T-COFFEE

Journal club in bioinformatics
by *Tõnu Margus*

PROTOCOL

Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures

Jean-Francois Taly[1,2], Cedrik Magis[1,2], Giovanni Bussotti[1], Jia-Ming Chang[1], Paolo Di Tommaso[1], Ionas Erb[1], Jose Espinosa-Carrasco[1], Carsten Kemena[1] & Cedric Notredame[1]

T-Coffee **T**ree-based **C**onsistency **O**bjective **F**unction **f**or alignm**E**nt **E**valuation

MIKS MA SELLEST RÄÄGIN?

- MSA on väga laialdaselt kasutatav meetod
- Mitmejärjestuse joondamiseks on palju programme
- 10a jooksul on tehtud enam kui 900 versiooniuuendust
- Praegu võib T-COFFEE’st rääkida kui joonduse stuudiost
- Praktika kui TÕE kriteerium

MSA peamised rakendused

- Fülogeneetilise ajaloo rekonstrueerimine
  - enamus puudearvutus programme võtab aluseks joonduse. Joonduse ja puude rekonstrueerimise omavaheline suhe on vastastikune ja komplitseritud
- Profiilide hindamine
  - konserveeruvus/varieeruvus mustri (valgu või RNA) hindamine võimaldab leida kaugeid homolooge; lisada funktsionaalset annotatsiooni

MITMEJÄRJESTUSE JOONDUS (MSA)

- The purpose of an MSA is to explicitly declare the relation of homology between all the residues within a set of related sequences, thus making it possible to identify highly conserved positions, or positions whose variability may have a functional meaning (e.g., substrate specificity).

JOONDUSE PROBLEMAATIKA

The MSA problem sits at the crossroads of biology and computer science. Its biological side lies in the difficulty of defining a mathematical function (scoring scheme) able to estimate the biological correctness of an MSA.

- This scoring scheme should reflect the relation-ship between sequence, structure and function; it should also integrate the effect of their mutual dependencies onto sequence mutations.

JOONDUSE PROBLEMAATIKA

The MSA problem sits at the crossroads of biology and computer science. Its biological side lies in the difficulty of defining a mathematical function (scoring scheme) able to estimate the biological correctness of an MSA.

BLOSUM

- This scoring scheme should reflect the relationship between sequence, structure and function; it should also integrate the effect of their mutual dependencies onto sequence mutations.

LOG-ODD SCORING SCHEME

- Log-odd matrices such as the BLOSUM are effective for closely related sequences
- They become rapidly non-informative when dealing with distantly related sequences (less than 25% for proteins and 70% for nucleic acids)
- Limited accuracy AND difficult to optimize

JOONDUSE PROBLEMAATIKA

The computational challenge of MSA modeling: the impossibility to guarantee the calculation of an optimal alignment, except in the simplest cases.

- In computer science terms, MSA computation is said to be a nondeterministic polynomial time complete problem (NP-complete)
- Vajaminev mälu hulk ja arvutusaeg on eksponent-siaalses sõltuvuses järjestuste arvust (~14)

TÖÖTAVAD LAHENDUSED - HEURISTIKA

HEURISTIKA

- Probleemi “lahendamatu” loomuse tõttu on loodud palju erinevaid heuristilisi meetodeid ja neile põhinevaid programme / pakette
- Wikipedia (aprill 2011 seisuga) toob ära 38 erinevat programmi ja paketti

4 main heuristic methods

1. *Standard progressive alignment methods*: ClustalW or Kalign
2. *Iterative methods*: MUSCLE and MAFFT
3. *Consistency-based methods*: T-Coffee, ProbCons and PROMALS
4. *Template-based method*

Consistency-based methods

- These methods combine the progressive alignment with a different scoring scheme.
- Their goal is to optimize the consistency (i.e., the agreement) between the MSA and a collection of weighted constraints
- The main difference between T-Coffee and a regular progressive aligner is that, rather than using a BLOSUM matrix to estimate the score of matches when incorporating the sequences into the MSA, these scores are evaluated using the scoring scheme defined by the library.

old ClustalW

Progressive aligner

```
seqA     GARFIELD THE LAST FA-T CAT
seqB     GARFIELD THE FAST CA-T ---
seqC     GARFIELD THE VERY FAST CAT
seqD     --------- THE ---- FA-T CAT
```

Consistency-based aligner

T-Coffee builds LIBRARY

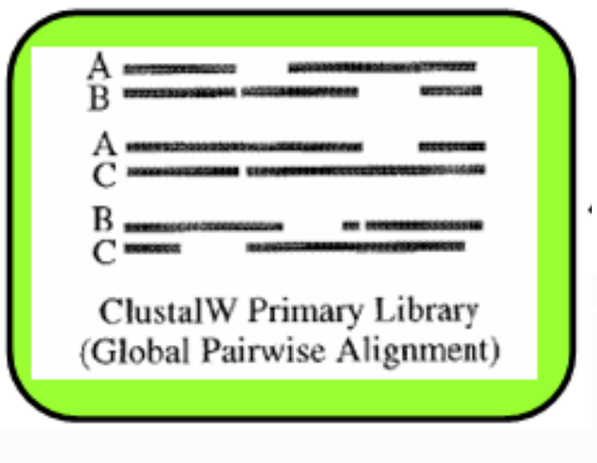
```
SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT ---                    Prim. Weight = 88

SeqA GARFIELD THE LAST FA-T CAT
SeqC GARFIELD THE VERY FAST CAT                   Prim. Weight = 77

SeqA GARFIELD THE LAST FAT CAT
SeqD ------- THE ---- FAT CAT                     Prim. Weight =100

SeqB GARFIELD THE ---- FAST CAT
SeqC GARFIELD THE VERY FAST CAT                   Prim Weight = 100

SeqB GARFIELD THE FAST CAT
SeqD ------- THE FA-T CAT                         Prim. Weight = 100

SeqC GARFIELD THE VERY FAST CAT
SeqD ------- THE ---- FA-T CAT                    Prim. Weight = 100
```

```
SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT                        Weight = 88

SeqA GARFIELD THE LAST FAT CAT
SeqC GARFIELD THE VERY FAST CAT                   Weight = 77
SeqB GARFIELD THE      FAST CAT

SeqA GARFIELD THE LAST FAT CAT
SeqD          THE      FAT CAT                    Weight = 100
SeqB GARFIELD THE      FAST CAT
```

} Primary Library

Consistency-based aligner

T-Coffee builds LIBRARY



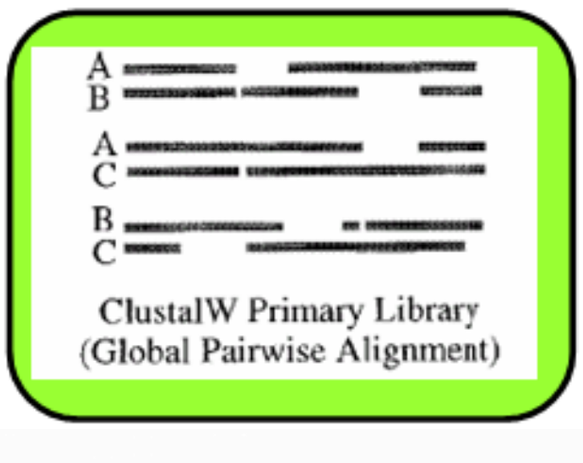
Global

```
SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT ---        Prim. Weight = 88

SeqA GARFIELD THE LAST FA-T CAT
SeqC GARFIELD THE VERY FAST CAT       Prim. Weight = 77

SeqA GARFIELD THE LAST FAT CAT
SeqD ------- THE ---- FAT CAT         Prim. Weight =100

SeqB GARFIELD THE ---- FAST CAT
SeqC GARFIELD THE VERY FAST CAT       Prim Weight = 100

SeqB GARFIELD THE FAST CAT
SeqD ------- THE FA-T CAT             Prim. Weight = 100

SeqC GARFIELD THE VERY FAST CAT
SeqD ------- THE ---- FA-T CAT        Prim. Weight = 100
```

Primary Library

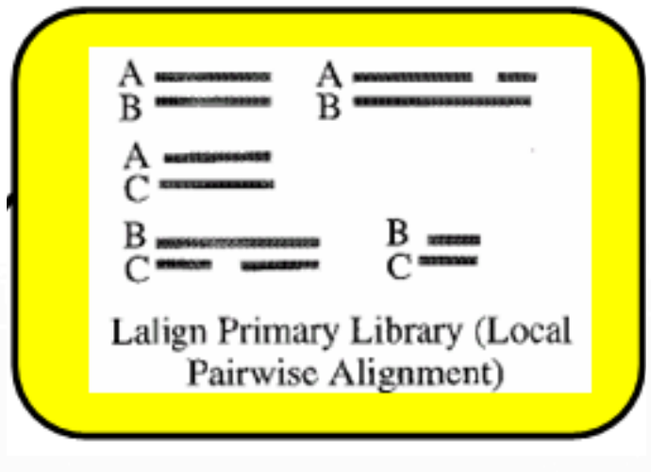
Consistency-based aligner

T-Coffee builds LIBRARY



Global

```
SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT ---        Prim. Weight = 88

SeqA GARFIELD THE LAST FA-T CAT
SeqC GARFIELD THE VERY FAST CAT       Prim. Weight = 77

SeqA GARFIELD THE LAST FAT CAT
SeqD ------- THE ---- FAT CAT         Prim. Weight =100

SeqB GARFIELD THE ---- FAST CAT
SeqC GARFIELD THE VERY FAST CAT       Prim Weight = 100

SeqB GARFIELD THE FAST CAT
SeqD ------- THE FA-T CAT             Prim. Weight = 100

SeqC GARFIELD THE VERY FAST CAT
SeqD ------- THE ---- FA-T CAT        Prim. Weight = 100
```

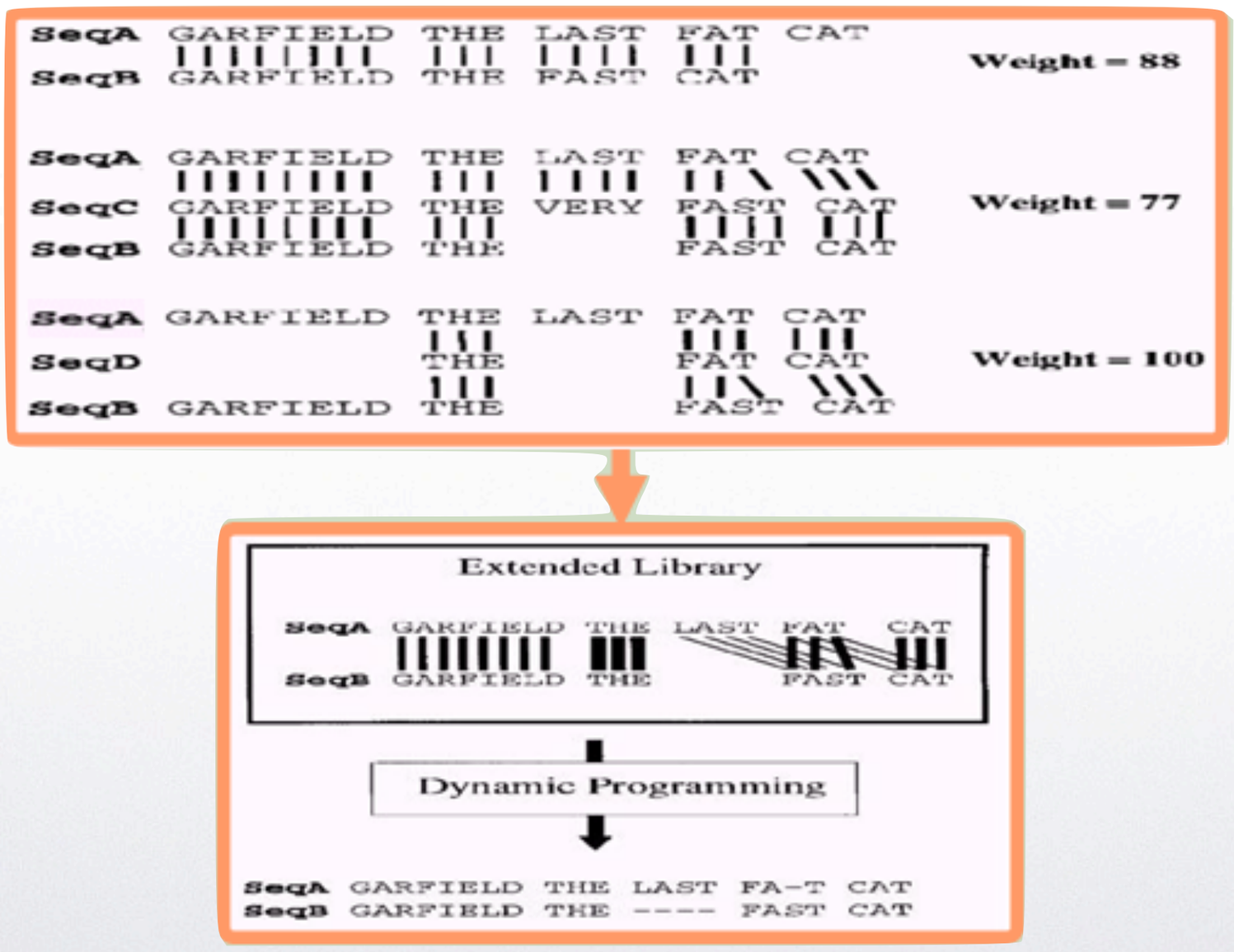


Local

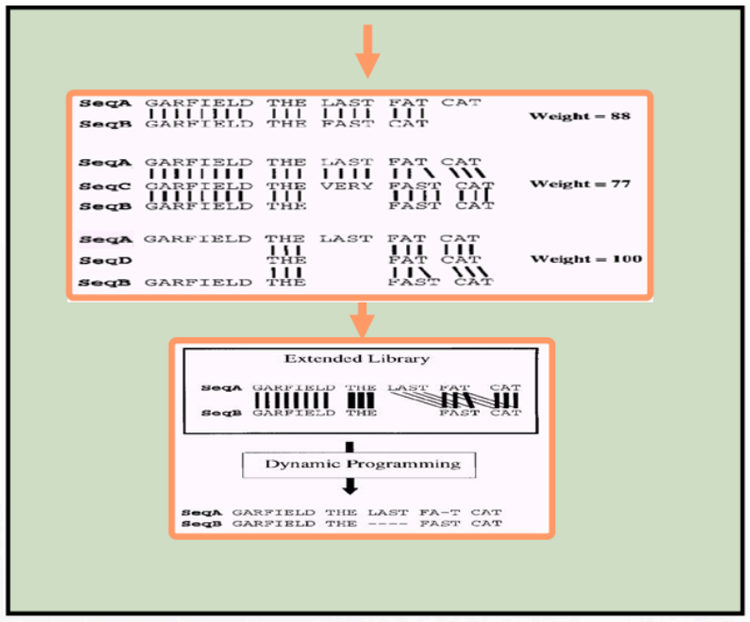
```
SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT            Weight = 88

SeqA GARFIELD THE LAST FAT CAT
SeqC GARFIELD THE VERY FAST CAT       Weight = 77
SeqB GARFIELD THE      FAST CAT

SeqA GARFIELD THE LAST FAT CAT
SeqD          THE      FAT CAT        Weight = 100
SeqB GARFIELD THE      FAST CAT
```

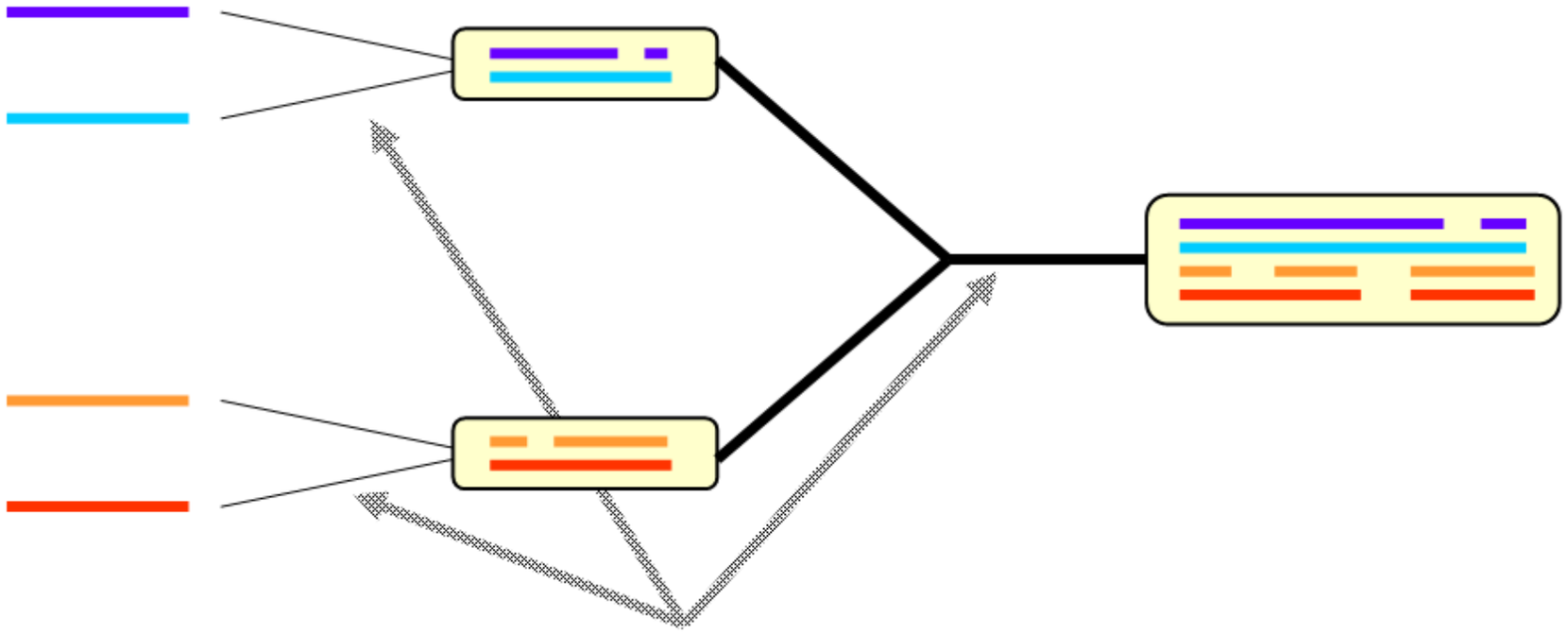
} Primary Library

T-Coffee constructs Extended LIBRARY

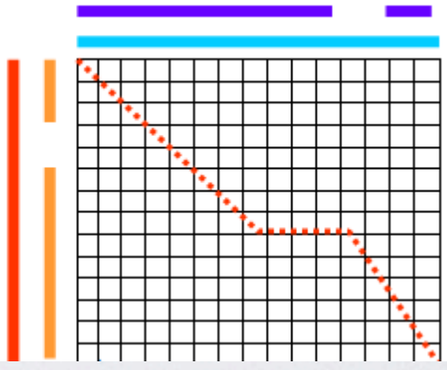


T-Coffee and Consistency



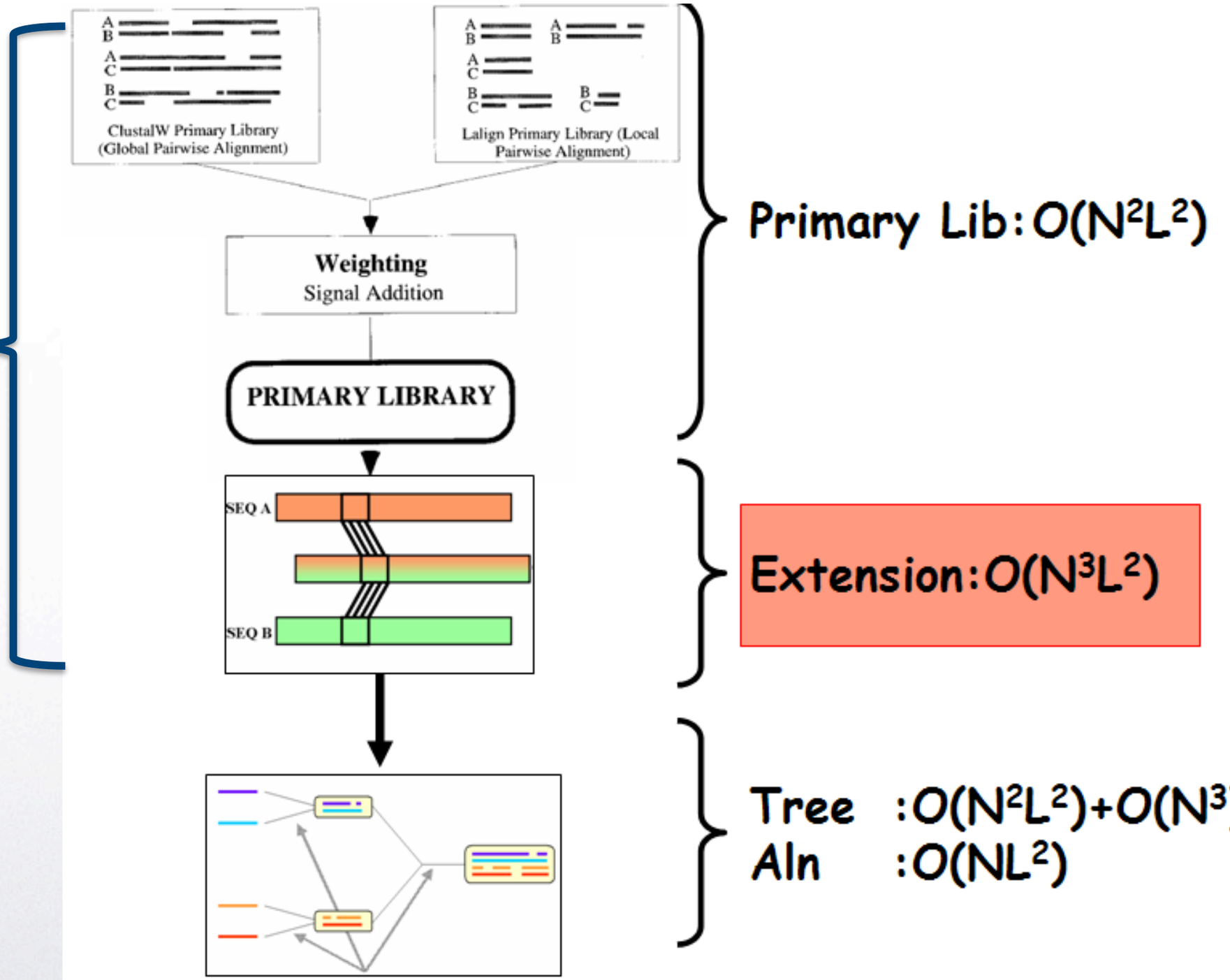


Dynamic Programming Using A Substitution Matrix

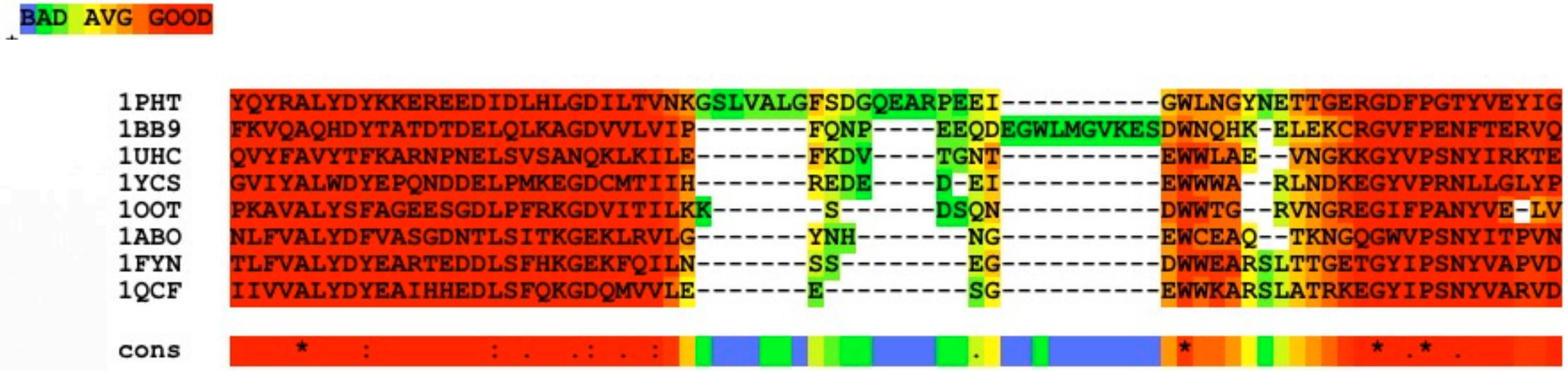


T-Coffee and Consistency

Väga oluline osa -
Programmi SÜDAmik



CORE INDEX vt. BOX 1



- The CORE index reflects the support given by the library to the alignment of every individual residue. Its value is normalized between 1 and 10.
- The CORE index is most informative when used to identify low-scoring portions within an MSA.
- It is also worth noting that the CORE index is not informative when aligning less than five sequences.

Praktiline osa

contents

- Istallation ~5 min
- Aligning protein sequences ~10 min
- RNA järjestuste joondamine ~1 min
- Promootor järjestuste joondamine ~1 min

Installing T-coffee

vt. artikkel lk. 1673 (5.)

- Supported OS
  - Mac OSX
  - UNIX/LINUX

- Software: http://tcoffee.org/Packages/Stable/Latest/

1. download T-coffee
2. installing (yes, yes, yes .....)
3. open new window - you are activating system variables needed by the program

Sample data

vt. artikkel lk. 1673 (5.)

You have two possibilities:

1. http://tcoffee.org/Projects/Datasets

2. `t_coffee -other_pg nature_protocol.pl`

käsurealt

Section 1: Protein MSAs

~10 min

	Mode	Description
A	T-Coffee default (`tcoffee`)	This default mode runs on all types of sequences and provides a reasonable tradeoff between speed and accuracy
B	Fast M-Coffee (`fmcoffee`)	This generic mode runs on all types of sequences. It applies the three fastest aligners (Kalign, MAFFT and MUSCLE) and combines their output into a unique model. It is much faster albeit less accurate than the default T-Coffee mode. It can align up to 1,000 sequences
C	PSI-Coffee (`psicoffee`)	This mode uses homology extension and is suitable for aligning remote homologs when no structural information is available. It can align up to 200 sequences
D	Expresso/3D-Coffee (`expresso`)	This mode automatically fetches and uses structural information to align protein sequences. It is best suited when all the sequences in the database have a close homolog with a known 3D structure. It is the most accurate but also the slowest mode of T-Coffee. It can align up to a 100 sequences

A Regular T-Coffee

käsurealt

```
% t_coffee -seq sh3.fasta
```

mac's

```
% open
```

```
sh3.dnd
sh3.html
sh3.aln
```

A Regular T-Coffee

käsurealt

```
% t_coffee -seq sh3.fasta
```

mac's

```
% open
```

```
sh3.dnd
sh3.html
sh3.aln
```

OPTIONAL

controls symbols per line in output

```
% export ALN_LINE_LENGTH=120
```

B Fast M-Coffee

Kalign, MAFFT and MUSCLE

```
% t_coffee -seq sh3.fasta -mode fmcoffee
```

more specific

```
% t_coffee -seq sh3.fasta method \
muscle_msa probcons_msa clustalw_msa
```

Up to 1000 sequences

C Psi-Coffee

BLAST'ib EBI's

BLAST'i kohta vt. BOX 3

```
% t_coffee -seq sh3.fasta -mode psicoffee
```

- Kõige uuem liige selles grupis
- Kasutada siis kui puuduvad struktuurid
- Koostab igale järjestusele BLAST'i otsingu tulemusena proofiili
- Library koostatakse profiilide põhjal pair-HMM või HH-align algoritme kasutades

NB! vt. CAUTION!

Up to 200 sequences

D Exspresso

BLAST'ib EBI's või NCBI's

```
% t_coffee -seq sh3.fasta -mode expresso
```

- **Poetntsiaalselt** kõige täpsem joondus
- ***Template based*** meetod - ***template***'id leitakse BLAST'a PDB vastu
- Librari koostatakse struktuuripõhise joondajaga SAP
- Ideaalis oleks ***template***'i identusus 50% siis Expresso poolt koostatud joondus on võrdne struktuuride MSA'a

Up to 100 sequences

BOX 5

Template File

The structural templates associated with each sequence of the input data set are listed in the file

`sh3_pdb1.template_list`

Sequence name

PDB structure

```
>1PHT _P_ 3I5SD
>1BB9 _P_ 1BB9A
```

Templete type

Cahin

D 3D-coffee & Exspresso

```
% t_coffee -seq sh3.fasta -method \
sap_pair -template_file sh3.template_file
```

3D-Coffee can be used as an alternative to Expresso, whereby users can specify the templates to be used for each sequence and the method to be used to align the templates. It can even be used to replicate Expresso. For instance, the above command line will produce an alignment similar to Expresso.

T-RMSD

~2 min

The T-RMSD can be used only if all sequences in the alignment have a structural template, then structure-based comparisons can be applied to estimate the distances between every sequence pair. The T-RMSD performs such analysis by comparing intra-molecular distances between pairs of ungapped positions (in the MSA). The clustering is therefore a structure-based classification of your sequences.

! CAUTION In order to insure a meaningful clustering, it may be useful to remove the sequences causing large number of indels (as estimated by visual inspection).

8 T-RMSD

~2 min

EXTRACT and MOVE DATA

```
% tar -zxvf PDBs.tar.gz
% mv PDBs/* .
```

CREATE STRUCTURE BASED ALIGNMENT (3D-Coffee)

```
% t_coffee -seq crd.fasta -template_file \
crd.template_file -method sap_pair \
TMalign_pair
```

9 T-RMSD

~2 min

```
% t_coffee -other_pg trmsd -aln crd.aln \
-template_file crd.template_file
```

Paariviisilised distantsid arvutatakse vsatavalt struktuuride erinevusele arvestades ainult neid positsioone kus joonduses puuduvad gapid

10 Analyze/Visualize the results

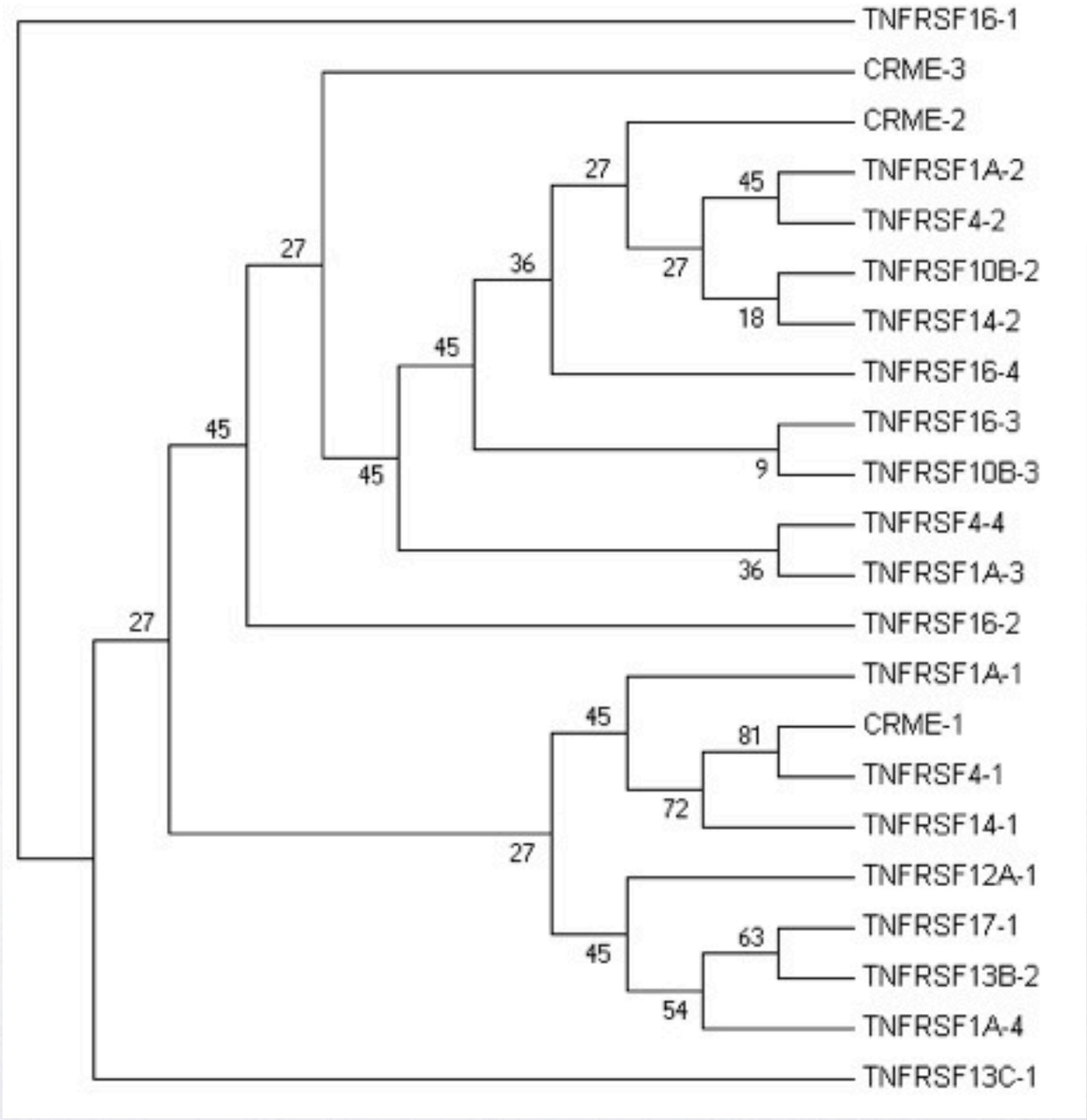
crd.struc_tree.list (text)	List of trees in Newick format. Each tree is estimated for a single ungapped column of the input MSA
crd.struc_tree.consensus (text)	Newick format file corresponding to the consensus tree made by Consense from the collection of trees contained in the previous file
crd.struc_tree.consense_output (text)	Statistics produced by Consense when producing the consensus tree
crd.struc_tree.html (html)	Color-coded version of the MSA indicating the level of support given by every ungapped column to the topology reported in crd.struc_tree.consensus

10 Analyze/Visualize the results

`crd.struc_tree.list`

avatud programmiga MEGA5

11 puu konsensus



inconsistencies between platforms

"Consistency, Bobby, consistency!"

Experiments with different compilations (version 9.01) on different platforms

1. mac precompiled installation
2. LINUX installation compiled from src
3. Protein Dataset supplied by authors (Nature Protocol)

```
% t_coffee -seq crd.fasta -template_file \
crd.template_file -method sap_pair \
TMalign_pair
```

Results in mac

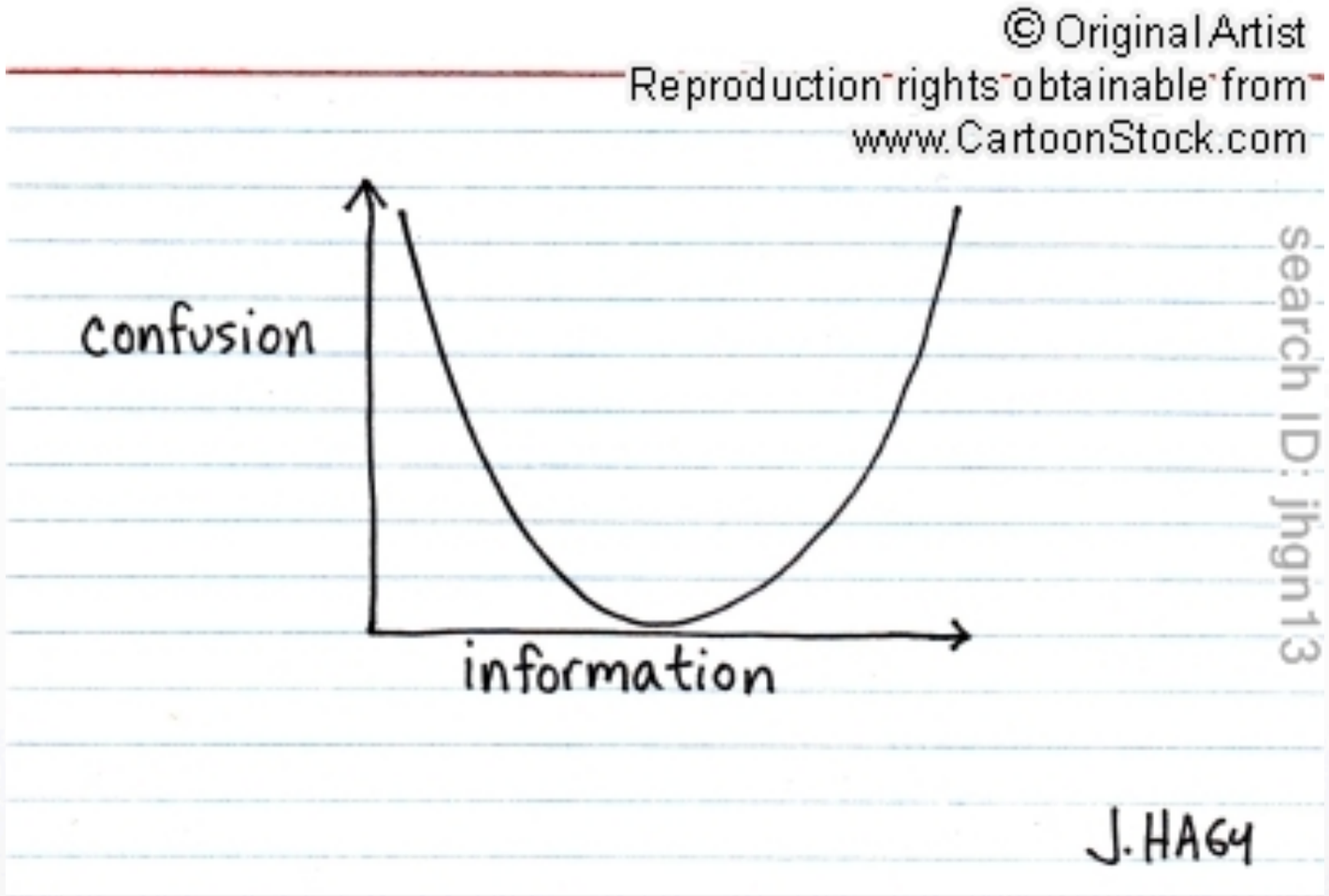
```
#  max OSX 10.6  precompiled version

CLUSTAL FORMAT for T-COFFEE Version_9.01 [http://www.tcoffee.org] [MODE:  ], CPU=14.61 sec,
SCORE=9, Nseq=22, Len=56

CRME-3          CD--S-N-SYCLLKASDG-NCVTCAPKT---KCGR--GYGKKGE-D-EM-G-NTIC
CRME-2          CP-S--D-TFTSIY----NRSPWCHSCRG--PCGT--NRVEVTPCT-PT-T-NRIC
CRME-1          CE--QGV-SYYNSQ--------ELKCCKL---CKP--GTYSDHRCD-KY-S-DTIC
TNFRSF4-4       CP-P--G-HFSPGD----NQ--ACKPWTN--C-TLA-GKHTLQPAS-NS-S-DAIC
TNFRSF4-2       CG--P-G-FYNDVV---S-SK-PCKPCT---WCNLRSGSERKQLCT-AT-Q-DTVC
TNFRSF4-1       CV--G-D-TYPS----------NDRCCH---ECRP--GNGMVSRCS-RS-Q-NTVC
TNFRSF12A-1     CS--R-G-SSWSAD------LDKCMDCAS---CR----ARPHSDFC--L-G----C
TNFRSF17-1      CS--Q-N-EYFDSL------LHACIPCQL---RCS--SNT--PPLT--C---QRYC
TNFRSF13B-2     CRKEQ-G-KFYDHL------LRDCISCA---SICG--QHP--K----QC-A-YF-C
TNFRSF1A-4      CH--A-G-FFLRE--------NECVSCS---NCKK--SLEC-TK-L---------C
TNFRSF13C-1     CV-P--A-ECFDLL----VR--HCVACGL--L-RTPRPK---PA-----------
TNFRSF1A-3      CR--K-N-QYRHYW---SENLFQCFNCSL---CL---NGTVHLSCQ-EK-Q-NTVC
TNFRSF1A-2      CE-S--G-SFTASE----NHLRHCLSCSK--C-RKEMGQVEISSCT-VD-R-DTVC
TNFRSF1A-1      CP-Q--G-KYIHPQ----NNS-ICCTKCH-----K--GTYLYNDCPGPG-Q-DTDC
TNFRSF16-4      CP--E-G-TYSDEA---N-HVDPCLPCTV---CED--TERQLRECT-PWAD--AEC
TNFRSF16-3      CA--Y-G-YYQDEE------TGHCEACSV---CEV--GSGLVFSCQ--D-KQNTVC
TNFRSF16-2      CL--DNV-TFSDVV---S-AT---EPCKPCTECL-G-LQSMSAPCV-EA-D-DAVC
TNFRSF16-1      CS--TGLYTHSG-----------ECCK---ACN-L-GEGVAQPCG--A-N-QTVC
TNFRSF14-2      CP--P-G-TYIAHL---N-GLSKCLQCQ---MCDPAMGLRASRNCS-RT-E-NAVC
TNFRSF14-1      CK--E-D-EYPV--------GSEC--CPK---CSP--GYRVKEACG-EL-T-GTVC
TNFRSF10B-3     CE--E-G-TFREED---S--PEMCRKCRT--GCPR--GMVKVGDCT-P-WS-DIEC
TNFRSF10B-2     CK-YG-Q-DYSTHW----NDLLFCLRCTR---CDS--GEVELSPCT-TT-R-NTVC
```

This alignment is NOT similar the article alignment

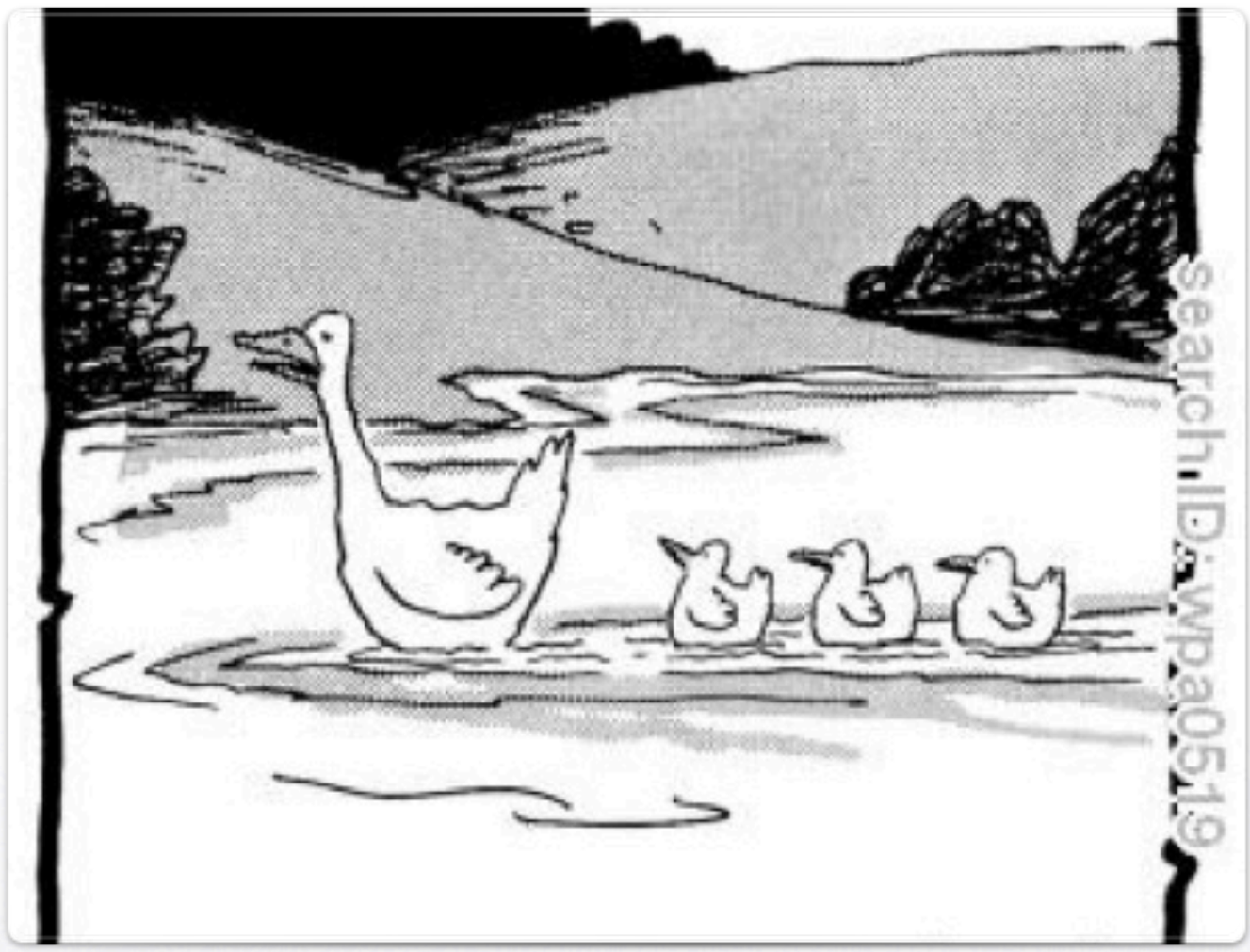
Results in LINUX src

```
#  LINUX multicore compiled from src

CLUSTAL FORMAT for T-COFFEE Version_9.01 [http://www.tcoffee.org] [MODE:  ], CPU=0.00 sec,
SCORE=80, Nseq=22, Len=50

CRME-3          -C--D-S-NSYCLLKASDGNCVTCAPKTKCGRG---YGKKGEDEMGNTIC
CRME-2          -C--P-S-DTFTSIYN---RSPWCHSCRGPCGT-NRVEVTPCTPTTNRIC
CRME-1          CE--Q-G-VSYYN-------SQELKCCKLCKP--GTYSDHRCDKYSDTIC
TNFRSF4-4       -C--P-P-GHFSPGDN-----QACKPWTNCTLA-GKHTLQPASNSSDAIC
TNFRSF4-2       -C--G-P-GFYNDVVS---SK-PCKPCTWCNLRSGSERKQLCTATQDTVC
TNFRSF4-1       -C--V-G-DTYPS-------NDRCCHE--CRP--GNGMVSRCSRSQNTVC
TNFRSF12A-1     -C--S-R-GSSWSAD-----LDKCMDCASCR----ARPHSDFCLGC----
TNFRSF17-1      -C--S-Q-NEYFDSL-----LHACIPCQLRCSS---NTPPLTCQR---YC
TNFRSF13B-2     -CRKE-Q-GKFYDH-----LLRDCISCASICGQ---HPKQCAYFC-----
TNFRSF1A-4      ---CH-A-GFFLRE-------NECVSCSNCKKS---LECTKL-------C
TNFRSF13C-1     -C--V-P-AECFDLLV-----RHCVACGLLRTPRPKPA-----------
TNFRSF1A-3      -C--R-K-NQYRHYWS--ENLFQCFNCSLCLN---GTVHLSCQEKQNTVC
TNFRSF1A-2      -C--E-S-GSFTASEN---HLRHCLSCSKCRKEMGQVEISSCTVDRDTVC
TNFRSF1A-1      -C--P-Q-GKYIHPQN---N-SICCTKCHKGTYL---YNDCPGPGQDTDC
TNFRSF16-4      -C--P-E-GTYSDEAN--H-VDPCLPCTVCED--TERQLRECTPWADAEC
TNFRSF16-3      -C--A-Y-GYYQDEET-----GHCEACSVCEV--GSGLVFSCQDKQNTVC
TNFRSF16-2      -C--L-DNVTFSDVVS---ATEPCKPCTECLG--LQSMSAPCVEADDAVC
TNFRSF16-1      -C--S-T-GLYT-------H--SGECCKACNL--GEGVAQPCG-ANQTVC
TNFRSF14-2      -C--P-P-GTYIAHLN---GLSKCLQCQMCDPAMGLRASRNCSRTENAVC
TNFRSF14-1      -C--K-E-DEYPVG-------SECCPKCSPG----YRVKEACGELTGTVC
TNFRSF10B-3     -C--E-E-GTFREEDS--P--EMCRKCRTGCPR-GMVKVGDCTPWSDIEC
TNFRSF10B-2     -C--KYG-QDYSTHWN---DLLFCLRCTRCDS--GEVELSPCTTTRNTVC
```

This alignment IS SIMILAR to the article's alignment



solving the problem is in progress ....



Ja nüüd oma andmetega!

'The trick is to make it look as easy as possible while, underneath, you're paddling like hell.'