Multiple Sequence Alignment

Group 1

Liu Kai

Why

- Identification of conserved sequence pattern and motif
- Essential prerequisite to building phylogenetic tree
- Requirement to predict protein secondary and tertiary structure
- Designing degenerate PCR

Algorithms

- Exhaustive Algorithm

Scoring matrices

- Heuristic Algorithm

Exhaustive Algorithm

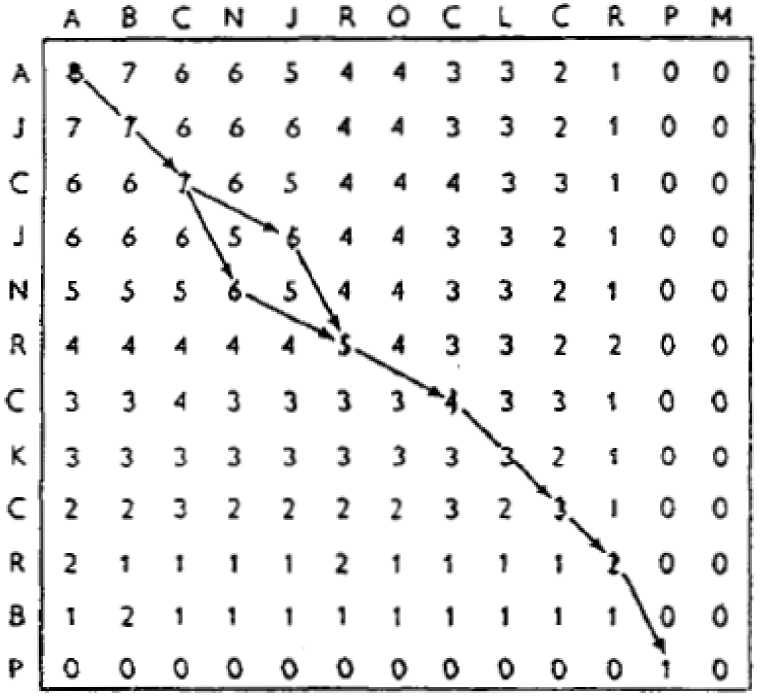
- Dynamic Programming

----Find OPTIMAL alignment

----But NOT practical for multiple sequence alignment

Dynamic Programming

- The maximum match can be determined by representing in a two-dimensional array, all possible pair combinations that can be constructed from the amino acid sequences of the proteins, A and *B,* being compared.



	A	B	C	N	J	R	O	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
J	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
J	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Needleman S. & Wunsch C.

***J. Mol. Bol. 48,* 443-453 (1970)**

Heuristic Algorithm

- Find solutions among all possible ones , but they do NOT guarantee that the best will be found, therefore they may be considered as approximately and not accurate algorithms. These algorithms, usually find a solution close to the best one and they find it fast and easily.

Heuristic Algorithm

- Progressive Alignment

Clustal and Tcoffee

- Iterative Alignment

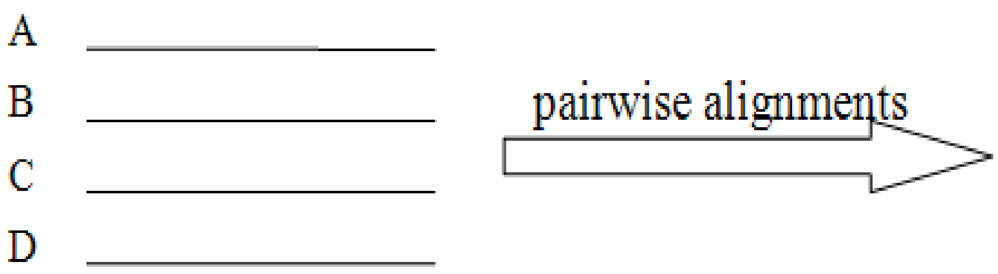
- Block-Based Alignment

Progressive Alignment

- Stepwise assembly of multiple alignment

First step:

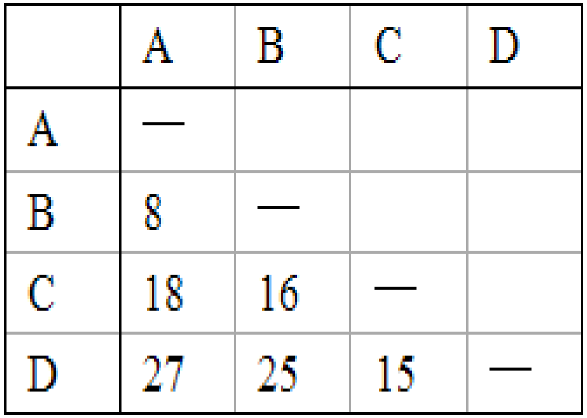
- Conduct pairwise alignments for each possible pair of sequences and get similarity scores from them.



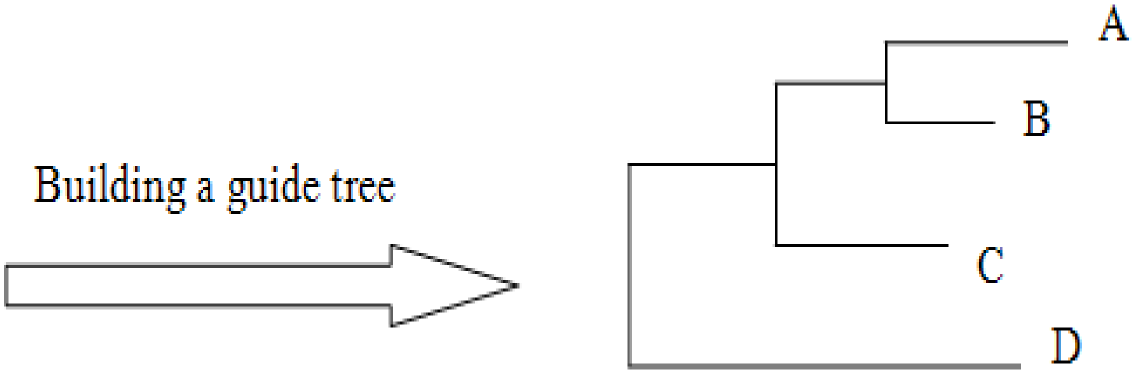
	A	B	C	D
A	—			
B	8	—		
C	18	16	—	
D	27	25	15	—

Second step:

- Calculate a guide tree



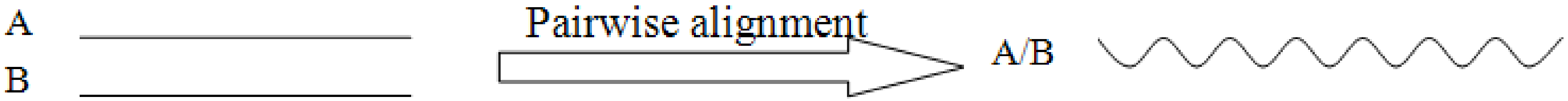
	A	B	C	D
A	—			
B	8	—		
C	18	16	—	
D	27	25	15	—



It’s only an approximate tree

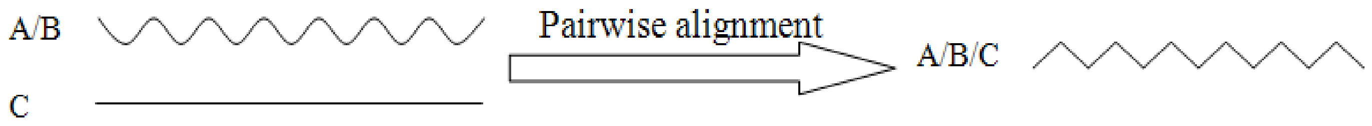
Third step:

- Align closest pair using dynimic programming and generate a consensus sequence A/B



Forth step:

- Align A/B and C using dynimic programming and generate a consensus NEW sequence A/B/C



Fifth step:

- Align A/B/C and D using dynimic programming and complete the multiple sequence alignment

Clustal

```
CLUSTAL W (1.83) multiple sequence alignment


FOS_RAT         MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_MOUSE       MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_CHICK       MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANF 60
FOSB_MOUSE      -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
FOSB_HUMAN      -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
                 *:..* .:*:: .***** **:.:*   * *..***.*  :.. :*: *:.*.  ...*

FOS_RAT         IPTVTAISTSPDLQWLVQPTLVSSVAPSQ-------TRAPHPYGLPTPS-TGAYARAGVV 112
FOS_MOUSE       IPTVTAISTSPDLQWLVQPTLVSSVAPSQ-------TRAPHPYGLPTQS-AGAYARAGMV 112
FOS_CHICK       VPTVTAISTSPDLQWLVQPTLISSVAPSQ-------NRG-HPYGVPAPAPPAAYSRPAVL 112
FOSB_MOUSE      VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTS----YSTPGLS 110
FOSB_HUMAN      VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTS----YSTPGMS 110
                :******:** **********:**:* **           .**.:*  :    *: ..:

FOS_RAT         KTMSGGRAQSIG--------------------RRGKVEQLSPEEEEKRRIRRERNKMAAA 152
FOS_MOUSE       KTVSGGRAQSIG--------------------RRGKVEQLSPEEEEKRRIRRERNKMAAA 152
FOS_CHICK       KAP-GGRGQSIG--------------------RRGKVEQLSPEEEEKRRIRRERNKMAAA 151
FOSB_MOUSE      AYSTGGASGSGGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAA 170
FOSB_HUMAN      GYSSGGASGSGGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAA 170
                    ** . * *                    ** : * *:********:******:***

FOS_RAT         KCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPNDLGF 212
FOS_MOUSE       KCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGF 212
FOS_CHICK       KCRNRRRELTDTLQAETDQLEEEKSALQAEIANLLKEKEKLEFILAAHRPACKMPEELRF 211
FOSB_MOUSE      KCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEG- 229
FOSB_HUMAN      KCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEG- 229
                *********** *********:**: *::***:* ****:***:*.**:*.**:* :

FOS_RAT         PEEMSVTS-LDLTGGLPEATTPESEEAFTLPLLNDPEPK-PSLEPVKNISNMELKAEPFD 270
FOS_MOUSE       PEEMSVAS-LDLTGGLPEASTPESEEAFTLPLLNDPEPK-PSLEPVKSISNVELKAEPFD 270
FOS_CHICK       SEELAAATALDLG----APSPAAAEEAFALPLMTEAPPAVPPKEPSG--SGLELKAEPFD 265
FOSB_MOUSE      PGPGPLAEVRDLPG-----STSAKEDGFGWLLPPPPPPP-----------------LPFQ 267
FOSB_HUMAN      PGPGPLAEVRDLPG-----SAPAKEDGFSWLLPPPPPPP-----------------LPFQ 267
                .   . :   **       :..  *:.*   *   . *                   **:
```

http://www.ebi.ac.uk/clustalw/

Clustal

- Important features:
  - ☺ Flexibility of using substitution matrices
  - ☺ Adjustable gap penalti
  - 0
  - ent

Closely related sequences— BL M62 or PAM120
Divergent sequences M45 or PAM250

A gap near a series of hydrophobic residues carries more penalties than the one next to a series of hydrophilic residues

Downweighting redundant and closely related groups of sequences in the alignment by a certain factor

Clustal

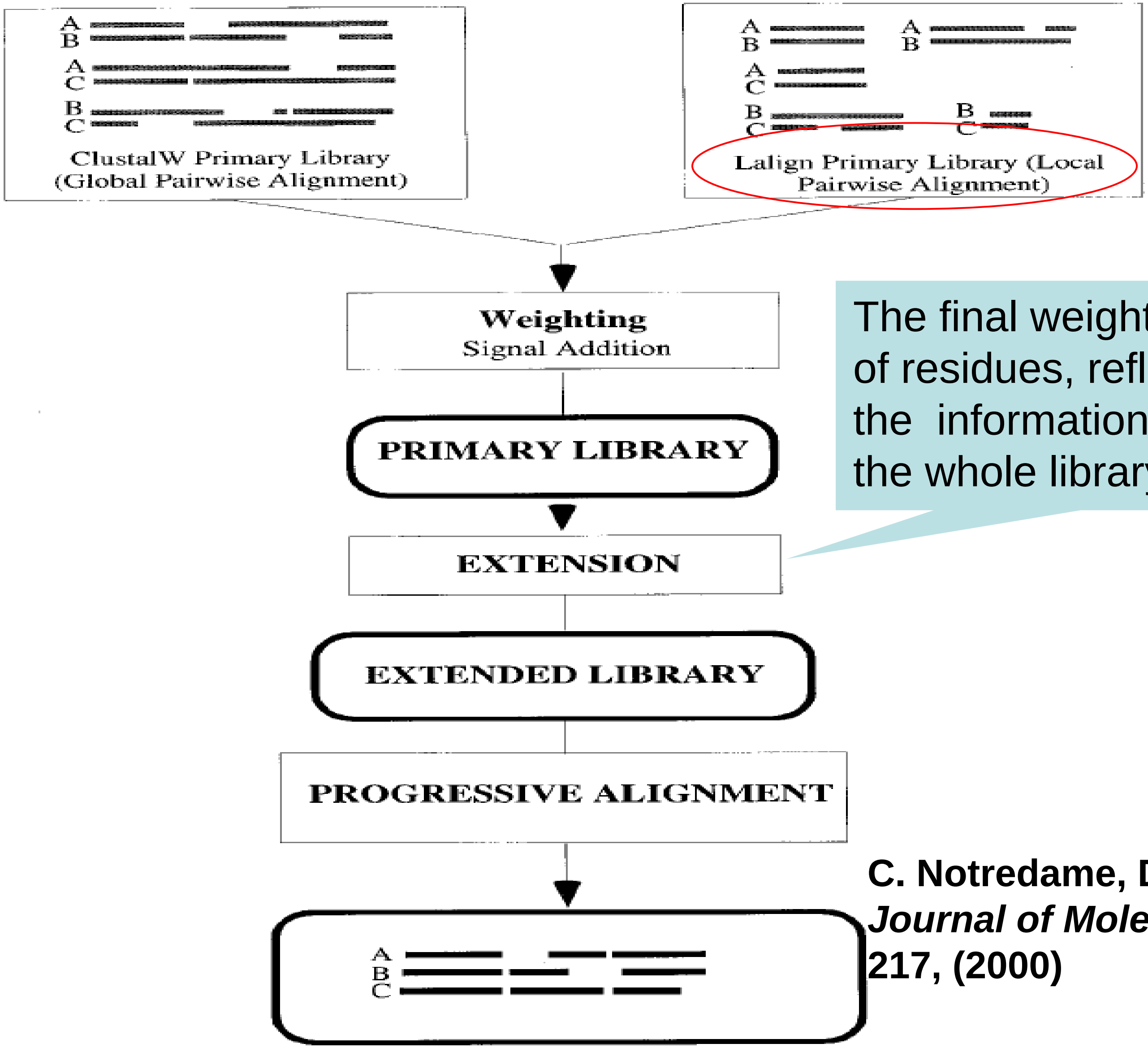
- Drawbacks:
  - ☻ Global alignment-based method

→not suitable for comparing sequences of different lengths

☻ “Greedy” nature: Once an error, always an error

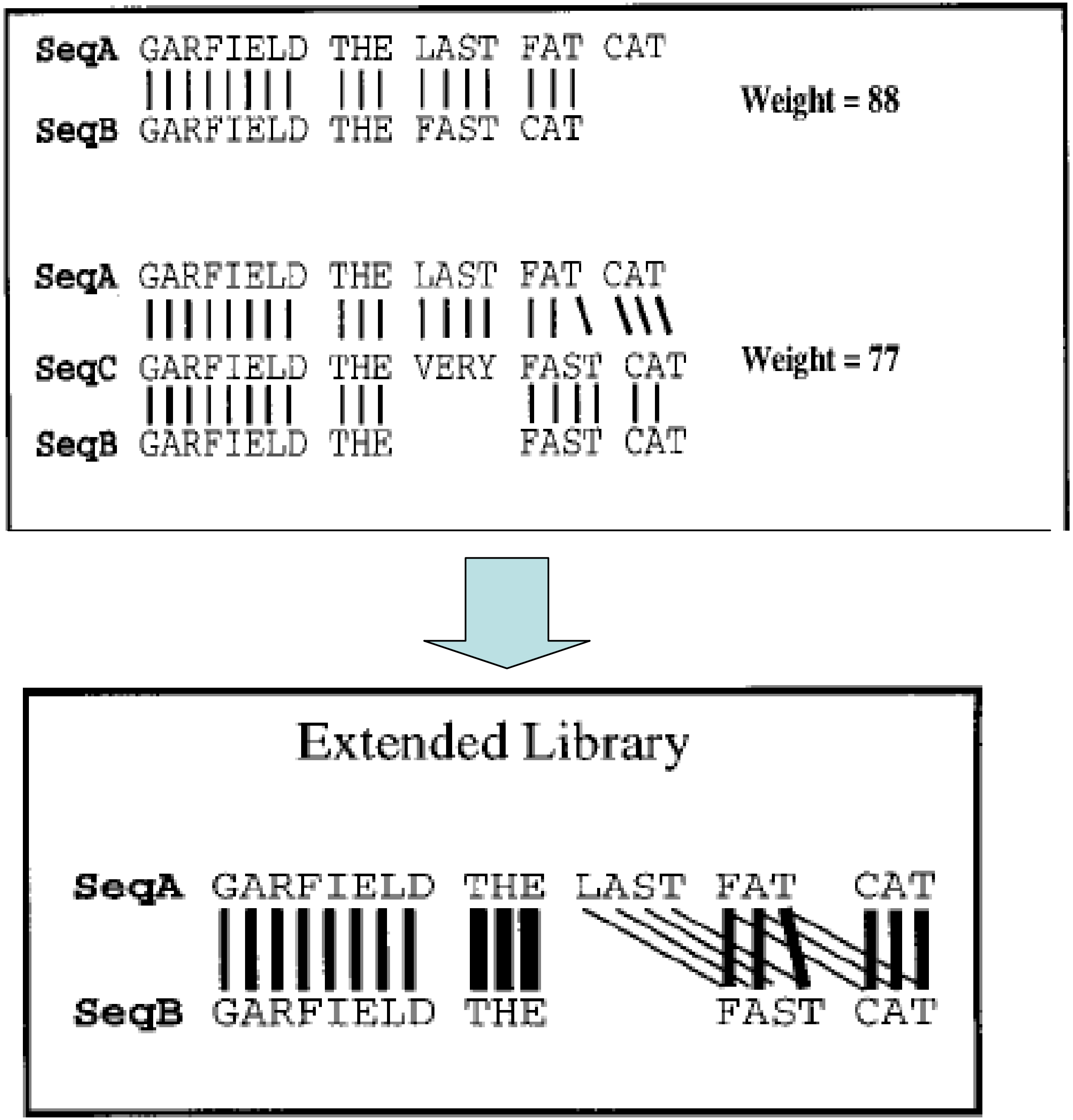
Tcoffee

Tcoffee



The final weight, for any pair of residues, reflects some of the information contained in the whole library.

C. Notredame, D. Higgins, J. Heringa
***Journal of Molecular Biology*, 302, 205-217, (2000)**



C. Notredame, D. Higgins, J. Heringa
Journal of Molecular Biology, **302**, 205-217, (2000)

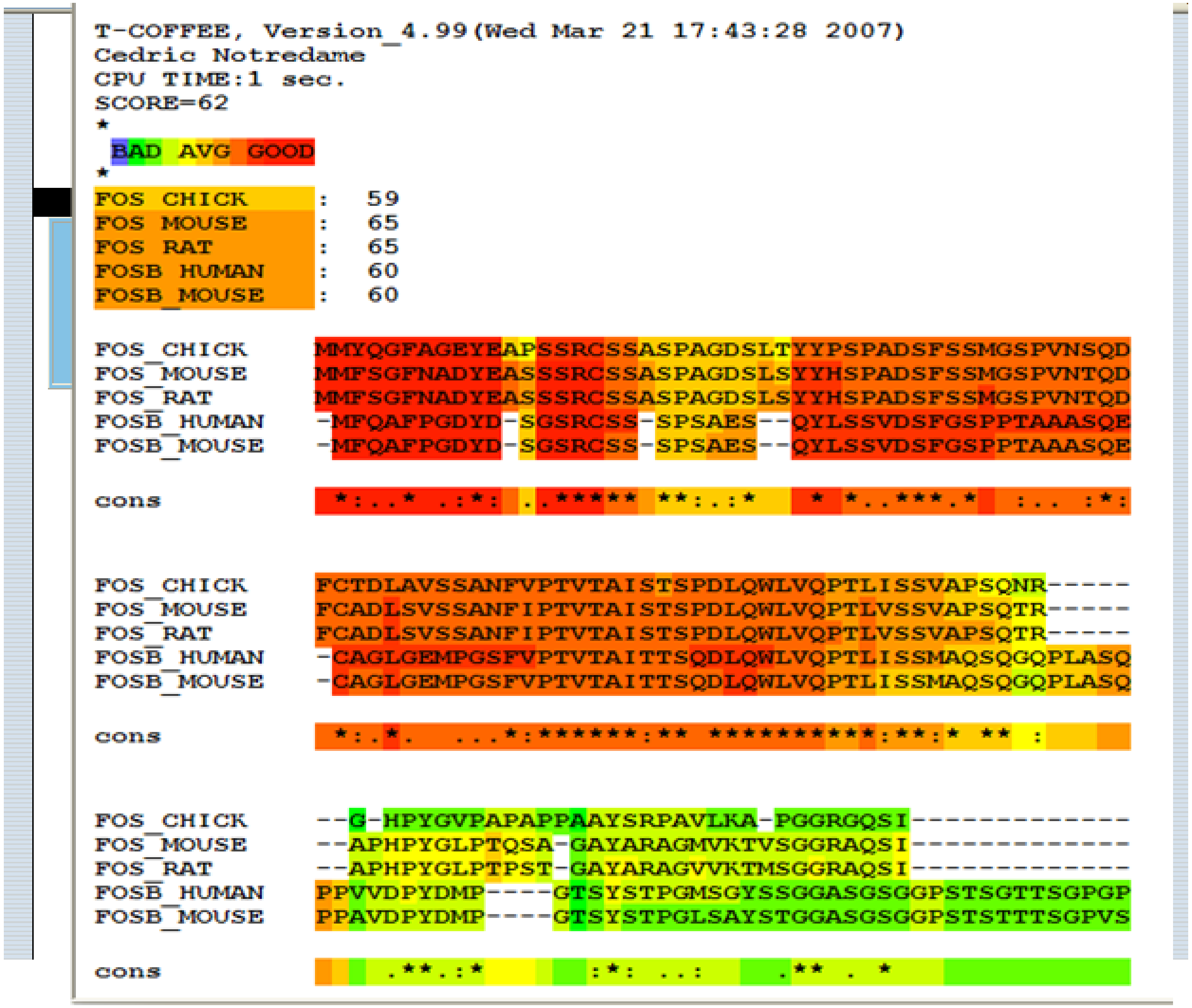
Tcoffee vs Clustal

- Tcoffee indeed outperforms Clustal when aligning moderately divergent sequences

- Slower

- NOT completely solve the problems

Tcoffee



```
T-COFFEE, Version_4.99(Wed Mar 21 17:43:28 2007)
Cedric Notredame
CPU TIME:1 sec.
SCORE=62
*
 BAD AVG GOOD
*
FOS_CHICK    :   59
FOS_MOUSE    :   65
FOS_RAT      :   65
FOSB_HUMAN   :   60
FOSB_MOUSE   :   60

FOS_CHICK    MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQD
FOS_MOUSE    MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQD
FOS_RAT      MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQD
FOSB_HUMAN   -MFQAFPGDYD-SGSRCSS-SPSAES--QYLSSVDSFGSPPTAAASQE
FOSB_MOUSE   -MFQAFPGDYD-SGSRCSS-SPSAES--QYLSSVDSFGSPPTAAASQE

cons          *:..* .:*: ..***** **:.:*   * *..***.*  :.. :*:


FOS_CHICK    FCTDLAVSSANFVPTVTAISTSPDLQWLVQPTLISSVAPSQNR-----
FOS_MOUSE    FCADLSVSSANFIPTVTAISTSPDLQWLVQPTLVSSVAPSQTR-----
FOS_RAT      FCADLSVSSANFIPTVTAISTSPDLQWLVQPTLVSSVAPSQTR-----
FOSB_HUMAN   -CAGLGEMPGSFVPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQ
FOSB_MOUSE   -CAGLGEMPGSFVPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQ

cons          *:.*.  ...*:******:** **********:**:* ** :


FOS_CHICK    --G-HPYGVPAPAPPAAYSRPAVLKA-PGGRGQSI-------------
FOS_MOUSE    --APHPYGLPTQSA-GAYARAGMVKTVSGGRAQSI-------------
FOS_RAT      --APHPYGLPTPST-GAYARAGVVKTMSGGRAQSI-------------
FOSB_HUMAN   PPVVDPYDMP----GTSYSTPGMSGYSSGGASGSGGPSTSGTTSGPGP
FOSB_MOUSE   PPAVDPYDMP----GTSYSTPGLSAYSTGGASGSGGPSTSTTTSGPVS

cons             .**.:*      :*: ..:    .** . *
```

http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi

Tips

1. DNA vs Protein

No frame shift error

More informative

2. Using a combination of multiple alignment programs

3. Refine results manually or automatically