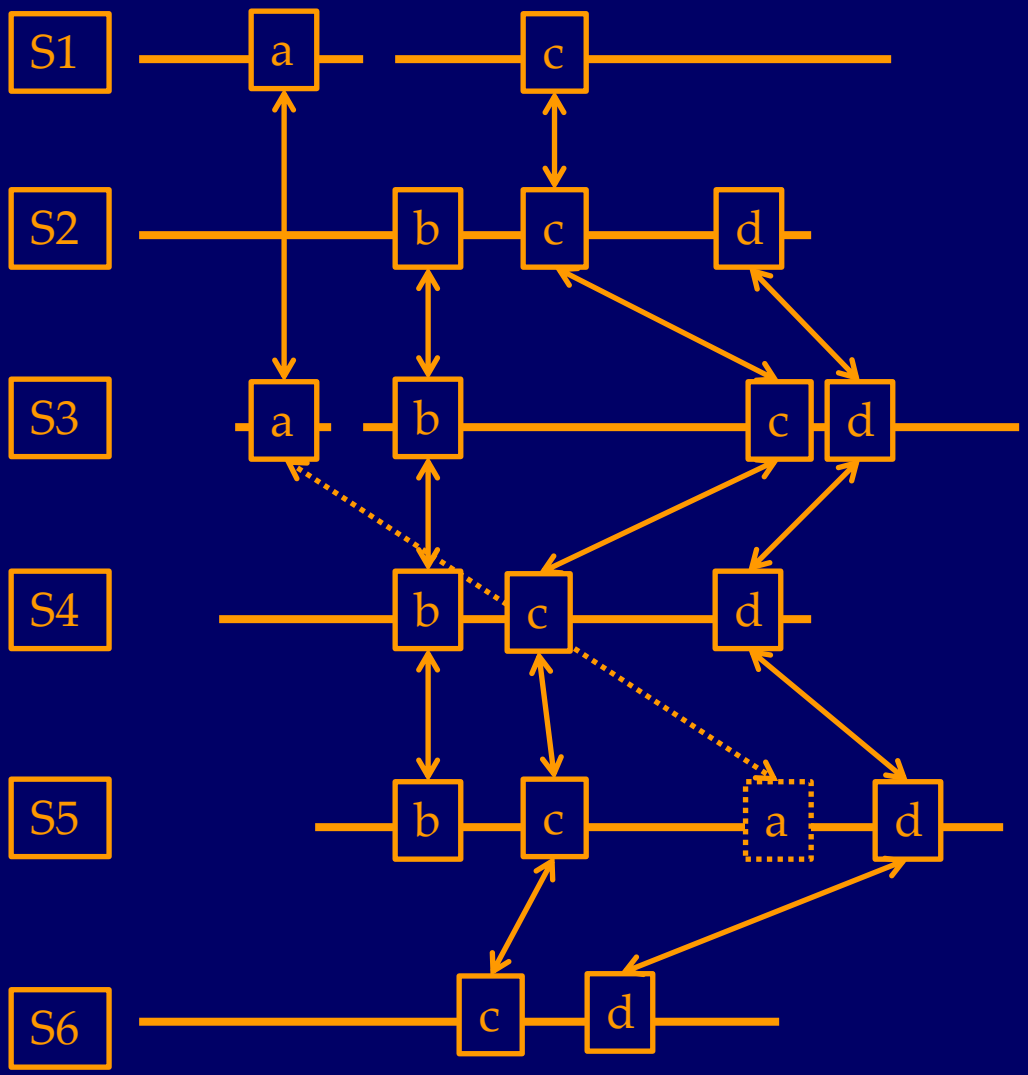

S1
a
c
S2
b
c
d
S3
a
b
c
d
S4
b
c
d
S5
b
c
a
d
S6
c
d

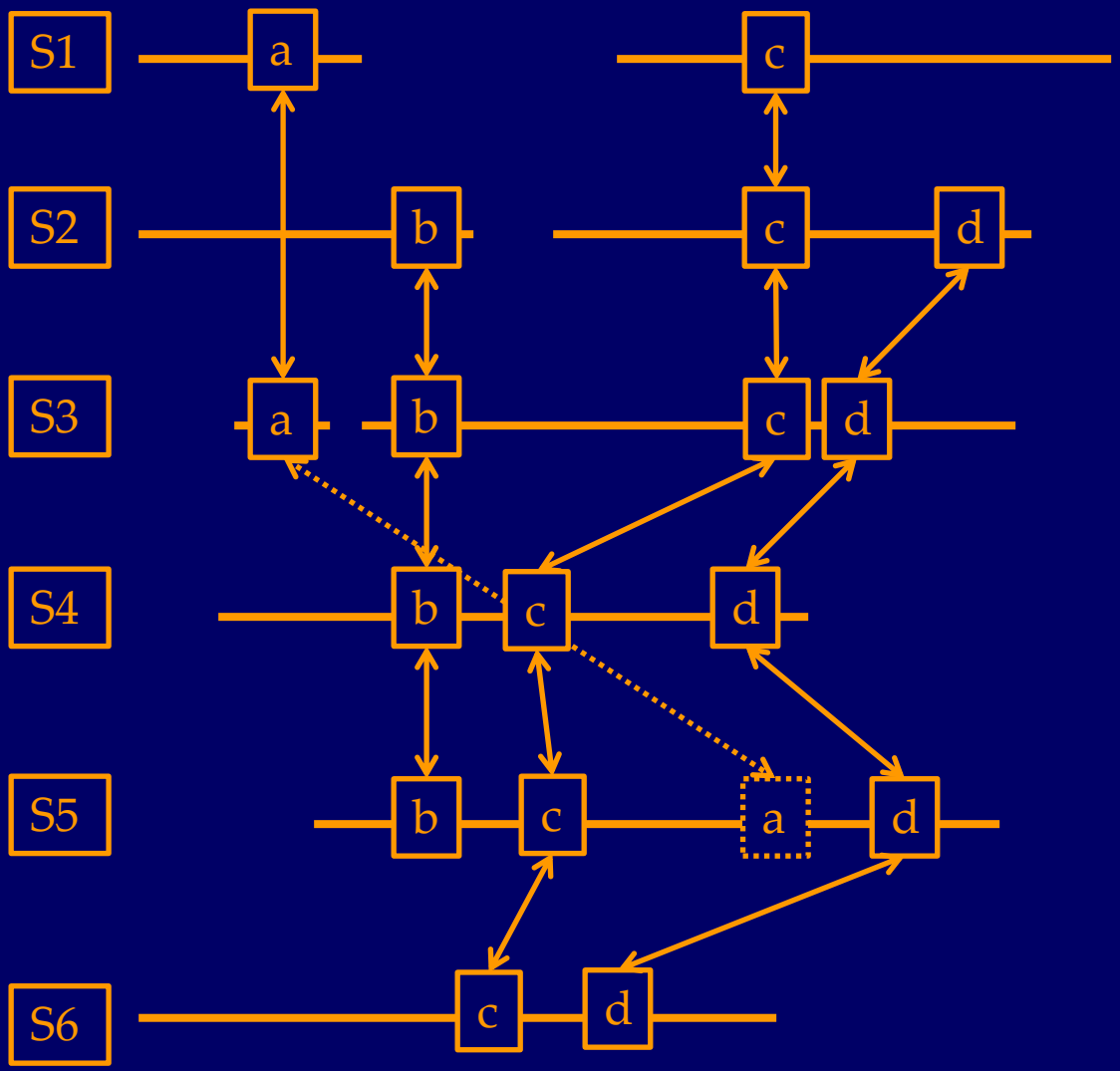

S1
a
c
S2
b
c
d
S3
a
b
c
d
S4
b
c
d
S5
b
c
a
d
S6
c
d

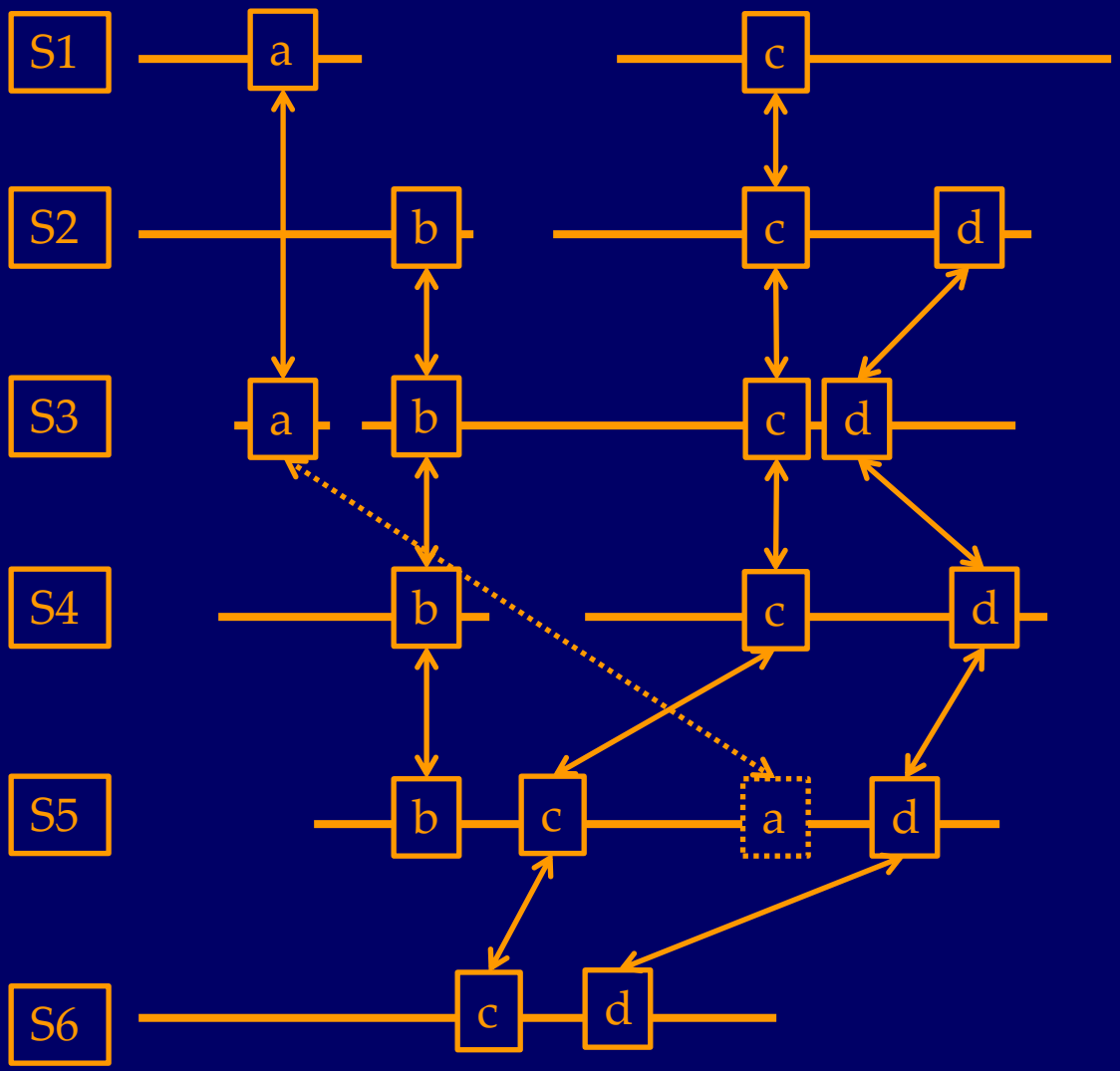

S1
a
c
S2
b
c
d
S3
a
b
c
d
S4
b
c
d
S5
b
c
a
d
S6
c
d

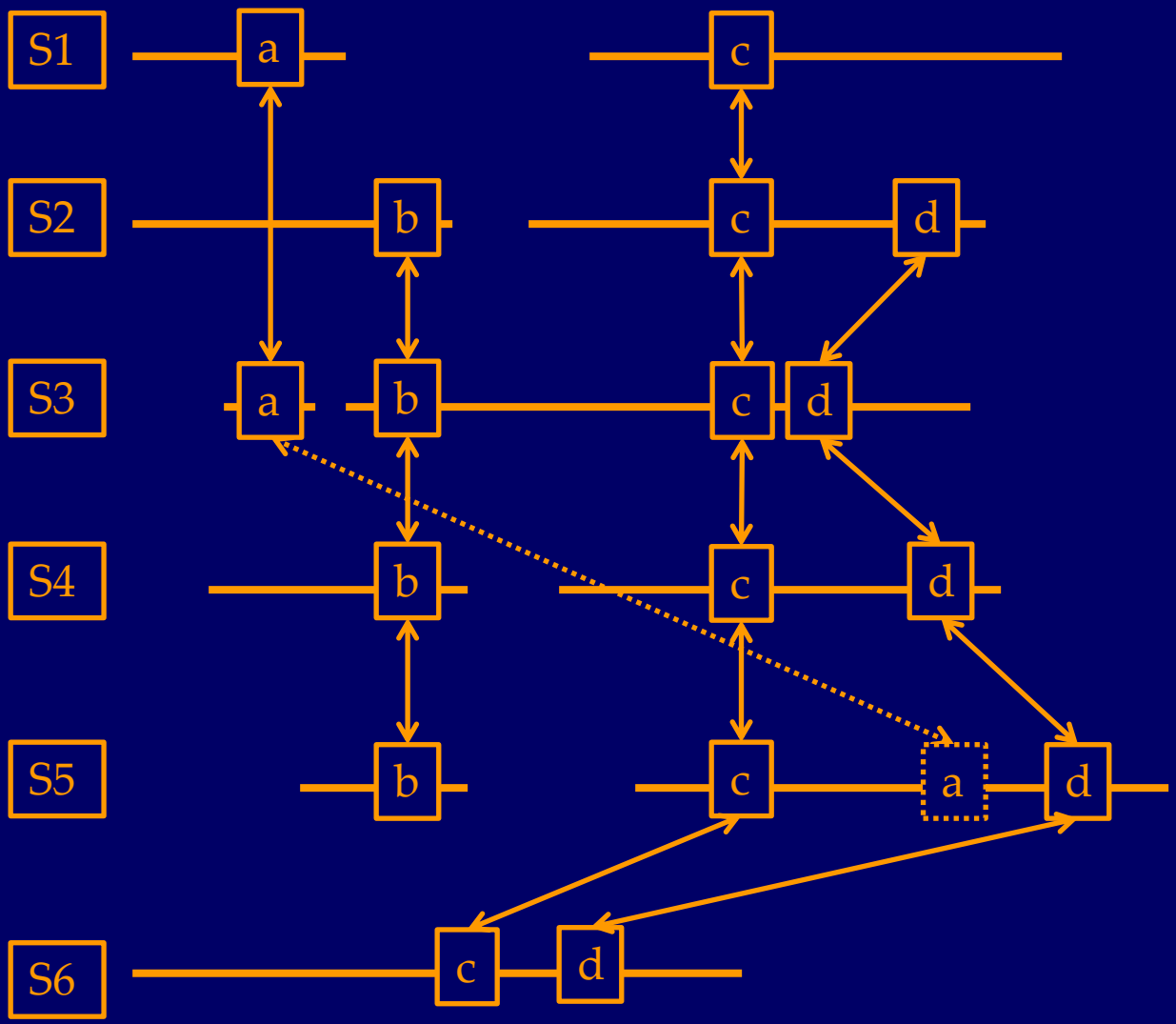

S1
a
c
S2
b
c
d
S3
a
b
c
d
S4
b
c
d
S5
b
c
a
d
S6
c
d

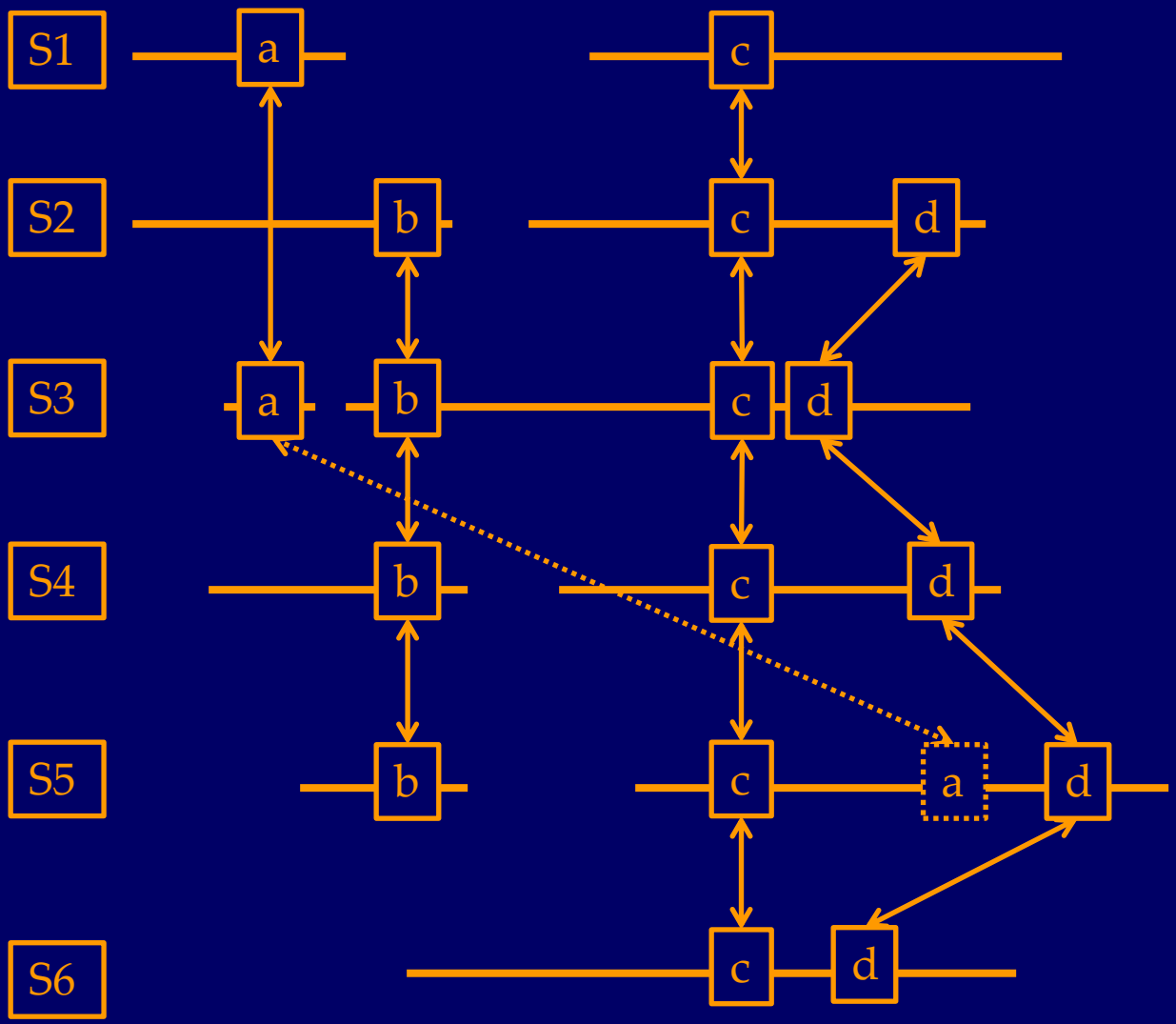

S1
a
c
S2
b
c
d
S3
a
b
c
d
S4
b
c
d
S5
b
c
a
d
S6
c
d

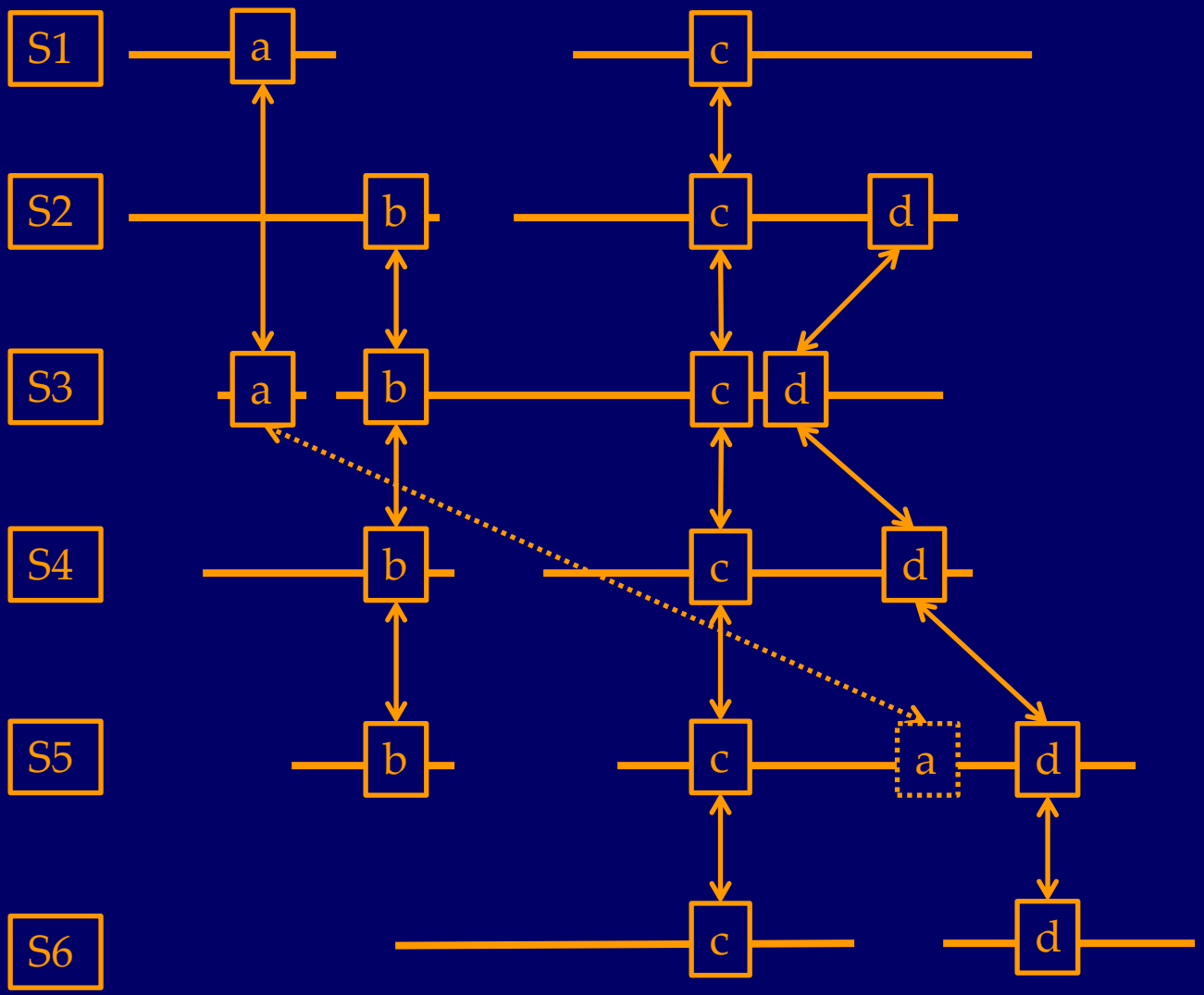

S1
a
c
S2
b
c
d
S3
a
b
c
d
S4
b
c
d
S5
b
c
a
d
S6
c
d

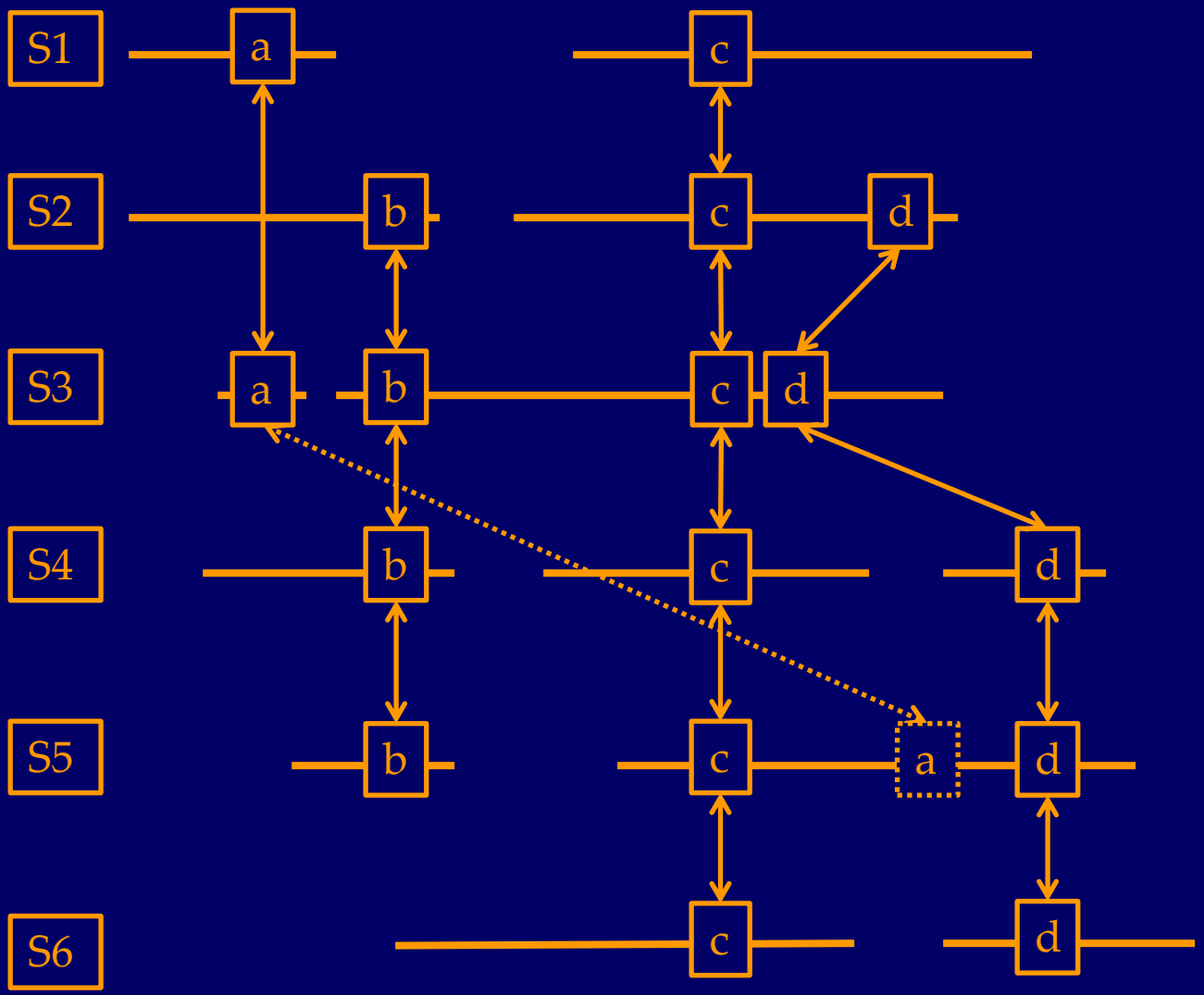

S1
a
c
S2
b
c
d
S3
a
b
c
d
S4
b
c
d
S5
b
c
a
d
S6
c
d

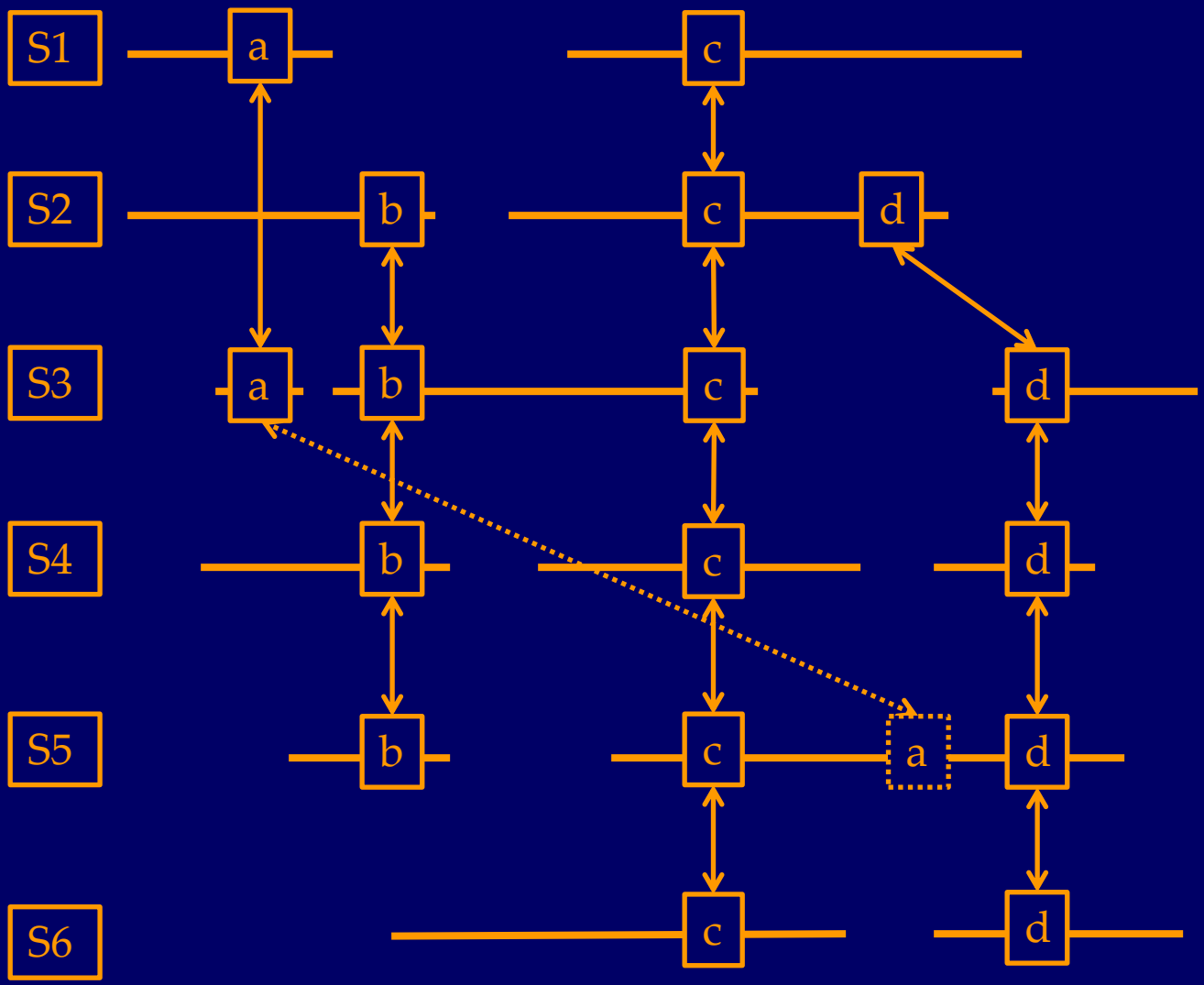

S1
a
c
S2
b
c
d
S3
a
b
c
d
S4
b
c
d
S5
b
c
a
d
S6
c
d

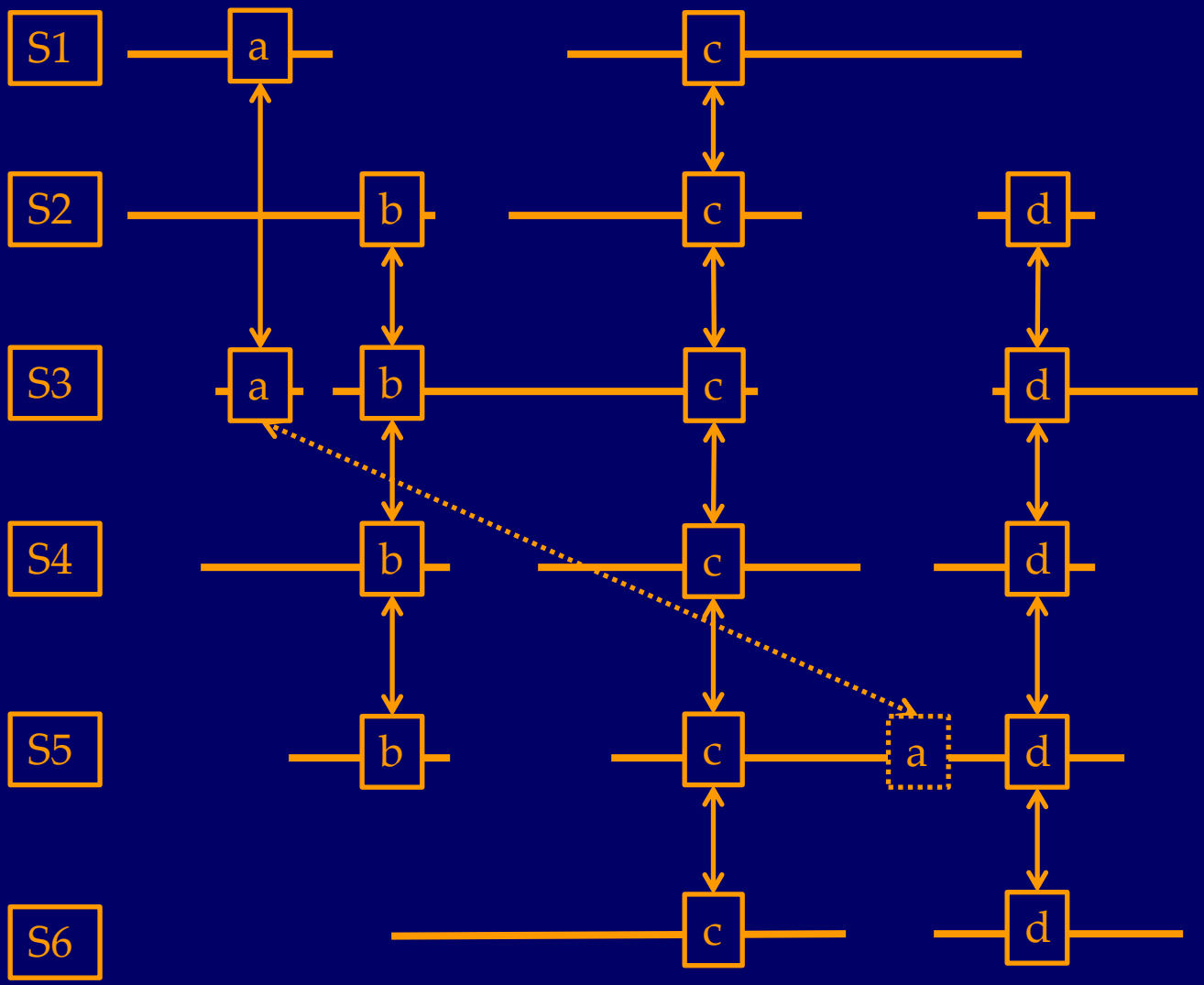

S1
S2
S3
S4
S5
S6
a
b
c
d
a
b
c
d
b
c
d
b
c
a
d
c
d

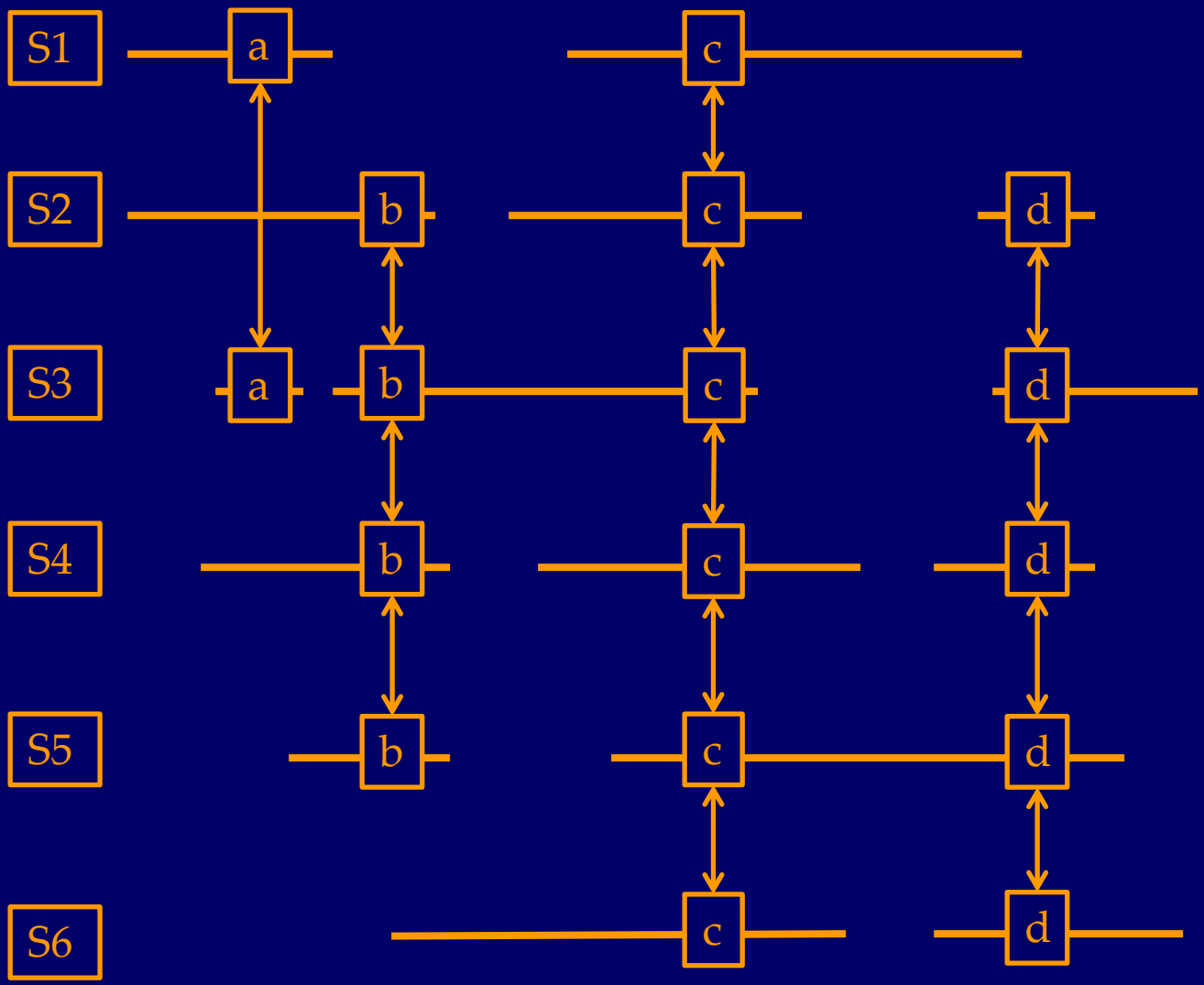

S1
S2
S3
S4
S5
S6
a
b
c
d
a
b
c
d
b
c
d
b
c
a
d
c
d

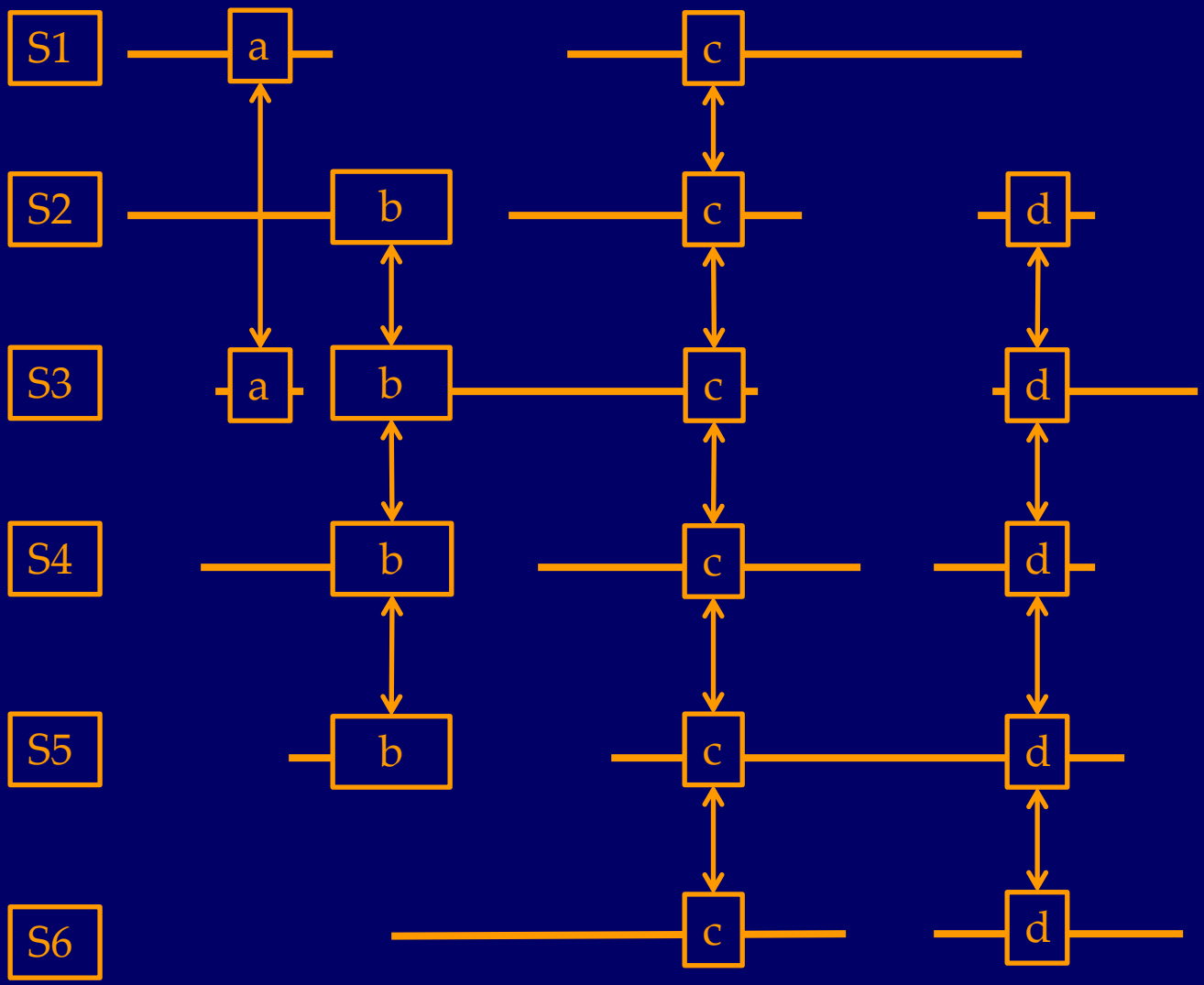

S1
S2
S3
S4
S5
S6
a
b
c
d
a
b
c
d
b
c
d
b
c
a
d
c
d

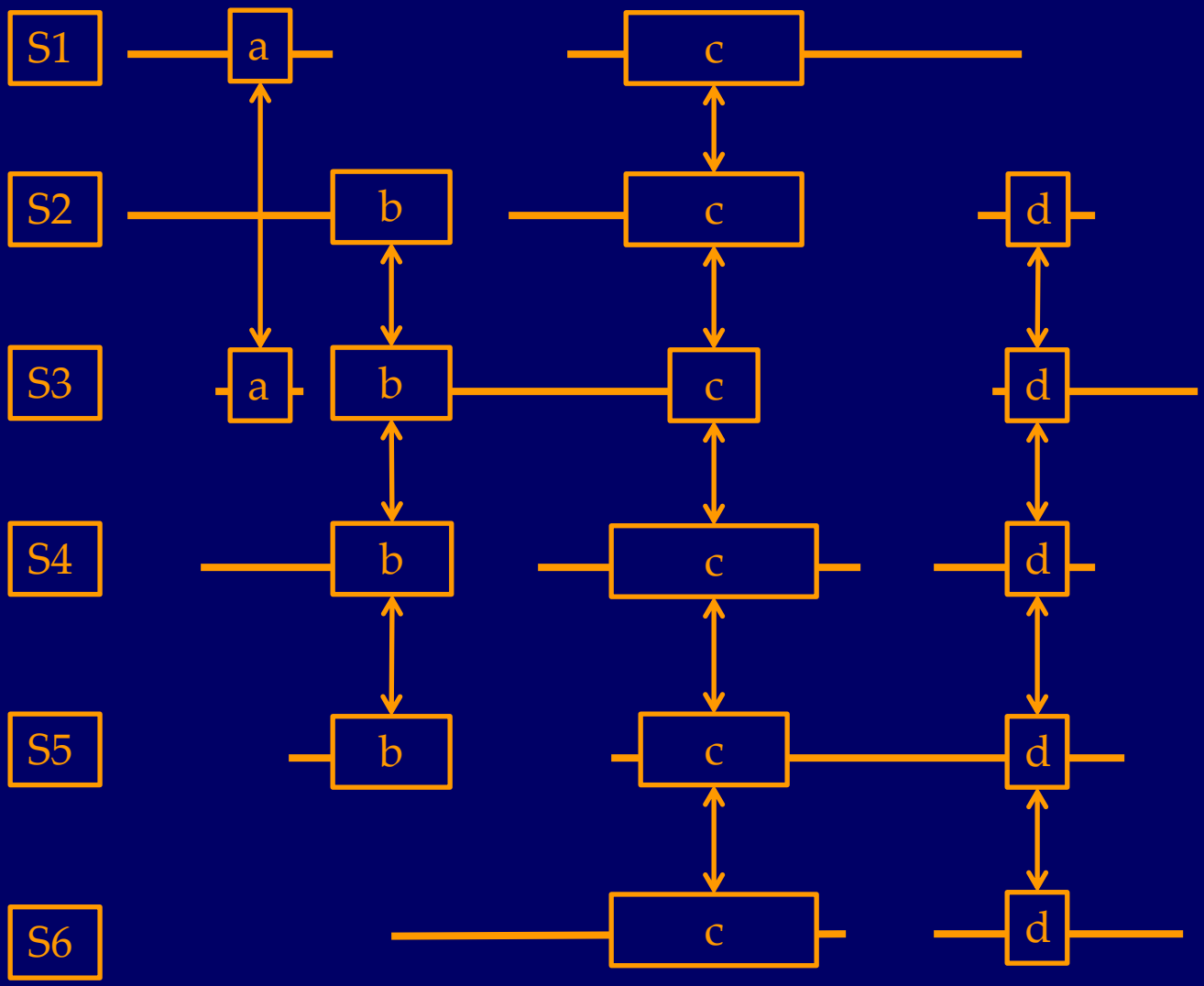

S1
S2
S3
S4
S5
S6
a
b
c
d
a
b
c
d
b
c
d
b
c
a
d
c
d

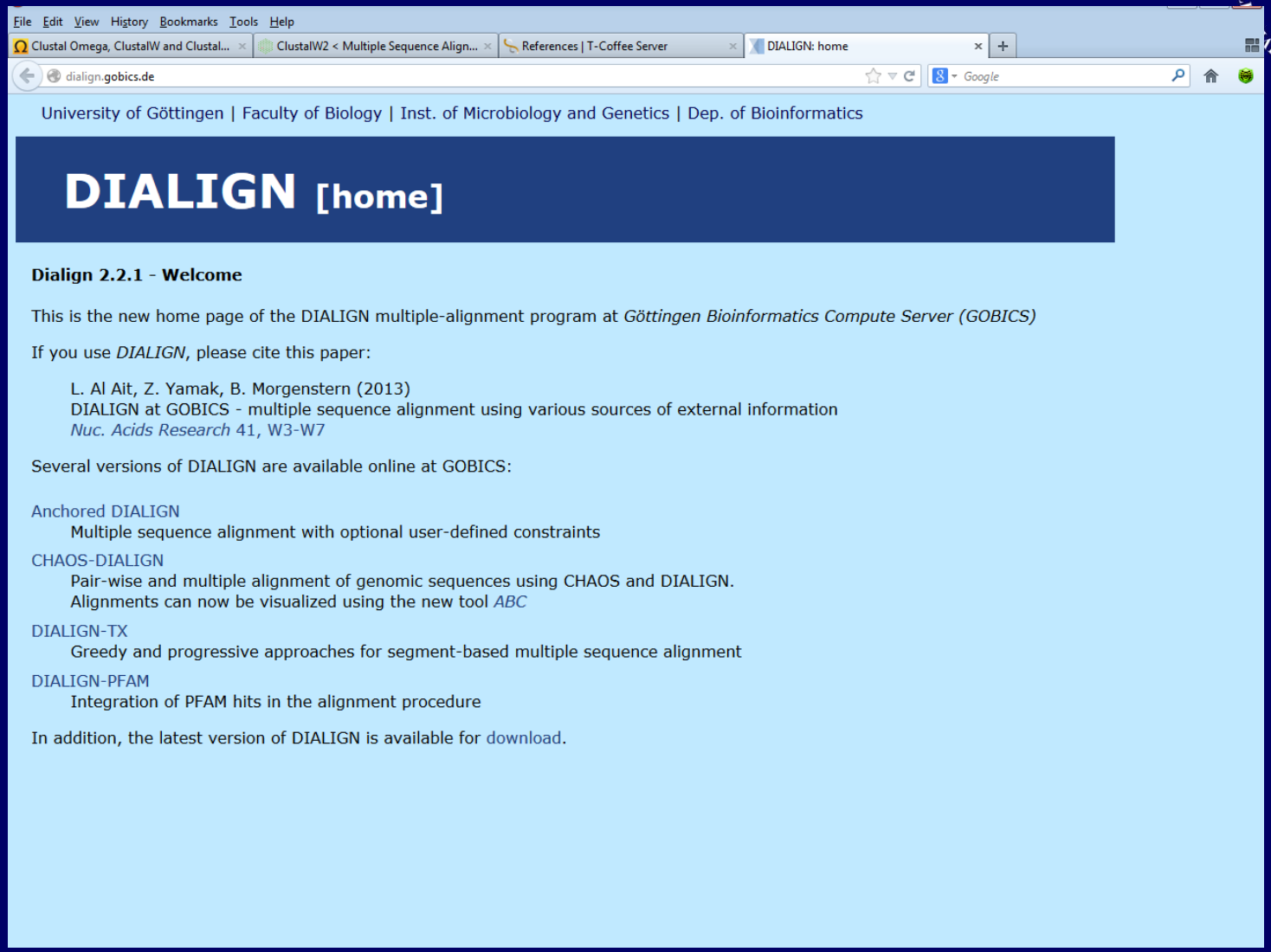

S1
S2
S3
S4
S5
S6
a
b
c
d
a
b
c
d
b
c
d
b
c
a
d
c
d


S1
S2
S3
S4
S5
S6
a
b
c
d
a
b
c
d
b
c
d
b
c
a
d
c
d


S1
S2
S3
S4
S5
S6
a
b
c
d


S1
S2
S3
S4
S5
S6
a
b
c
d


S1
S2
S3
S4
S5
S6
a
b
c
d
a
b
c
d
b
c
d
b
c
d
c
d


File Edit View History Bookmarks Tools Help
Clustal Omega, ClustalW and Clustal...
ClustalW2 < Multiple Sequence Align...
References | T-Coffee Server
DIALIGN: home
dialign.gobics.de
Google
University of Göttingen | Faculty of Biology | Inst. of Microbiology and Genetics | Dep. of Bioinformatics
DIALIGN [home]
Dialign 2.2.1 - Welcome
This is the new home page of the DIALIGN multiple-alignment program at Göttingen Bioinformatics Compute Server (GOBICS)
If you use DIALIGN, please cite this paper:
L. Al Ait, Z. Yamak, B. Morgenstern (2013)
DIALIGN at GOBICS - multiple sequence alignment using various sources of external information
Nuc. Acids Research 41, W3-W7
Several versions of DIALIGN are available online at GOBICS:
Anchored DIALIGN
Multiple sequence alignment with optional user-defined constraints
CHAOS-DIALIGN
Pair-wise and multiple alignment of genomic sequences using CHAOS and DIALIGN.
Alignments can now be visualized using the new tool ABC
DIALIGN-TX
Greedy and progressive approaches for segment-based multiple sequence alignment
DIALIGN-PFAM
Integration of PFAM hits in the alignment procedure
In addition, the latest version of DIALIGN is available for download.

Szünet

Iteratív eljárások

- A progresszív eljárásban az egyszer már elfogadott illesztés nem módosul
- Ha egy hiba bekerül egyszer, az benn is marad
- Esetleg még további hibákat okoz
- Az itaratív módszerek felülvizsgálják a már illesztett részeket is
- Így javítják a végeredményt

A MUSCLE eljárás

- Honlap: http://www.drive5.com/muscle/
- Web szolgáltatás az EBI felületen keresztül
- Letölthető Linux és Windows változatban is
- Bőséges dokumentáció elérhető a web-en
- A program szöveges felületen keresztül használható

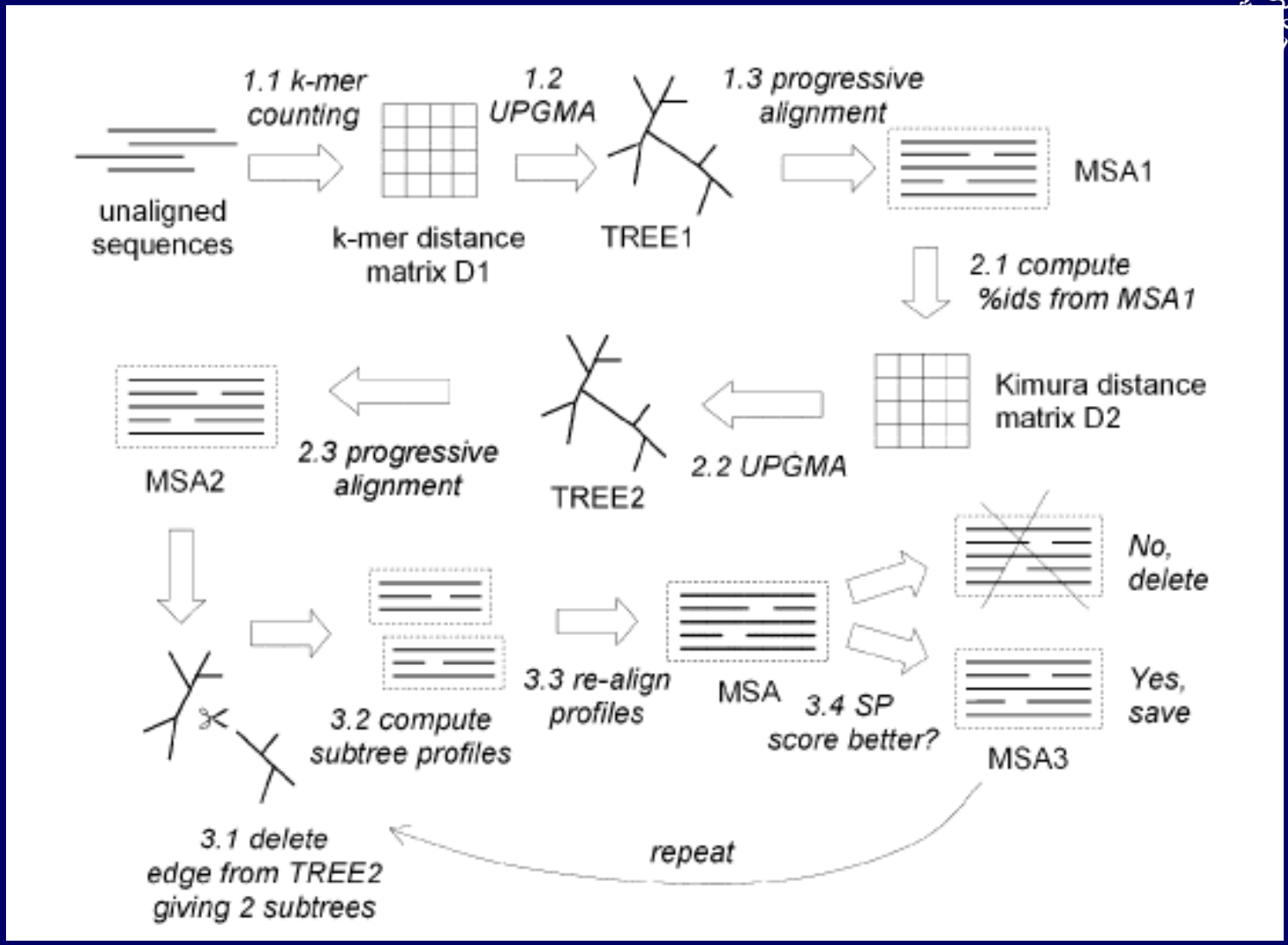
Az algoritmus

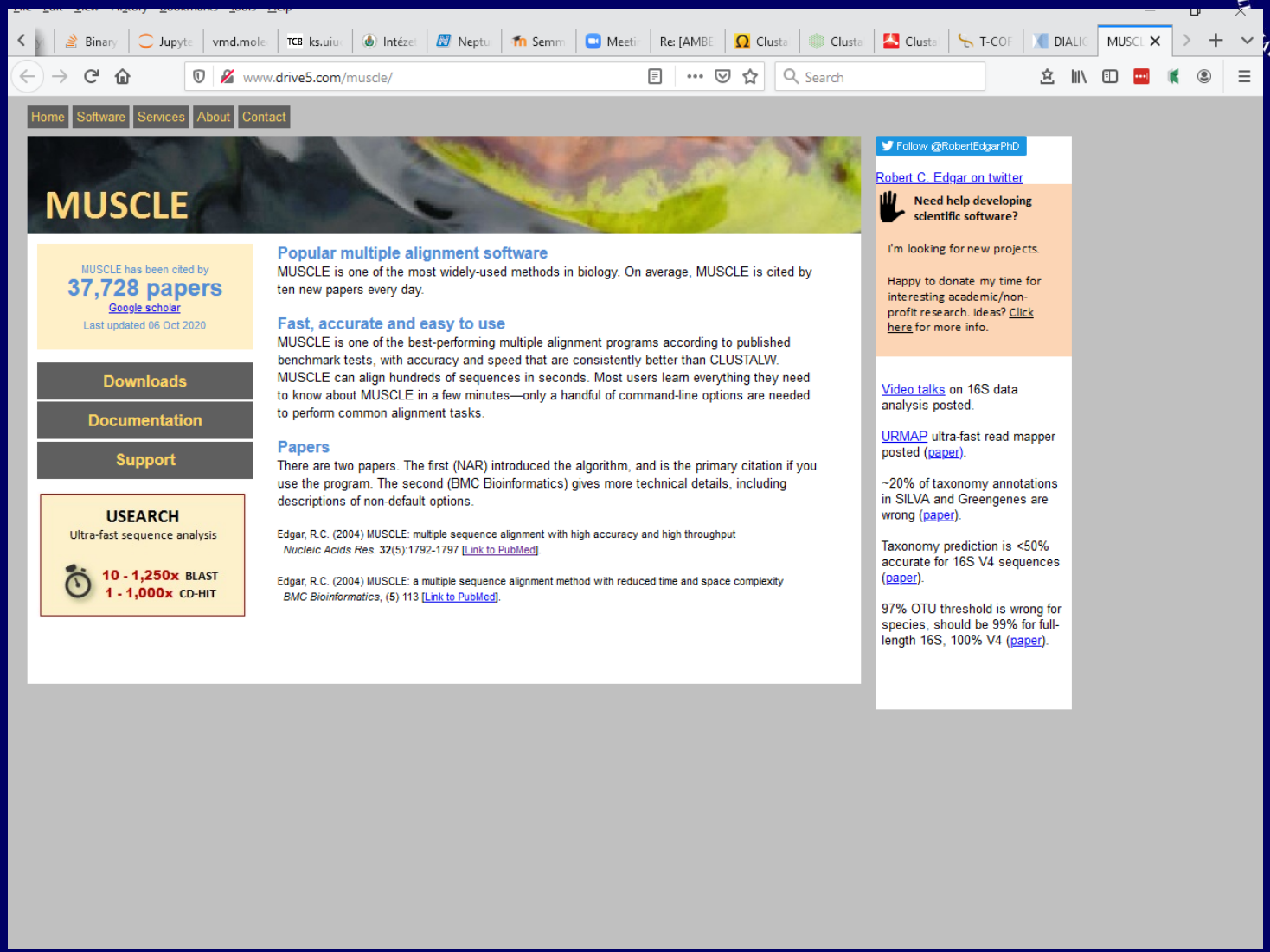
A folyamat 3 fázisban működik:

- Vázlatos illsztés:
  - K-mer szavak alapján számolt távolság-mátrix
  - Vázlatos származástani fa
  - Progresszív illesztés
- Finomított illesztés:
  - Az illesztés alapján számolt távolság-mátrix
  - Finomított származástani fa
  - Finomított progresszív illesztés

Iteratív finomítás

- A fa egy ágát eltöröljük
- A két részt külön használva új illesztést készítünk
- Kombináljuk a két illesztést
- Ha javult az illesztés, megtartjuk az eredményt és új ciklust kezdünk
- Addig csináljuk, amíg nem javul tovább


1.1 k-mer counting
1.2 UPGMA
1.3 progressive alignment
unaligned sequences
k-mer distance matrix D1
TREE1
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
3.1 delete edge from TREE2 giving 2 subtrees
3.2 compute subtree profiles
3.3 re-align profiles
MSA
3.4 SP score better?
No, delete
Yes, save
MSA3
repeat


www.drive5.com/muscle/
Search
Home
Software
Services
About
Contact
MUSCLE
MUSCLE has been cited by
37,728 papers
Google scholar
Last updated 06 Oct 2020
Downloads
Documentation
Support
USEARCH
Ultra-fast sequence analysis
10 - 1,250x BLAST
1 - 1,000x CD-HIT
Popular multiple alignment software
MUSCLE is one of the most widely-used methods in biology. On average, MUSCLE is cited by ten new papers every day.
Fast, accurate and easy to use
MUSCLE is one of the best-performing multiple alignment programs according to published benchmark tests, with accuracy and speed that are consistently better than CLUSTALW. MUSCLE can align hundreds of sequences in seconds. Most users learn everything they need to know about MUSCLE in a few minutes—only a handful of command-line options are needed to perform common alignment tasks.
Papers
There are two papers. The first (NAR) introduced the algorithm, and is the primary citation if you use the program. The second (BMC Bioinformatics) gives more technical details, including descriptions of non-default options.
Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput Nucleic Acids Res. 32(5):1792-1797 [Link to PubMed].
Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity BMC Bioinformatics, (5) 113 [Link to PubMed].
Follow @RobertEdgarPhD
Robert C. Edgar on twitter
Need help developing scientific software?
I'm looking for new projects.
Happy to donate my time for interesting academic/non-profit research. Ideas? Click here for more info.
Video talks on 16S data analysis posted.
URMAP ultra-fast read mapper posted (paper).
~20% of taxonomy annotations in SILVA and Greengenes are wrong (paper).
Taxonomy prediction is <50% accurate for 16S V4 sequences (paper).
97% OTU threshold is wrong for species, should be 99% for full-length 16S, 100% V4 (paper).

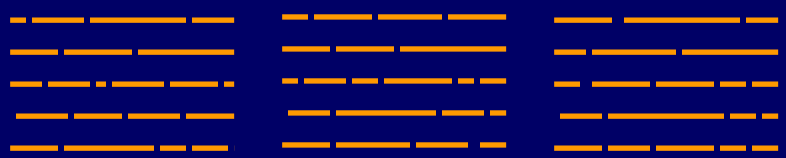
További módszerek

- Iterációs eljárások:
  - Genetikus algoritmus
  - Szimulált dermedés (simulated annealing)
- Az egesz illesztés változik, semmi sem rögzített
- A globális megoldás a cél
- Nagy számítási kapacitást igényel

Genetikus algoritmus

"biológiai analógia"

- Létrehozunk véletlenszrű változatokat - szülők
- Ezeket egymással kombináljuk - utódok
- További véletlenszrű változtatások – mutáció
- Kiválasztjuk a legjobbakat – szelekció
- Ezek lesznek a szülők a kovetkező ciklusban



„szülők”

„szülők”

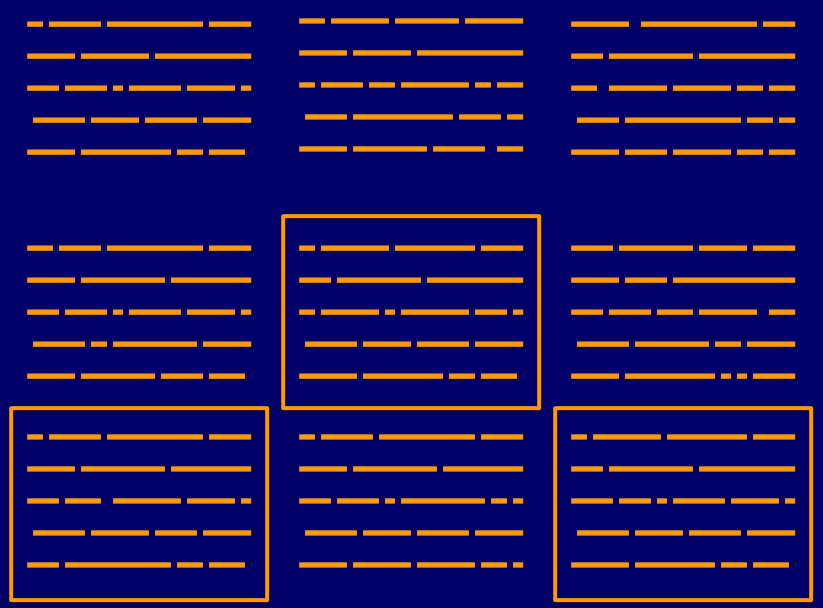
„szülők”

„szülők”

„utódok”

„szülők”

„utódok”



„szülők”

„utódok”

kiválasztás

„szülők”

„utódok”

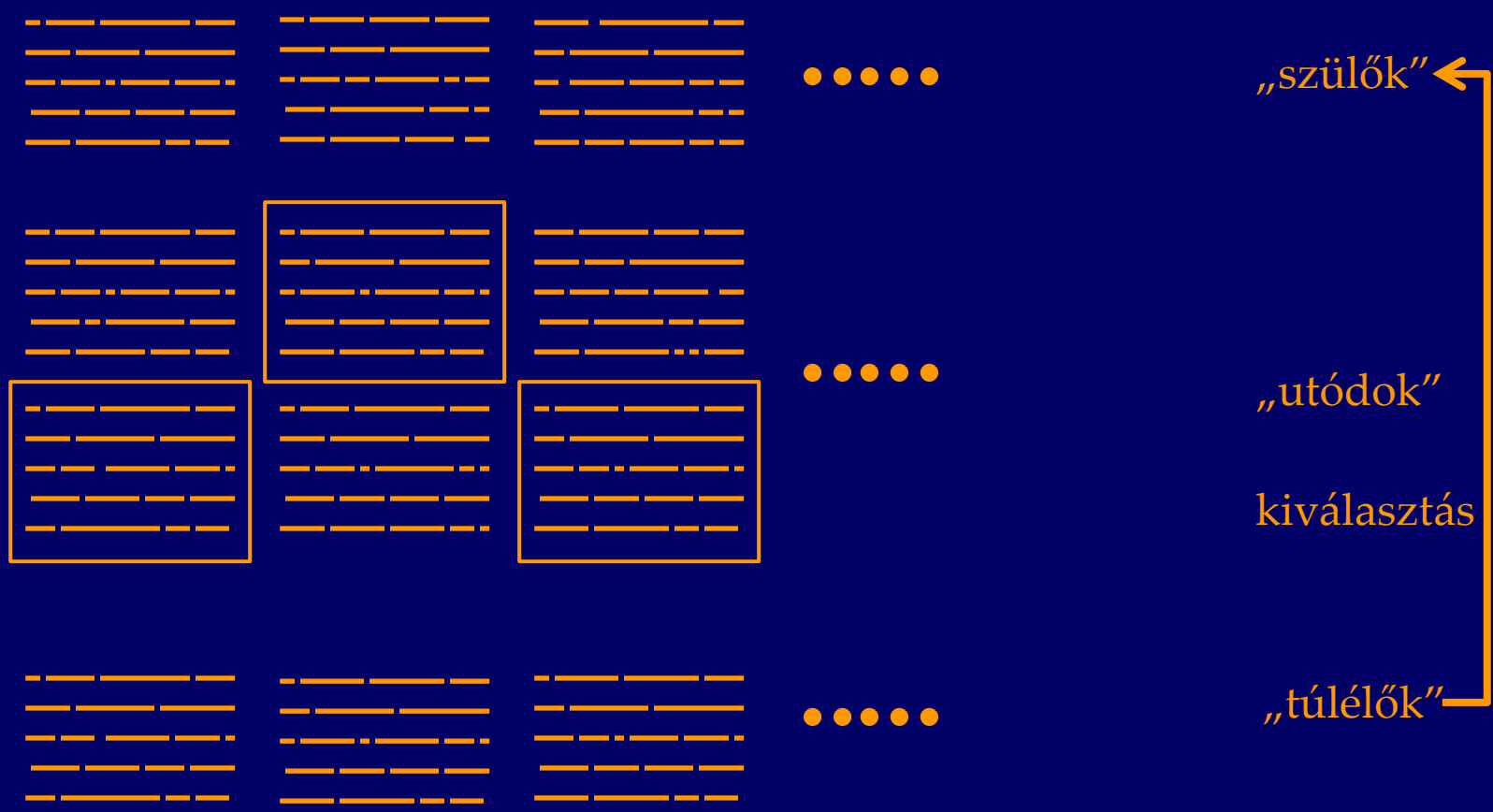
kiválasztás

„szülők”

„utódok”

kiválasztás

„túlélők”


„szülők”
„utódok”
kiválasztás
„túlélők”

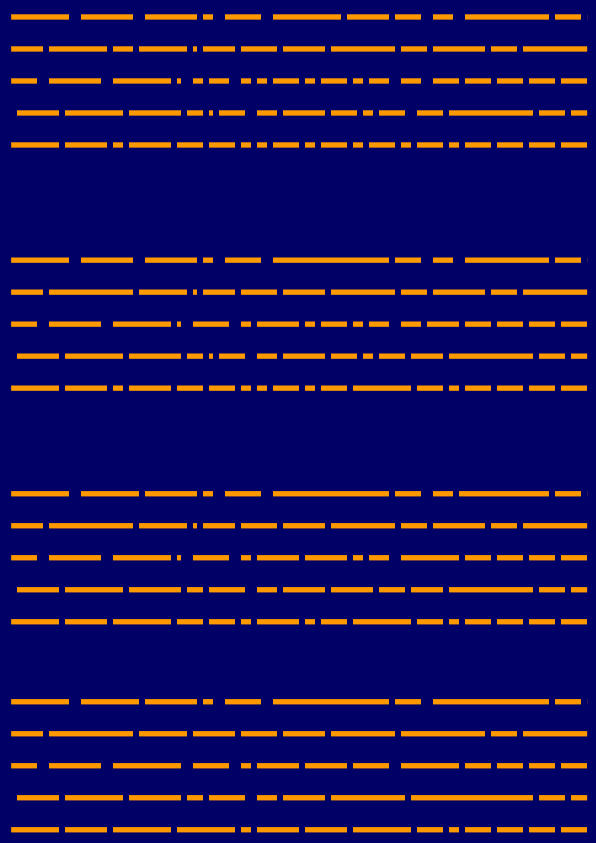
Szimulált dermedés

"fizikai analógia"

- "felmelegítjük" a rendszert
- Hagyuk "kihűlni"
- A rendszer "megdermed"
- Beáll a fázisátmenet – "kikristályosodik"
- Megint "felmelegítjük", de most nem annyira
- Stb…

Többszörös illesztés esetén

- A szekvenciákat telerakjuk betoldással
- Kis adagokban elvesszük a felesleges betoldásokat
- Ha az illesztés már nem javul tovább, visszateszünk egy adag betoldást. Stb..



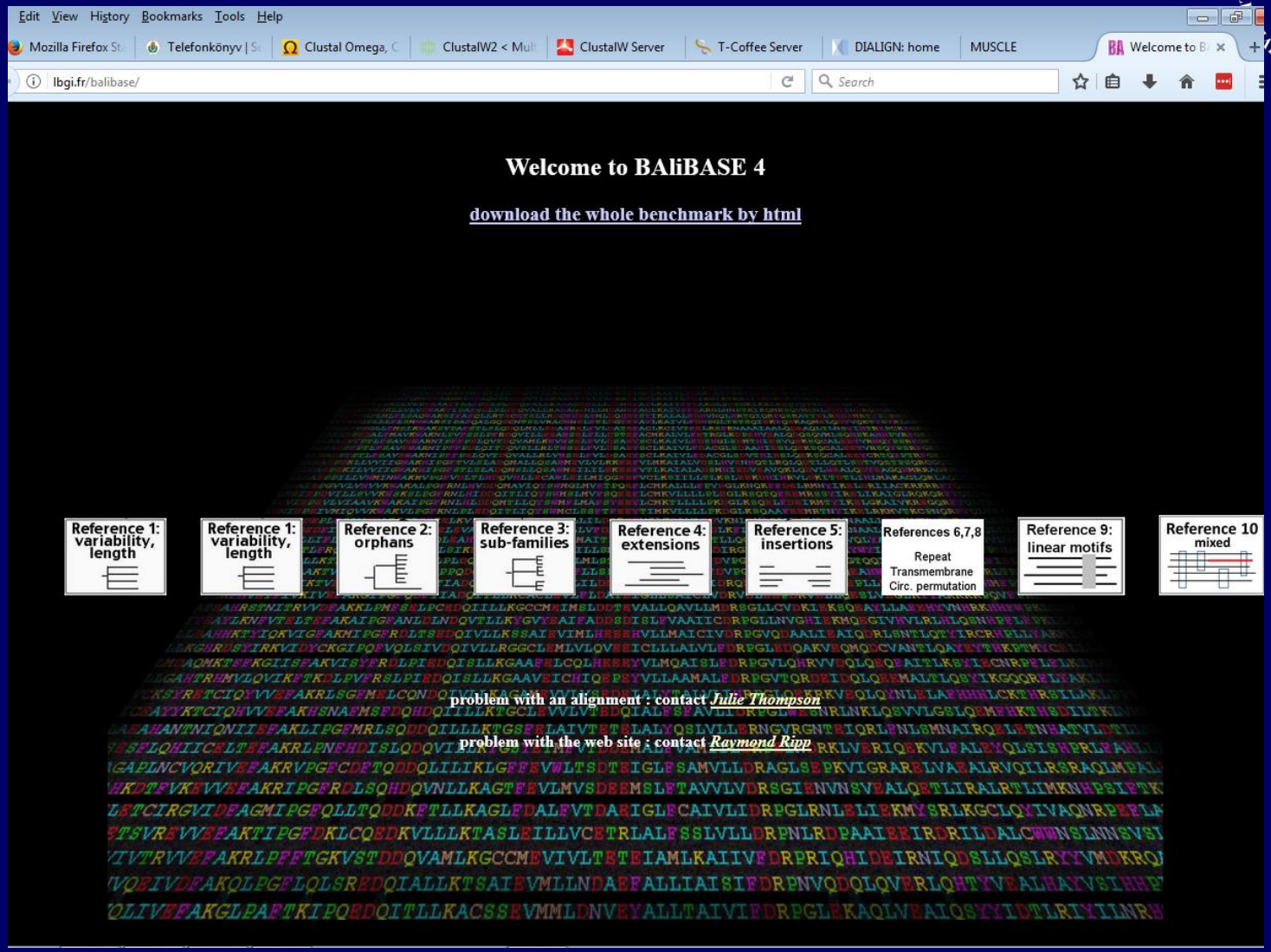
Ellenőrző adatbázisok

- Sok lehetőség – zavaróan sok megoldás
- Melyik a „legjobb”?
- Minden szerző saját teszteket közöl
- Szükség van egy közös összehasonlítási alapra
- Több ilyen is van...

A BALiBASE adatbázis

- Kifejezettem többszörös illesztések kibróbálására és minősítésére készült
- 10 kategóriában ad meg előre elkészített referencia illsztéseket
- A legtöbbhöz elérhető a szerkezet is
- Az illeszések kézzel készültek
- Az adatok szabadon letölthetők:

  http://lbgi.fr/balibase/


Edit View History Bookmarks Tools Help
Mozilla Firefox St
Telefonkönyv | Se
Clustal Omega, C
ClustalW2 < Mult
ClustalW Server
T-Coffee Server
DIALIGN: home
MUSCLE
Welcome to B
lbgi.fr/balibase/
Search
Welcome to BAliBASE 4
download the whole benchmark by html
Reference 1: variability, length
Reference 1: variability, length
Reference 2: orphans
Reference 3: sub-families
Reference 4: extensions
Reference 5: insertions
References 6,7,8
Repeat
Transmembrane
Circ. permutation
Reference 9: linear motifs
Reference 10 mixed
problem with an alignment : contact Julie Thompson
problem with the web site : contact Raymond Ripp

Egyenlő evolúciós távolságú szekvenciák

Reference 1

short, <25% identity

1aboA	SH3
1idy	myb dna-binding domain
1r69	repressor
1tvxA	pertussis toxin
1ubi	ubiquitin
1wit	twitchin
2trx	thioredoxin

short, 20%=40% identity

1aab	high mobility group protein
1fjlA	homeodomain
1hfh	factor h
1hpi	high-potential iron-sulfur protein
1csy	SH2
1pfc	immunoglobulin PFc fragment
1tgxA	cardiotoxin
1ycc	cytochrome e
3cyr	cytochrome c3
451c	cytochrome c

short, >35% identity

1aho	toxin II
1csp	cold shock protein
1dox	ferredoxin [2fe-2s]
1fkj	immunophilin
1fmb	hiv-1 protease
1krn	serine protease
1plc	plastocyanin
2fxb	ferredoxin
2mhr	hemerythrin
9rnt	ribonuclease

medium, <25% identity

1bbt3	foot-and-mouth disease virus
1sbp	sulfate binding protein
1havA	hepatitis proteinase
1uky	uridylate kinase
2hsdA	3 Alpha, 20 beta-hydroxysteroid dehydrogenase
2pia	phtalate reductase
3grs	glutathione reductase
kinase	protein kinase

medium, 20-40% identity

1ad2	ribosomal protein l1
1aym3	rhinovirus 16 coat protein
1gdoA	glucosamine 6-phosphate synthase
1ldg	lactate dehydrogenase
1mrj	alpha tricosanthin
1pgtA	glutathione
1pii	anthranilate isomerase
1ton	tonin
2cba	anhydrase

medium, >35% identity

1amk	triose phosphate isomerase
1ar5A	superoxide dismutase
1ezm	elastase
1led	lectin
1ppn	papain
1pysA	phenylalanyl-trna synthetase
1thm	serine protease
1tis	thymidylate sythase
1zin	adenylate kinase
5ptp	serine protease

Fehérjecsaládok egy kilógó, távoli taggal

short

1aboA	SH3
1csy	SH2
1idy	myb dna-binding domain
1r69	repressor
1tgxA	cardiotoxin
1tvxA	pertussis toxin
1ubi	ubiquitin
1wit	twitchin
2trx	thioredoxin

medium

1sbp	sulfate binding protein
1havA	hepatitis proteinase
1uky	uridylate kinase
2hsdA	3 Alpha, 20 beta-hydroxysteroid dehydrogenase
2pia	phtalate reductase
3grs	glutathione reductase
kinase	protein kinase

long

1ajsA	aminotransferase
1cpt	cytochrome p450
1lvl	dihydrolipoamide dehydrogenase
1pamA	cyclodextrin
1ped	alcohol dehydrogenase
2myr	myrosinase
4enl	enolase

Alcsaládok gyenge homológiával

short

1idy	myb dna-binding domain
1r69	repressor
1ubi	ubiquitin
1wit	twitchin

medium

1uky	uridylate kinase
2pia	phtalate reductase
kinase	protein kinase

long

1ajsA	aminotransferase
1pamA	cyclodextrin
1ped	alcohol dehydrogenase
2myr	myrosinase
4enl	enolase

A terminálisokon túlnyúló végek

Reference 4 - Extensions

1ckaA	SH3
1csp	major cold shock protein
1dynA	pleckstrin homology domain
1lkl	SH2
1mfa	immunoglobulin fab fragment
1pfc	immunoglobulin PFc fragment
1pysA	glycyl-tRNA synthetase
1vln	concanavalin a
1ycc	cytochrome e
2abk	endonuclease iii
kinase1	protein kinase
kinase2	protein kinase

Hosszú betoldások középen

Reference 5 - Insertions

1eft	eftu
1ivy	human protective protein
1pysA	glycyl-tRNA synthetase
1qpg	3-Phosphoglycerate kinase
1thm1	thermitase
1thm2	thermitase
2cba	carbonic anhydrase ii
S51	nuclear receptor
S52	nuclear receptor
kinase1	protein kinase
kinase2	protein kinase
kinase3	protein kinase

Repetitív elemeket tartalmazó szekvenciák

Reference 6 - Repeats

sh3	SH3
zf	Zinc Finger
apo	Apolipoprotein
sushi	Sushi
myb	Myb dna-binding domain
ank	Ankyrin
kringle	KRINGLE
dead	DEAD/DEAH box helicase
trk	TrkA potassium uptake protein
ktn+trk	KTN NAD-binding domain, plus TrkA potassium uptake protein
faa	Fumarylacetoacetate hydrolase
lrr	Leucine-rich repeat
ion	ION

Transzmembrán szekvenciák

Reference 7 - Transmembrane Sequences

ion	ION
acr	ACR
dtd	DTD
ptga	PTGA
Nat	Nat
photo	PHOTO
msl	MSL
7tm	7 TM

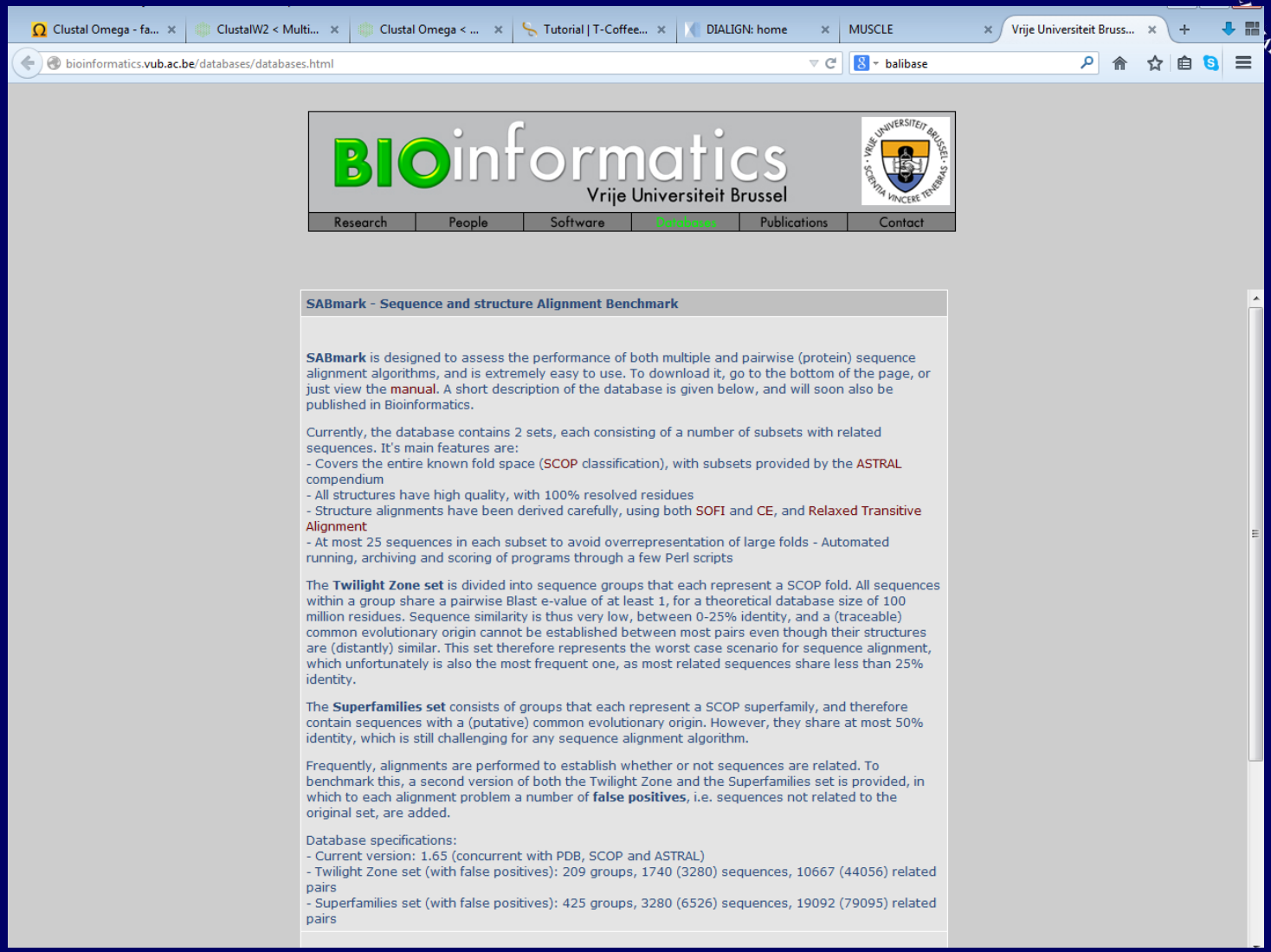
Példák cirkuláris permutációra

Reference 8 - Circular Permutations

sh3/sh2	SH3 / SH2
gsh	glutathione synthetase
lectin	Lectin1 / Lectin2
ptga	PTSEIIB / PTGA
cellulase	Cellulase / CBD_1

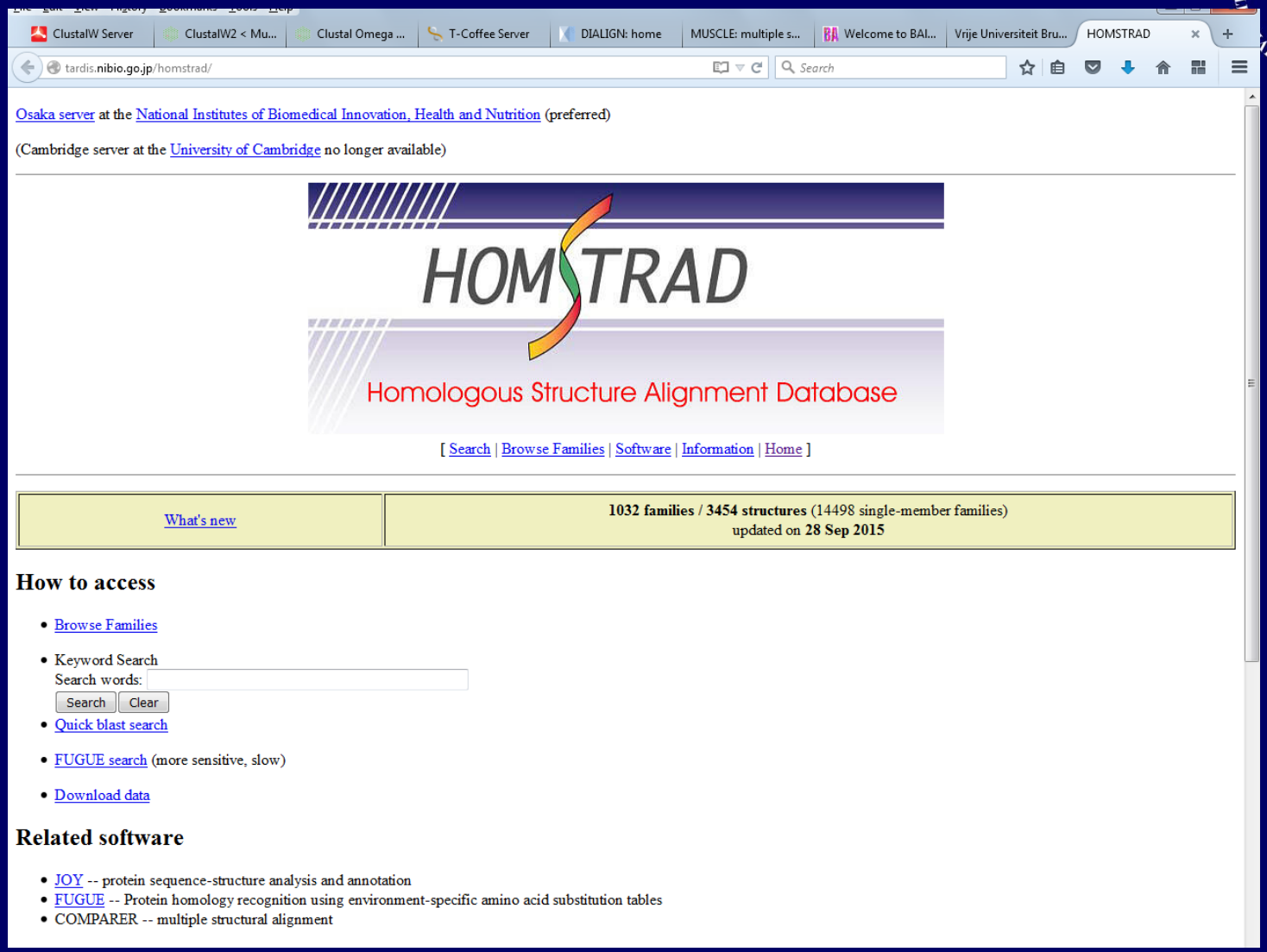
SABmark adatbázis

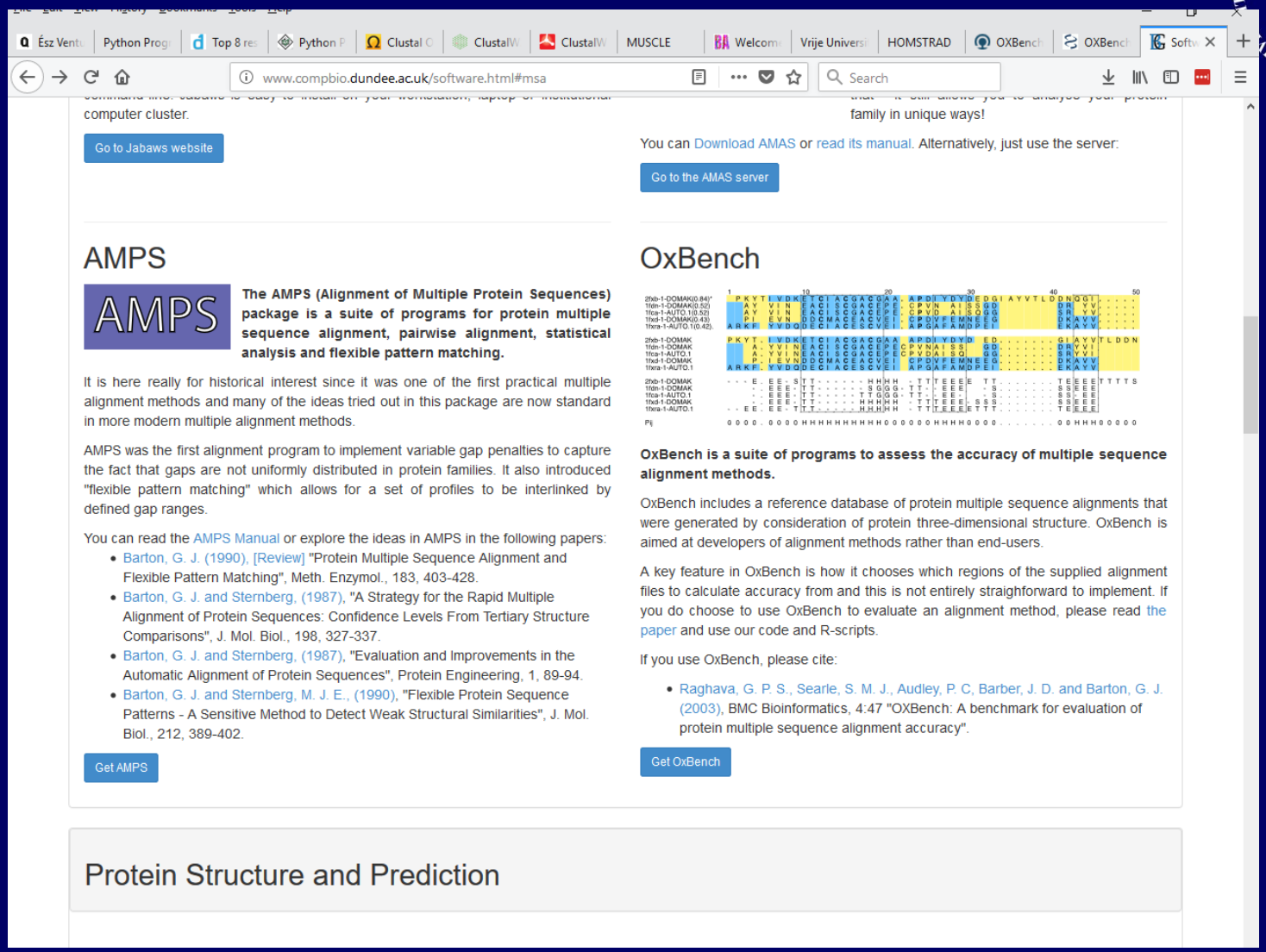
- A teljes protein univerzumot lefedi
- Csak 25 szekvenciát tartalmaz családonként
- "alkonyzóna"-gyüjtemény – nagyon alacsony fokú rokonság
- "nagycsalád"-gyüjtemény – kicsit nagyobb fokú rokonság
- A gyüjtemények szándékosan tartalmaznak nem odaillő szekvenciákat
- http://bioinformatics.vub.ac.be/databases/databases.html


Clustal Omega - fa...
ClustalW2 < Multi...
Clustal Omega < ...
Tutorial | T-Coffee...
DIALIGN: home
MUSCLE
Vrije Universiteit Bruss...
bioinformatics.vub.ac.be/databases/databases.html
balibase
BIOinformatics
Vrije Universiteit Brussel
Research
People
Software
Databases
Publications
Contact
SABmark - Sequence and structure Alignment Benchmark
SABmark is designed to assess the performance of both multiple and pairwise (protein) sequence alignment algorithms, and is extremely easy to use. To download it, go to the bottom of the page, or just view the manual. A short description of the database is given below, and will soon also be published in Bioinformatics.
Currently, the database contains 2 sets, each consisting of a number of subsets with related sequences. It's main features are:
- Covers the entire known fold space (SCOP classification), with subsets provided by the ASTRAL compendium
- All structures have high quality, with 100% resolved residues
- Structure alignments have been derived carefully, using both SOFI and CE, and Relaxed Transitive Alignment
- At most 25 sequences in each subset to avoid overrepresentation of large folds - Automated running, archiving and scoring of programs through a few Perl scripts
The Twilight Zone set is divided into sequence groups that each represent a SCOP fold. All sequences within a group share a pairwise Blast e-value of at least 1, for a theoretical database size of 100 million residues. Sequence similarity is thus very low, between 0-25% identity, and a (traceable) common evolutionary origin cannot be established between most pairs even though their structures are (distantly) similar. This set therefore represents the worst case scenario for sequence alignment, which unfortunately is also the most frequent one, as most related sequences share less than 25% identity.
The Superfamilies set consists of groups that each represent a SCOP superfamily, and therefore contain sequences with a (putative) common evolutionary origin. However, they share at most 50% identity, which is still challenging for any sequence alignment algorithm.
Frequently, alignments are performed to establish whether or not sequences are related. To benchmark this, a second version of both the Twilight Zone and the Superfamilies set is provided, in which to each alignment problem a number of false positives, i.e. sequences not related to the original set, are added.
Database specifications:
- Current version: 1.65 (concurrent with PDB, SCOP and ASTRAL)
- Twilight Zone set (with false positives): 209 groups, 1740 (3280) sequences, 10667 (44056) related pairs
- Superfamilies set (with false positives): 425 groups, 3280 (6526) sequences, 19092 (79095) related pairs

További adatbázisok

- HOMSTAD: http://mizuguchilab.org/homstrad/
- OXBENCH: http://www.compbio.dundee.ac.uk/


ClustalW Server
ClustalW2 < Mu...
Clustal Omega ...
T-Coffee Server
DIALIGN: home
MUSCLE: multiple s...
Welcome to BAI...
Vrije Universiteit Bru...
HOMSTRAD
tardis.nibio.go.jp/homstrad/
Search
Osaka server at the National Institutes of Biomedical Innovation, Health and Nutrition (preferred)
(Cambridge server at the University of Cambridge no longer available)
HOMSTRAD
Homologous Structure Alignment Database
[Search | Browse Families | Software | Information | Home]
What's new
1032 families / 3454 structures (14498 single-member families)
updated on 28 Sep 2015
How to access
Browse Families
Keyword Search
Search words:
Search
Clear
Quick blast search
FUGUE search (more sensitive, slow)
Download data
Related software
JOY -- protein sequence-structure analysis and annotation
FUGUE -- Protein homology recognition using environment-specific amino acid substitution tables
COMPARER -- multiple structural alignment



computer cluster.

Go to Jabaws website

family in unique ways!

You can Download AMAS or read its manual. Alternatively, just use the server:

Go to the AMAS server

AMPS

The AMPS (Alignment of Multiple Protein Sequences) package is a suite of programs for protein multiple sequence alignment, pairwise alignment, statistical analysis and flexible pattern matching.

It is here really for historical interest since it was one of the first practical multiple alignment methods and many of the ideas tried out in this package are now standard in more modern multiple alignment methods.

AMPS was the first alignment program to implement variable gap penalties to capture the fact that gaps are not uniformly distributed in protein families. It also introduced "flexible pattern matching" which allows for a set of profiles to be interlinked by defined gap ranges.

You can read the AMPS Manual or explore the ideas in AMPS in the following papers:

- Barton, G. J. (1990), [Review] "Protein Multiple Sequence Alignment and Flexible Pattern Matching", Meth. Enzymol., 183, 403-428.
- Barton, G. J. and Sternberg, (1987), "A Strategy for the Rapid Multiple Alignment of Protein Sequences: Confidence Levels From Tertiary Structure Comparisons", J. Mol. Biol., 198, 327-337.
- Barton, G. J. and Sternberg, (1987), "Evaluation and Improvements in the Automatic Alignment of Protein Sequences", Protein Engineering, 1, 89-94.
- Barton, G. J. and Sternberg, M. J. E., (1990), "Flexible Protein Sequence Patterns - A Sensitive Method to Detect Weak Structural Similarities", J. Mol. Biol., 212, 389-402.

Get AMPS

OxBench

OxBench is a suite of programs to assess the accuracy of multiple sequence alignment methods.

OxBench includes a reference database of protein multiple sequence alignments that were generated by consideration of protein three-dimensional structure. OxBench is aimed at developers of alignment methods rather than end-users.

A key feature in OxBench is how it chooses which regions of the supplied alignment files to calculate accuracy from and this is not entirely straighforward to implement. If you do choose to use OxBench to evaluate an alignment method, please read the paper and use our code and R-scripts.

If you use OxBench, please cite:

- Raghava, G. P. S., Searle, S. M. J., Audley, P. C, Barber, J. D. and Barton, G. J. (2003), BMC Bioinformatics, 4:47 "OXBench: A benchmark for evaluation of protein multiple sequence alignment accuracy".

Get OxBench

Protein Structure and Prediction

Mit tanultunk ma?

- A többszörös szekvencia illesztés a leghatékonyabb bioinformatikai módszer
- A feladatnak nincs egzakt megoldása, annyira komplex
- Heurisztikus megoldások vannak: elfogadható megoldást kapunk elfogadható idő alatt
- Az eltérő eljárások más eredményt adnak

Feladat 4

- Próbálj ki több szervert a többszörös illesztési feladatok elvégzéséhez és válaszd ki a neked megfelelőt.