

# Multiple Alignments

**Laboratory of Bioinformatics I  
Module 2**

20 March, 2020

**Emidio Capriotti**

<http://biofold.org/>



**Biomolecules  
Folding and  
Disease**

Department of Pharmacy and  
Biotechnology (FaBiT)  
University of Bologna



# Multiple Structure Alignment

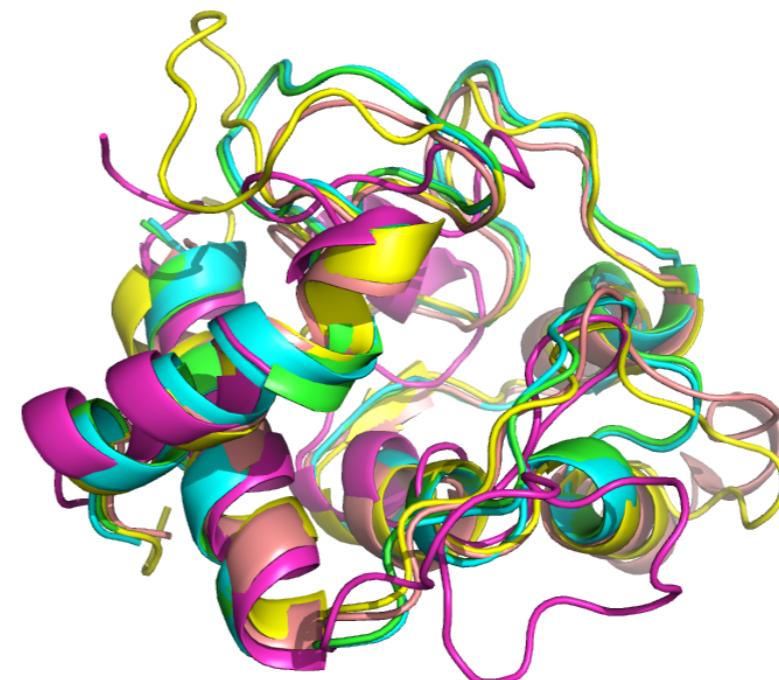
Align the structure of 5 structures of Cytochromes

3zcf:A --GDVEK**GKKI**FIMKC**sQ**CHTVEkgg-----khKTG--PNLHG--**L**fgRKTgqapgysyt---aank  
3o20:A --GDVEK**GKKI**FVQK**Ca**QCHTVEkgg-----khKTG--PNLHG--**L**fgRKTgqapgftyt---dank  
2ce0:A --LDI**QRGATLFNRACa**A**CHDTG**g-----nIIQpgATLFTkd**L**--ERN-----  
1cxc:A qe**GDPEAGAKAFNQCQ-TCH**VIVddsgttiagrnaKTG--PNLYG--**V**vgRTAgtqadfkgygegmkeag  
1i8o:A --**EDAKAGEAVFKQCM-TCHRAD**k-----nMVG--PALAG--**V**vgRKAgttaagftysp-lhnsg

3zcf:A nkgiIW-GEDTLMEYLENPKkyi-----pgTK**Mi**FvGiK-----KKEERAD  
3o20:A nkgiTW-KE**E**TLMEYLENPKkyi-----pgTK**Mi**FaGiK-----KKTERED  
2ce0:A ---GVdTEEEIYRVTYFGK-----GRM-PgF-GekctprgqctfgprlQDEEIKL  
1cxc:A akglAW-DEEHFVQYVQDPTkflkeyt-----gdakak**GK****Mt**F-K1K-----KEADAHN  
1i8o:A eaglVW-TADNIVPYLADPNaf1kkfltekkadqavgv**TK****Mt**F-K1A-----NEQQRKD

3zcf:A **L**IAY**L**KKATne----  
3o20:A **L**IAY**L**KKATne----  
2ce0:A **LAEFV**KFQAdqgwpt  
1cxc:A **I**WAY**L**QQVAvrp---  
1i8o:A **VVAY**LATLK-----

Functional sites  
Conserved sites  
Similar substitutions



# Important Information

The comparison among multiple structures **highlight** the most conserved and the most variable sites.

Conserved sites **could be** functionally or structurally important.

For each site the residue distribution is estimated

The information is not general, but family specific



# Multiple Alignment

A representation of a **set of sequences, where equivalent residues** (e.g. functional, structural) are aligned in columns.

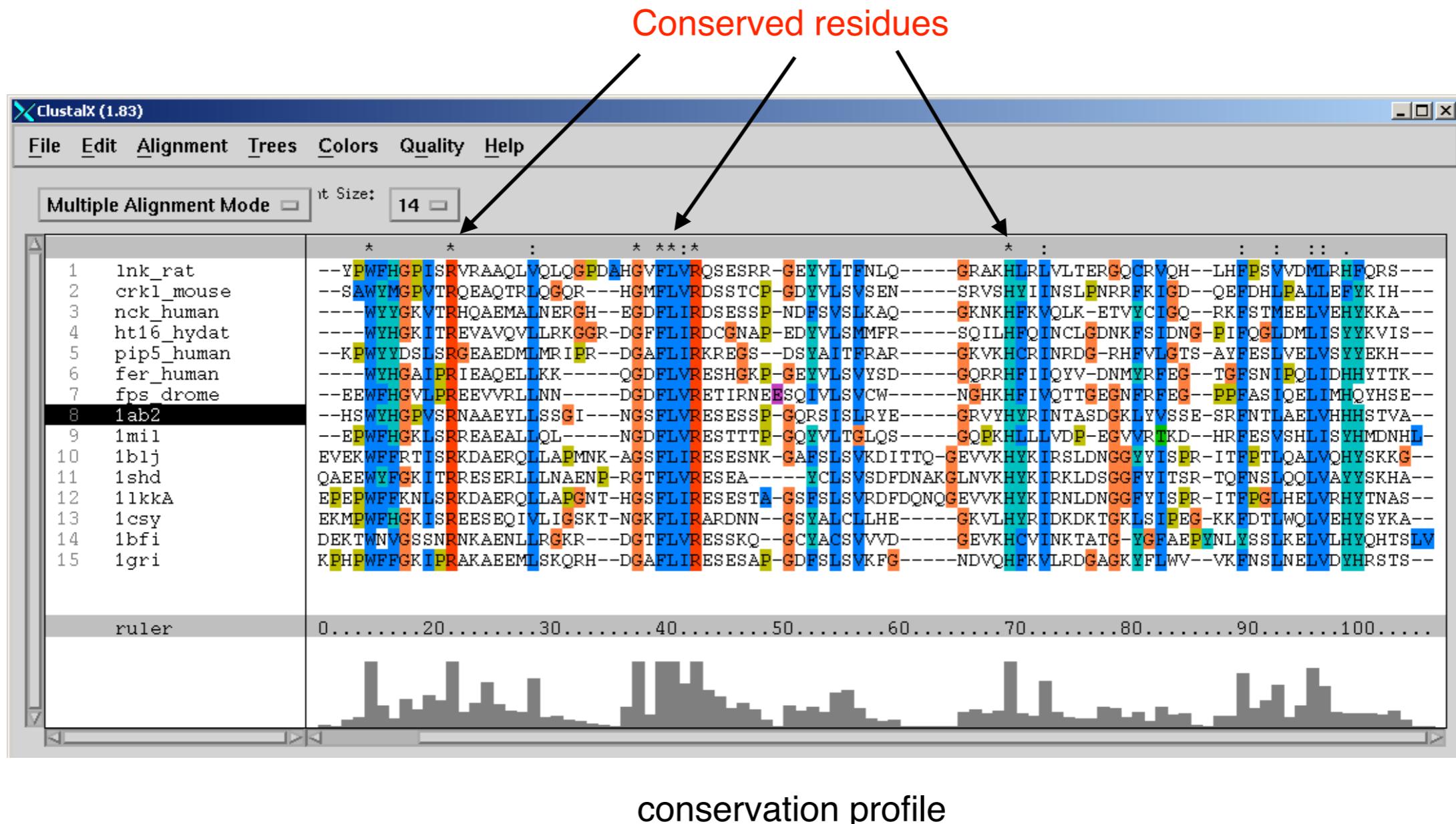
Part of an alignment of SH2 domains from 14 sequences

	*	*	:	*	**:	*	:	:	:	:	:
Ink_rat	-----	YPWFHGPISRVRAAQLVQLQGPDAHGVFLVRQSESR	-GEYVLTFNLQ-----	GRAKHLRLVLTERGQCRVQH--LHFPSVVDML							
crk1_mouse	-----	SAWYMPVTRQEAQTRLQGQR--	HGMFLVRDSSTCP-GDYVLSVSEN-----	SRVSHYIINSLPNRRFKIGD--QEFDHPALL							
nck_human	-----	WYYGKVTRHQAEMLNERGH--	EGDFLIRDSESSP-NDFSVSLKAQ-----	GKNKHFKVQLK-ETVYCIGQ--RKFSTMEELV							
ht16_hydat	-----	WYHGKITREVAVQVLLRKGGR-DGFFLIRDGNAP-EDYVLSMMFR-----	SQILHFQINCLGDNKFSIDNG-PIFQGLDMLI								
pip5_human	-----	KPWYYDSLSRGAEADMLMRIPR--DGAFLIRKREGS--DSYAITFRAR-----	GKVKHCRINRDG-RHFVLGTS-AYFESLVELV								
fer_human	-----	WYHGAIPRIEAQELLKK----QGDFLVRESHGKP-GEYVLSVYSD-----	GQRRHFIIQYV-DNMYRFEG--TGFSNIPQLI								
1ab2	-----	EEWFHGVLPREEVVRLNN----DGDFLVRETIRNEESQIVLSVCW-----	NGHKHFIVQTTGEGNFRFEG--PPFASIQELI								
1mil	-----	HSWYHGPVSRNAAEYLLSSGI---NGSFLVRESESSP-GQRSISLRYE-----	GRVYHYRINTASDGKLYVSSE-SRFNTLAEV								
1bjj	-----	EPWFHGKLRSRREAEALLQL----NGDFLVRESTTP-GQYVLTGLQS-----	GQPKHLLVDP-EGVVRTKD--HRFESVSHLI								
1shd	-----	GSVAPVETLEVEKWFFRTISRKDAERQLLAPMNK-AGSFLIRESESINK-GAFSLSVKDITQ-GEVVKHYKIRSLDNGGYIISPR-ITFPQLQALV									
1lkkA	-----	S IQAEEWYFGK ITRRESERLLLNAENP-RGTFLVRESEA-----	YCLSVSDFDNAKGLNVKHYKIRKLDGGFYITSR-TQFNSLQQLV								
1csy	-----	LEPEPWFFKNLSRKDAERQLLAPGNT-HGSFLIREESTA-GSFSLSVRDFDQNQGEVVKHYKIRNLDNGGFYIISPR-ITFPGLHELV									
1bfi	-----	SHEKMPWFHGKISREESEQIVLIGSKT-NGKFLIRARDNN--GSYALCLLHE-----	GKVLHYRIDKDKTGKLSIPEG-KKFDTLWQLV								
1gri	-----	HHDEKTWNVGSSNRNKAENLLRGKR--DGTFLVRESSKQ--GCYACSVVVD-----	GEVKHCVINKTATG-YGFAEPYNLYSSLKELV								
	-----	EMKPHPWFFGKIPRAKAEEMLSQRH--DGAFLIRESESAP-GDFSLSVKFG-----	NDVQHFKVLRDGAGKYFLWV--VKFNSLNELV								

\* conserved identical residues  
: conserved similar residues

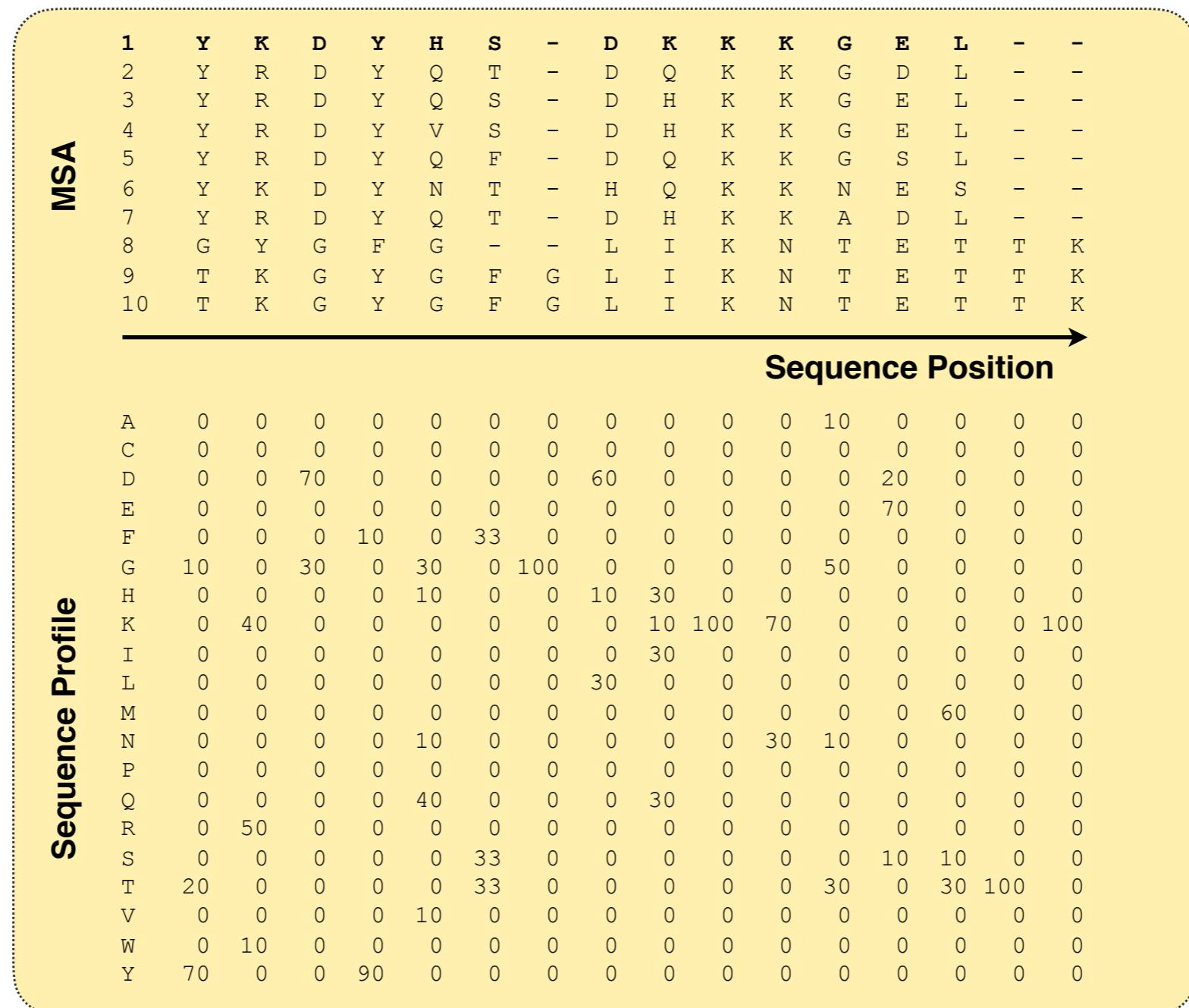
# Why is important?

Multiple sequence alignment is important to find conserved residues



# Sequence Profile

A multiple sequence alignment can be represented by sequence profile

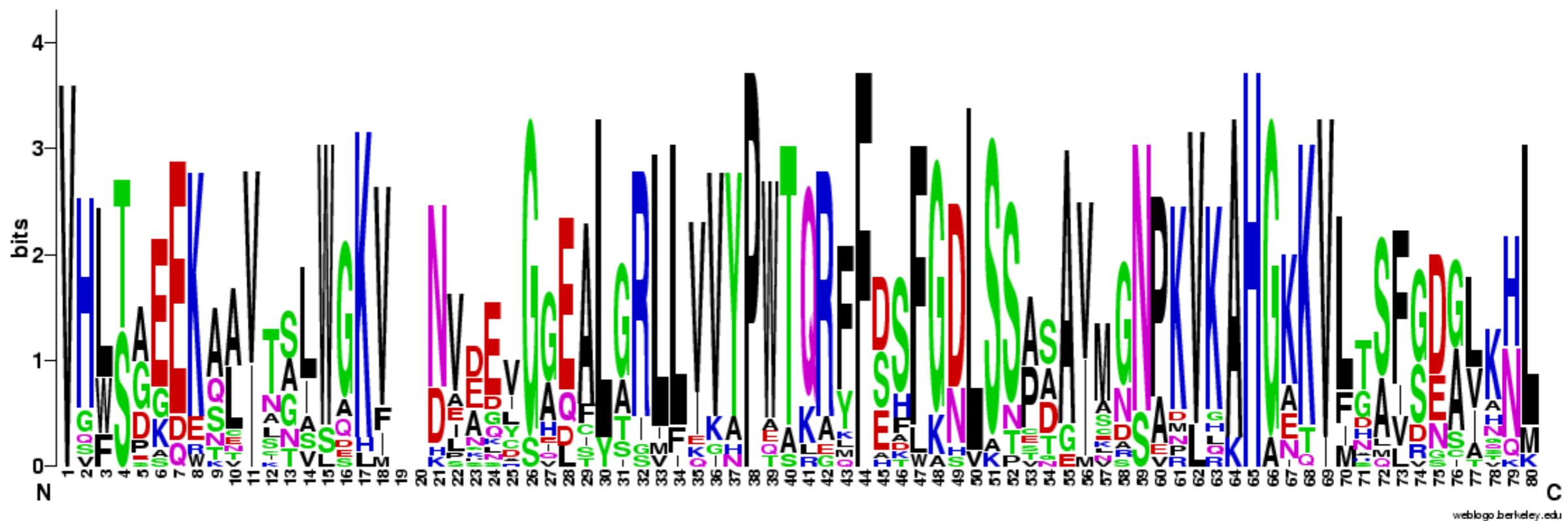


# Sequence Logo

Plot drawn with score related to the information entropy in each position

$$I = \log_2 20 - S(p)$$

$$S(p) = \sum_{i=1}^{20} -p_i \ln p_i$$



# Alignment Scoring

How to score an alignment of many sequences?

Given M sequences  $A_i$ , we can define a score for the multiple sequence alignment as the **sum of the scores of all the induced pair alignments**

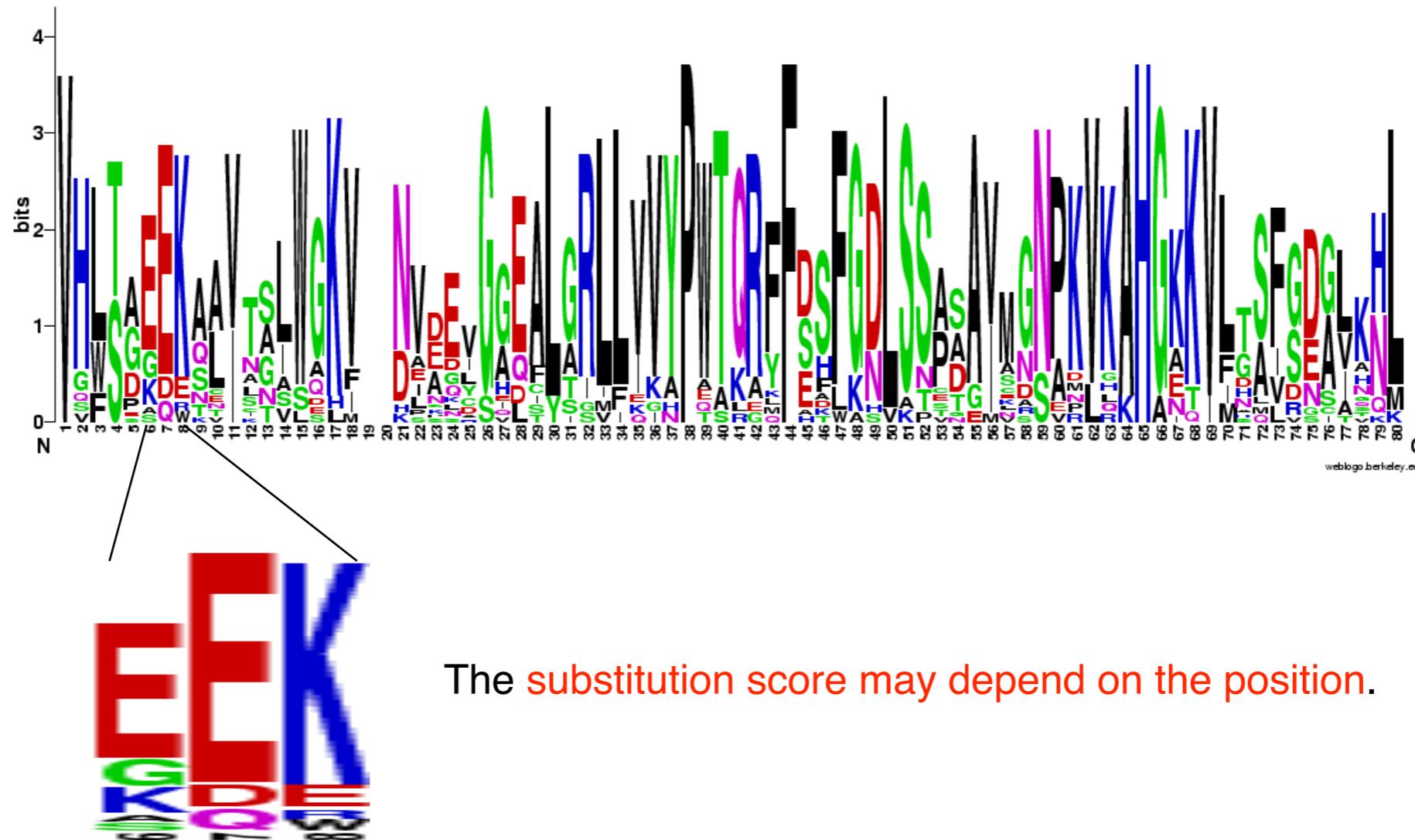
$$S = \sum_{i < j} S(A_i, A_j)$$

$$S \left[ \begin{array}{l} 1>\text{ASPTLPLSLA} \\ 2>\text{SS-TLPA--A} \\ 3>\text{SSPTLPA--A} \end{array} \right] = +S \left[ \begin{array}{l} 1>\text{ASPTLPLSLA} \\ 2>\text{SS-TLPA--A} \\ 3>\text{SSPTLPA--A} \end{array} \right] +S \left[ \begin{array}{l} 2>\text{SS-TLPA--A} \\ 3>\text{SSPTLPA--A} \end{array} \right]$$

# Entropy Score

The multiple sequence alignment can be obtained minimizing the entropy

$$S = \sum_{j=1}^{Ncolumns} \sum_{i=1}^{20} - p_{ji} \ln p_{ji}$$



# Profile-Based Alignment

Given the position  $i$  along a sequence profile, it is represented by a 20-element vector  $P_i = P_i(A) \ P_i(C) \ \dots \ P_i(Y)$

A	0
C	85
D	0
E	0
F	5
G	0
H	0
I	0
K	0
L	2
M	0
N	8
P	0
Q	0
R	0
S	0
T	0
V	0
W	0
Y	0

Given the residue in position  $j$  along the sequence to align:  $S_j$   
The score for aligning  $S_j$  to the vector  $P_i$  is:

$$Score(i, j) = \sum_{k=1}^{20} P_i(r_k) \cdot M(r_k, s_j)$$

where  $M$  is a matrix score (BLOSUM or PAM)

The score can be used in dynamic programming procedures  
(Needleman-Wunsch, Smith-Waterman)

# Sequence to Profile Score

Alignment score between  $P_i$  and  $S_i$  is

$P_i$

A	0
C	85
D	0
E	0
F	5
G	0
H	0
I	0
K	0
L	2
M	0
N	8
P	0
Q	0
R	0
S	0
T	0
V	0
W	0
Y	0

$$\begin{aligned}
 &= 0.85 * M(C, C) + 0.05 * M(C, F) + 0.02 * M(C, L) + 0.08 * M(C, N) = \\
 &= 0.85 * (9) + 0.05 * (-2) + 0.02 * (-1) + 0.08 * (-3) = 7.29
 \end{aligned}$$

$$Score(i, j) = \sum_{k=1}^{20} P_i(r_k) \cdot M(r_k, s_j)$$

$S_i = "C"$

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	2	-2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	

# Alignment Strategies

## Multiple sequence alignment (MSA)

The algorithmic problem is to find the alignment with the maximum score

## Exact algorithms

Algorithms based of multi-dimensional dynamic programming have been implemented. However they are too slow when many sequences have to be compared.

## Progressive alignments

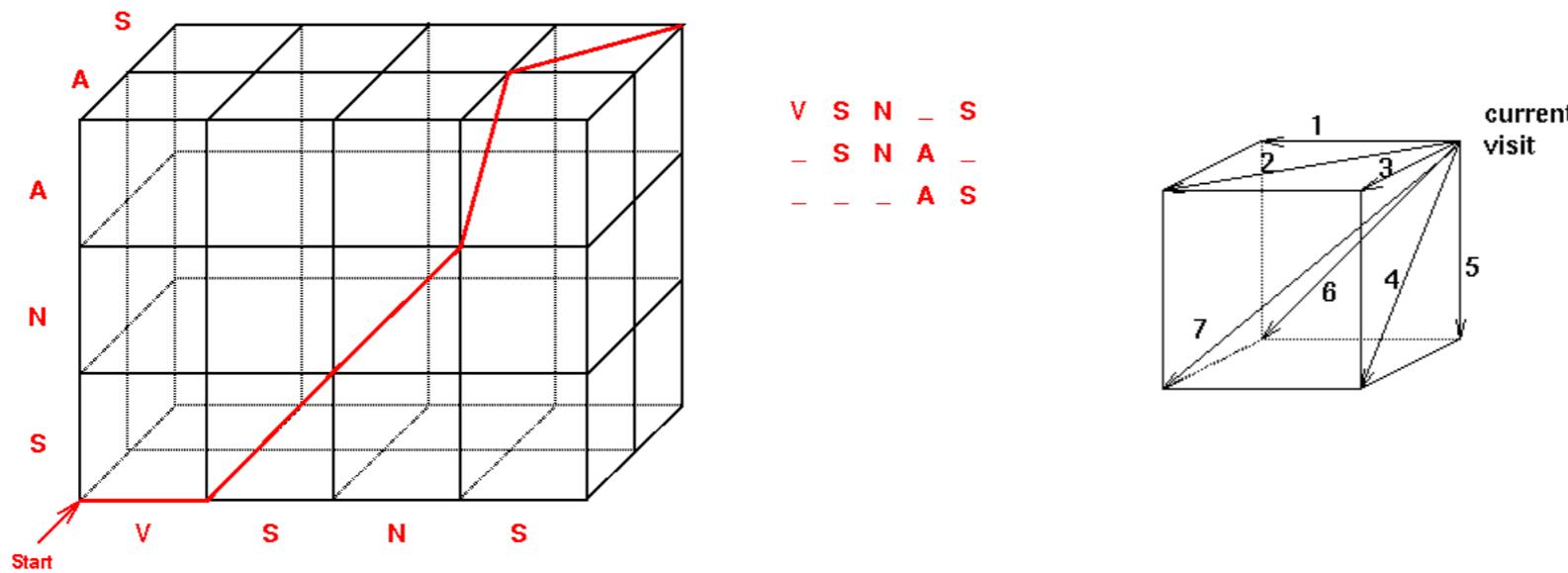
## Iterative algorithms

## Consistency-based algorithms

# Optimal Alignment

Optimal Multiple Alignment: MSA (Lipman et al. 1989, Gupta et al. 1995)

Extension of dynamic programming for 2 sequences => N dimensions



Problem: calculation time and memory requirements

Time proportional to  $N^k$  for k sequences of length N => limited to less than 10 sequences

# Progressive MSA

Idea: Progressively align pairs of sequences (or groups of sequences)

Problem:

Start with which sequences? How to decide order of alignment?

First align the most closely related sequences

How to measure the similarity of the sequences?

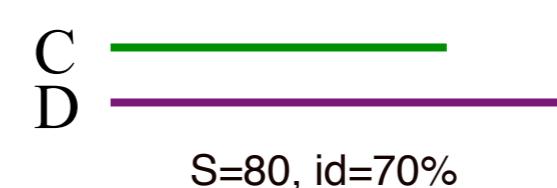
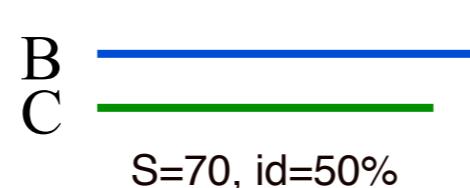
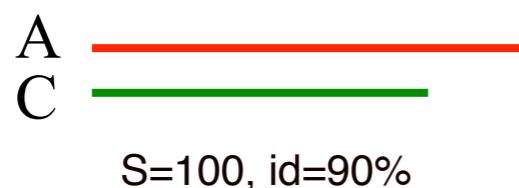
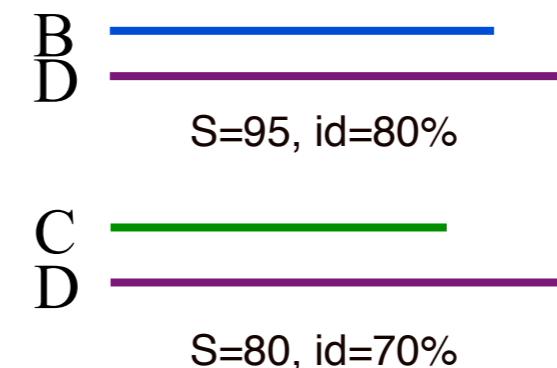
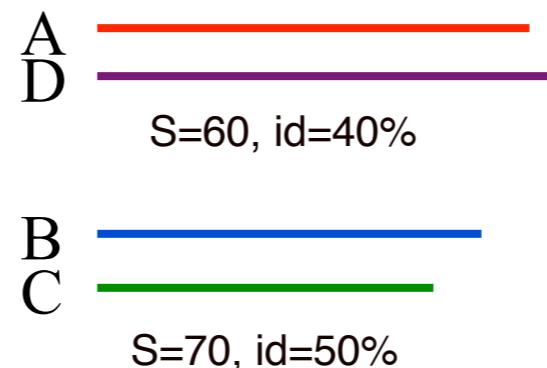
align all the sequences pairwise

calculate the similarity between each pair from the alignment

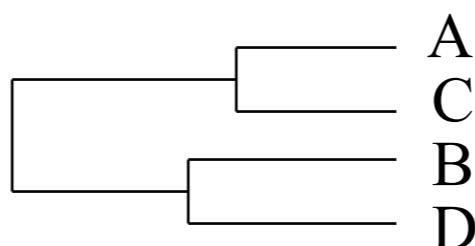
# Progressive MSA - Start



**Step1:** Pairwise sequence alignment: exact, all-against-all



**Step1:** Build a similarity tree



# Progressive MSA - End

**Step 3:** Exact alignment of the most similar sequences, following the tree



**Step 4:** Build the profile from the sub alignments

**Step 5:** Perform profile-to-profile alignment following the similarity tree, until comprising all the sequences



# Profile-Profile Alignment

The position  $i$  along the first sequence profile, it is represented by a 20-element vector  
 $P^1_i = P^1_i(A) \ P^1_i(C) \ \dots \ P^1_i(Y)$

The position  $j$  along the second sequence profile, it is represented by a 20-element vector  
 $P^2_j = P^2_j(A) \ P^2_j(C) \ \dots \ P^2_j(Y)$

The score for aligning the two positions is:

$$Score(i, j) = \sum_{m=1}^{20} \sum_{k=1}^{20} P^1_i(r_m) P^2_j(r_k) \cdot M(r_m, r_k)$$

where  $M$  is a matrix score (BLOSUM or PAM)

The score can be used in dynamic programming procedures  
(Needleman-Wunsch, Smith-Waterman)

# Adding Gaps

- Where gaps are added is a critical question
- Gaps are often added to the **first two (closest) sequences**
- To **change the initial gap choices** later on corresponds to give more **weight** to distantly related sequences
- To maintain the **initial gap choices** means that the initial gaps are the most believable

# Limitations

- Dependence of the final MSA on the initial pairwise sequence alignment with the highest score
- Errors in initial alignments are propagated
- Gaps can proliferate, if not carefully evaluated
- Gaps can be amino-acid specific, so that you penalize introduction of gaps into segments that are less likely to have gaps (e.g. hydrophobic core)

# Alignment Evaluation

## How many conserved sites?

CLUSTAL 2.1 multiple sequence alignment

sp|P99999|CYC\_HUMAN  
sp|P00004|CYC\_HORSE  
sp|P0C0X8|CYC2\_RHOSH  
sp|P00091|CYC22\_RHOPA  
sp|Q93VA3|CYC6\_ARATH

----- MGDVEKGKKIFIMK**C**S ----- Q**C**HT 20  
----- MGDVEKGKKIFVQ**K**CA ----- Q**C**HT 20  
----- QEGDPEAGAKAFNQC**Q**T**C**H VIVDDSGT 27  
----- MVKKLLTILSIAATAGSLSIGTASAQDAKAGEAVFK**Q****C**MT 40  
MRLVLSGASSFTSNLFCSSQQVNNGR**K**ELKNPISLNHNKDLD**F**LLKKLAP 50  
\*

sp|P99999|CYC\_HUMAN  
sp|P00004|CYC\_HORSE  
sp|P0C0X8|CYC2\_RHOSH  
sp|P00091|CYC22\_RHOPA  
sp|Q93VA3|CYC6\_ARATH

sp|P99999|CYC\_HUMAN  
sp|P00004|CYC\_HORSE  
sp|P0C0X8|CYC2\_RHOSH  
sp|P00091|CYC22\_RHOPA  
sp|Q93VA3|CYC6\_ARATH

LMEYLENP-----	KKYIPG-----	TK <b>M</b> IFVGI	86
LMEYLENP-----	KKYIPG-----	TK <b>M</b> IFAGI	86
FVQYVQDPTK-----	FLKEYTGD-----	AKAKGK <b>M</b> TFK-L	104
IINYLNPDNA-----	FLKKFLTDKGKADQAVGVTK <b>M</b> TFK-L		122
ATLFTKDLERNGVDTEEEIYRVTYFGKGR <b>M</b> PGFG--	EKCTPRGQCTFGPR		148
: ::	*	.	: *

sp|P99999|CYC\_HUMAN  
sp|P00004|CYC\_HORSE  
sp|P0C0X8|CYC2\_RHOSH  
sp|P00091|CYC22\_RHOPA  
sp|Q93VA3|CYC6\_ARATH

# Alternative Alignment

How many conserved sites?

CLUSTAL 2.1 multiple sequence alignment

sp|P99999|CYC\_HUMAN\_Cytochrome  
sp|P00004|CYC\_HORSE\_Cytochrome  
sp|P0C0X8|CYC2\_RHOSH\_Cytochrom  
sp|P00091|CYC22\_RHOPA\_Cytochro  
sp|Q93VA3|CYC6\_ARATH\_Cytochrom

-MGDVEKGKKIFIMK**CSQCH**TVE-----KGGKHKTGPNLHGLFGRKTG 42  
-MGDVEKGKKIFVQK**CAQCH**TVE-----KGGKHKTGPNLHGLFGRKTG 42  
QEGDPEAGAKAFN-QCQT**CH**VIVDDSGTTIAGRNAKTGPNL~~YGVVGR~~TAG 49  
--QDAKAGEAVFK-QCMT**CHR**-----ADKN-MVGPALGGVVGRKAG 37  
QTLDIQRGATLFNRAC**IGCH**DTG-----GNIIQPGATLFTKDLERNG 42  
\* : \* \* \* \*\* . \* . \* . \*

sp|P99999|CYC\_HUMAN\_Cytochrome  
sp|P00004|CYC\_HORSE\_Cytochrome  
sp|P0C0X8|CYC2\_RHOSH\_Cytochrom  
sp|P00091|CYC22\_RHOPA\_Cytochro  
sp|Q93VA3|CYC6\_ARATH\_Cytochrom

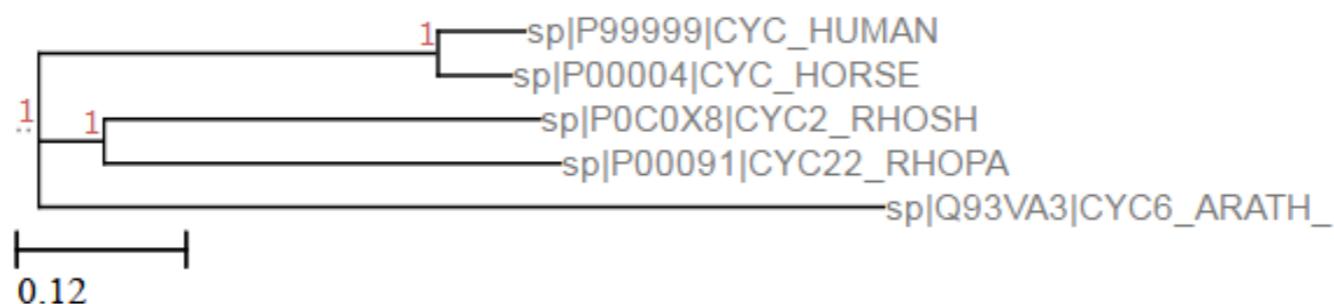
QAPGYS-YTAANKN---KGIIWGEDTLMEYLENPKKYIP----- 77  
QAPGFT-YTDANKN---KGITWKEETLMYEYLENPKKYIP----- 77  
TQADFKGYGEGMKEAGAKGLAWDEEHFVQYVQDPDKFLKEYTGD---AK 95  
TAAGFT-YSPLNHNSGEAGLVWTADNIINYLNPDNAFLKKFLTDKGKADQ 86  
VDTEEEIYRVTYFGKG-RMPGFGEKCTPRGQCTFGPRLQ----- 80  
. \* : . . . :

sp|P99999|CYC\_HUMAN\_Cytochrome  
sp|P00004|CYC\_HORSE\_Cytochrome  
sp|P0C0X8|CYC2\_RHOSH\_Cytochrom  
sp|P00091|CYC22\_RHOPA\_Cytochro  
sp|Q93VA3|CYC6\_ARATH\_Cytochrom

--GTK**M**IFVGIKKKEERADLIAYLKKATNE- 105  
--GTK**M**IFAGIKKKTEREDLIAYLKKATNE- 105  
AKG--K**MT**FKLKKEADAHNIWAYLQQVAVRP 124  
AVGVTK**MT**FKLANEQQRKDVVAYLATLK--- 114  
----DEEIKLLAEFVKFQADQGWPTVSTD- 105

# Improve the Alignment

The alignment is based on a guide tree computed on the basis of the pairwise distances (guide tree).



The sequence distances computed starting from the MSA can be different ("phylogenetic" tree)



# Phylogenetic Tree

If the trees are very different, the final MSA is somehow incoherent with respect of the procedure used to derive it.

It is then possible to iterate the progressive alignment procedure, using the “phylogenetic” tree as guide.

# Iterative Alignment Method

Iterative Methods: MUSCLE

MUSCLE (MUltiple Sequence Comparison by Log Expectation), 3 steps:

## draft progressive:

consists of a progressive sequence alignment

- I (accuracy) it uses log-expectation score instead of PPS score in profile-profile alignment;
- I (efficiency) uses k-mer distance instead of alignment score for sequence similarity (a k-mer is a substring of length k)
- I instead of neighbour joining, it uses UPGMA (Unweighted Pair Group Method with Arithmetic Mean)

## improved progressive:

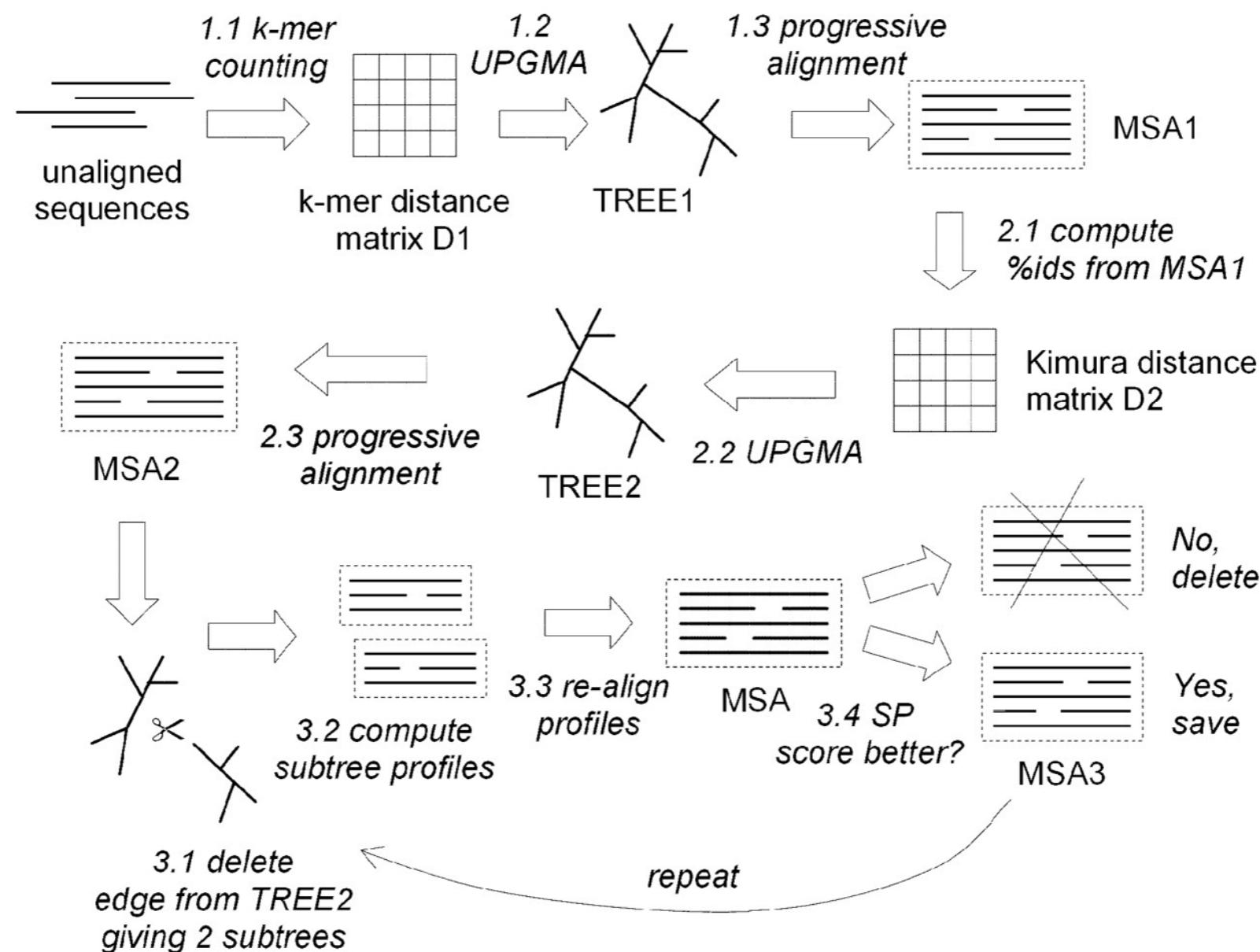
- use alignment to compute more accurate pairwise distance between sequences, Kimura distance:  $-\ln(1 - D - D^2/5)$ , where D is the fraction of identical bases between the pair of sequences.
- from new distance matrix, build the guide tree and a new alignment.

## refinement: tries to improve alignment

refines multiple alignment using the tree-dependent restricted partition technique - a process of deleting edges of guide tree, and re-combine the alignment of the disjoint trees, if better.

# Muscle

The first guide tree is not based on pairwise alignment, but on the comparison between the vectors containing the k-mer compositions of each sequence (faster)



# Alignment with Muscle

Calculate the alignment of five sequences of the Cytochromes used before with Muscle

sp|Q93VA3|CYC6\_ARATH  
sp|P99999|CYC\_HUMAN  
sp|P00004|CYC\_HORSE  
sp|P0C0X8|CYC2\_RHOSH  
sp|P00091|CYC22\_RHOA

MRLVLSGASSFTSNLFCSSQQVNNGRGKELKNPISLNHNKDLDFLKKLAPPLTAVLLAVS  
-----MG-----  
-----MG-----  
-----QEG-----  
-----MVKLLTILSIAATAGSLSIGTASAQ-----  
\*

sp|Q93VA3|CYC6\_ARATH  
sp|P99999|CYC\_HUMAN  
sp|P00004|CYC\_HORSE  
sp|P0C0X8|CYC2\_RHOSH  
sp|P00091|CYC22\_RHOA

PICFPPESLGQTLDIQRGATLFNRACIGCH-----DTGGNI-----  
-----DVEKGKKIFIMKCSQCH-----TVEKGGKHKTGPNLHGLFGRKTG  
-----DVEKGKKIFVQKCAQCH-----TVEKGGKHKTGPNLHGLFGRKTG  
-----DPEAGAKAFNQ-CQTCHVIVDDSGTTIAGRNAKTGPNLGVVGRTAG  
-----DAKAGEAVFKQ-CMTCH-----RADKNMVGPALGGVVGRKAG  
\* : \* \* \* \*\* . \* :

sp|Q93VA3|CYC6\_ARATH  
sp|P99999|CYC\_HUMAN  
sp|P00004|CYC\_HORSE  
sp|P0C0X8|CYC2\_RHOSH  
sp|P00091|CYC22\_RHOA

IOPGATLFTKDLER---NGVDTEEEIYRVTYFGKGRM-----PGFGEKCTPRGQCTF  
QAPGYS-YTAANKN---KGIIWGEDTL-MEYLENPKKYI-----PG-----TKMIF  
QAPGFT-YTDANKN---KGITWKEETL-MEYLENPKKYI-----PG-----TKMIF  
TQADFKGYGEGMKEAGAKGLAWDEEHF-VQYVQDPTKFL-----KEYTGDAKAKGKMTF  
TAAGFT-YSPLNHNSGEAGLVWTADNI-INYLNDPNAFLKKFLDKGKADQAVGVTKMTF  
... : \*: : \*.. : \* :

sp|Q93VA3|CYC6\_ARATH  
sp|P99999|CYC\_HUMAN  
sp|P00004|CYC\_HORSE  
sp|P0C0X8|CYC2\_RHOSH  
sp|P00091|CYC22\_RHOA

-GPRLQDEEIKLLAEFVKFQADQGWPTVSTD  
VGICKKKEERADLIAYLKKATNE-----  
AGIKKKTEREDLIAYLKKATNE-----  
-KLKKEADAHNIWAYLQQVAVRP-----  
-KLANEQQRKDVVAYLATLK-----  
: : .. \* :

# Consistency

For any multiple alignment, the induced pairwise alignments are necessarily consistent;

given a multiple alignment containing three sequences  $x$ ,  $y$ , and  $z$ , if position  $x_i$  aligns with position  $z_k$  in the projected  $x-z$  alignment and position  $z_k$  aligns with  $y_j$  in the projected  $z-y$  alignments, then  $x_i$  must align with  $y_j$  in the projected  $x-y$  alignment.

Consistency-based techniques apply this principle in reverse, using evidence from intermediate sequences to guide the pairwise alignment of  $x$  and  $y$ , such as needed during the steps of a progressive alignment.

# Transitive Relation

In mathematics, a binary relation  $R$  over a set  $X$  is transitive if whenever an element **a** is related to an element **b**, and **b** is in turn related to an element **c**, then **a** is also related to **c**.

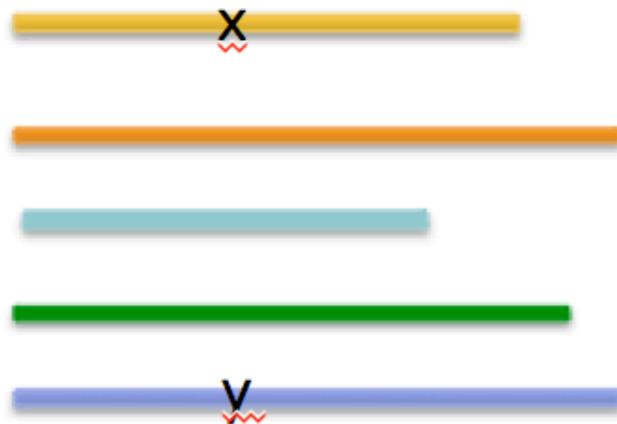
$$\forall a,b,c \in X : (aRb \wedge bRc) \Rightarrow aRc$$

# Transitivity in Alignments

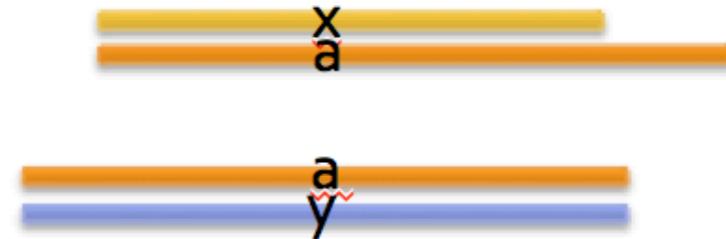
$$\forall a, b, c \in X : (aRb \wedge bRc) \Rightarrow aRc$$

$$\forall x, y, z \in \text{alned} : (x \text{Aln } z \wedge z \text{Aln } y) \Rightarrow x \text{Aln } y$$

Multiple Sequence Alignment

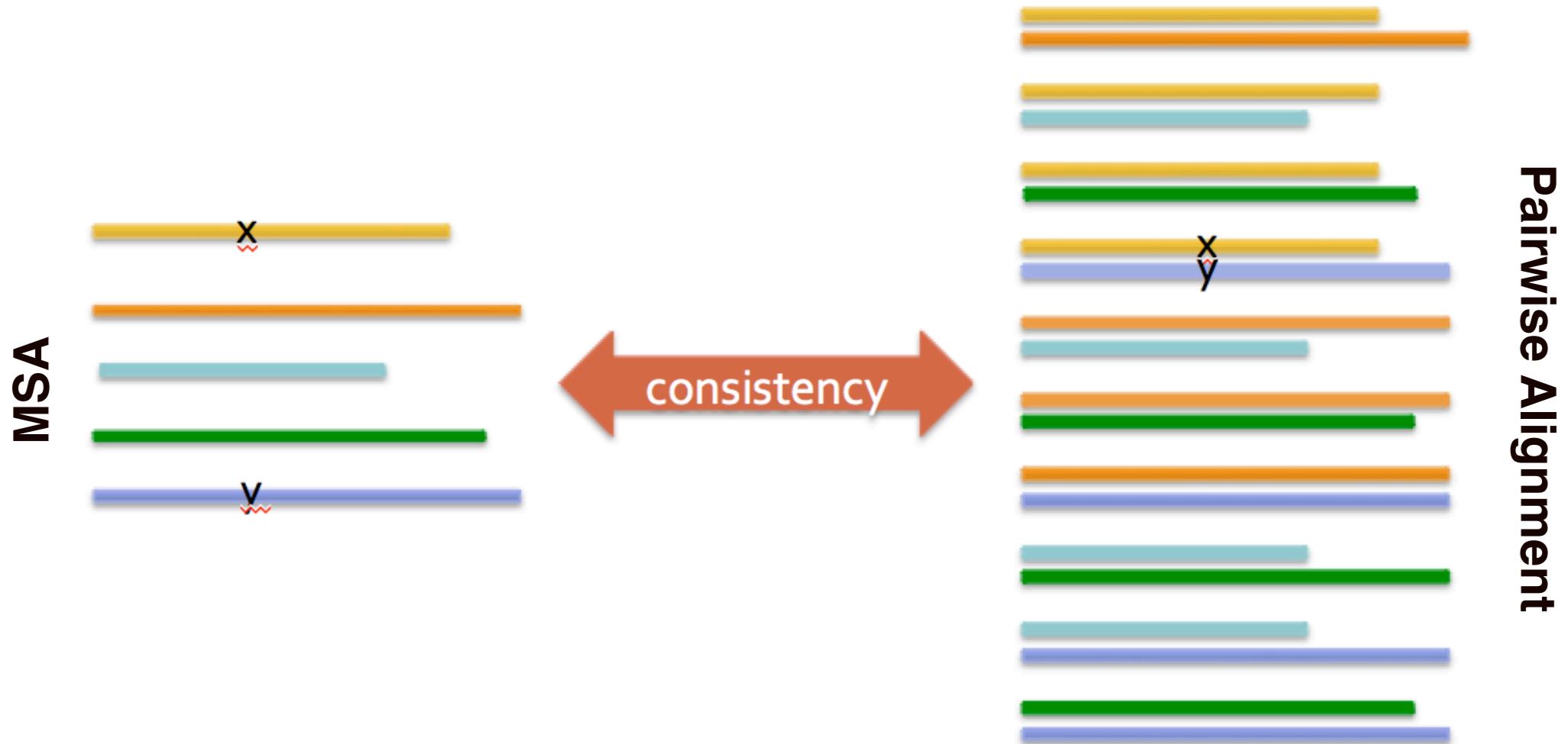


Pairwise Alignment



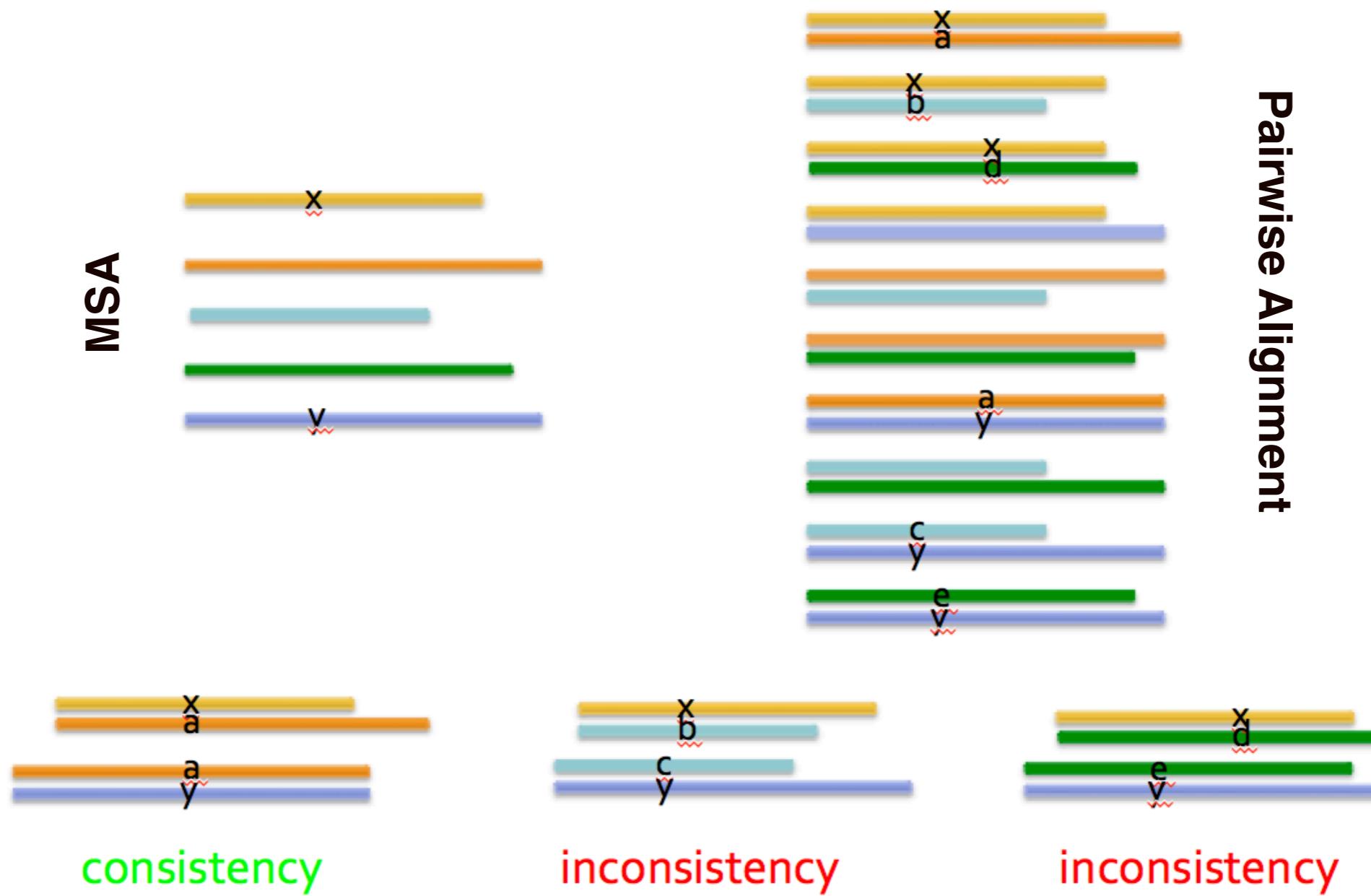
# How it can be applied?

Consistency between MSA & pairwise alignment : 0/1  
How can we increase the resolution of confidence?



# Consistent Alignments

The information are in the pairwise alignments



# How to Improve?

- MSA from **progressive alignments** can be largely inconsistent with respect to the set of pairwise alignments used to build the guide tree
- Consistency-based methods try **to build the tree in a more consistent way**

# T-Coffee

Tree-based Consistency Objective Function for alignment Evaluation

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

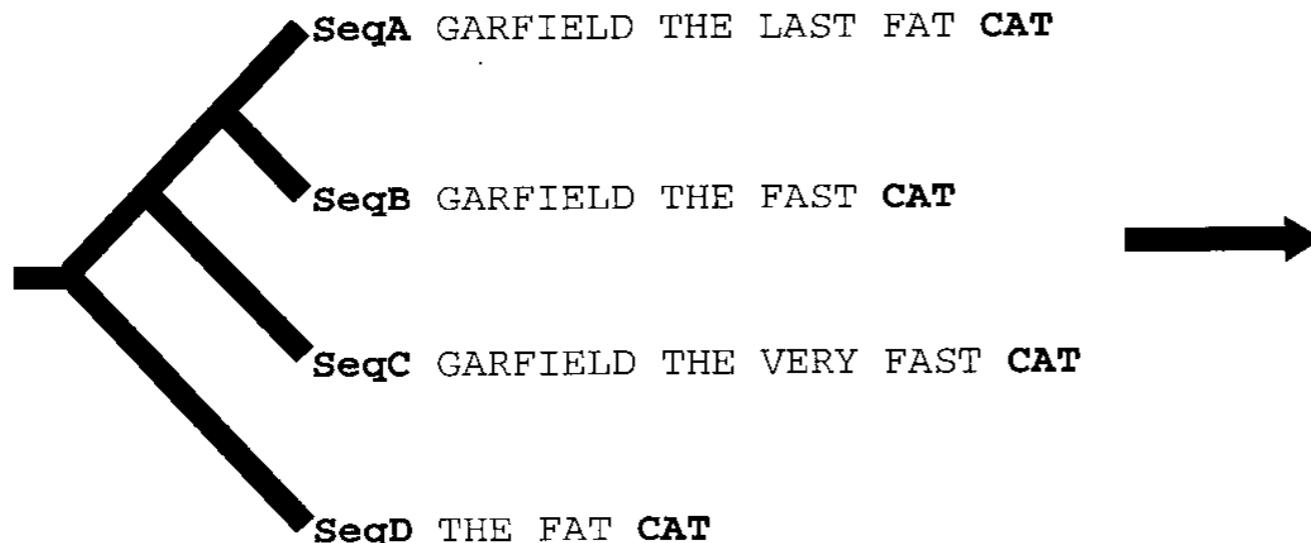
**SeqA** GARFIELD THE LAST FA-T CAT    Prim. Weight = 77  
**SeqC** GARFIELD THE VERY FAST CAT

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

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

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

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



**SeqA** GARFIELD THE LAST FA-T **CAT**  
**SeqB** GARFIELD THE FAST **CA-T** ---  
**SeqC** GARFIELD THE VERY FAST **CAT**  
**SeqD** ----- THE ---- FA-T **CAT**

This would be the ClustalW alignment of the four sequences.  
**CAT** is evidently misaligned

# T-Coffee - Start

The T-Coffee strategy starts from pairwise alignments as well.

Each pair of aligned residues is associated with a weight equal to the average identity among matched residues (gapped positions are neglected).

Identity values are used instead of alignment scores.

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

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

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

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

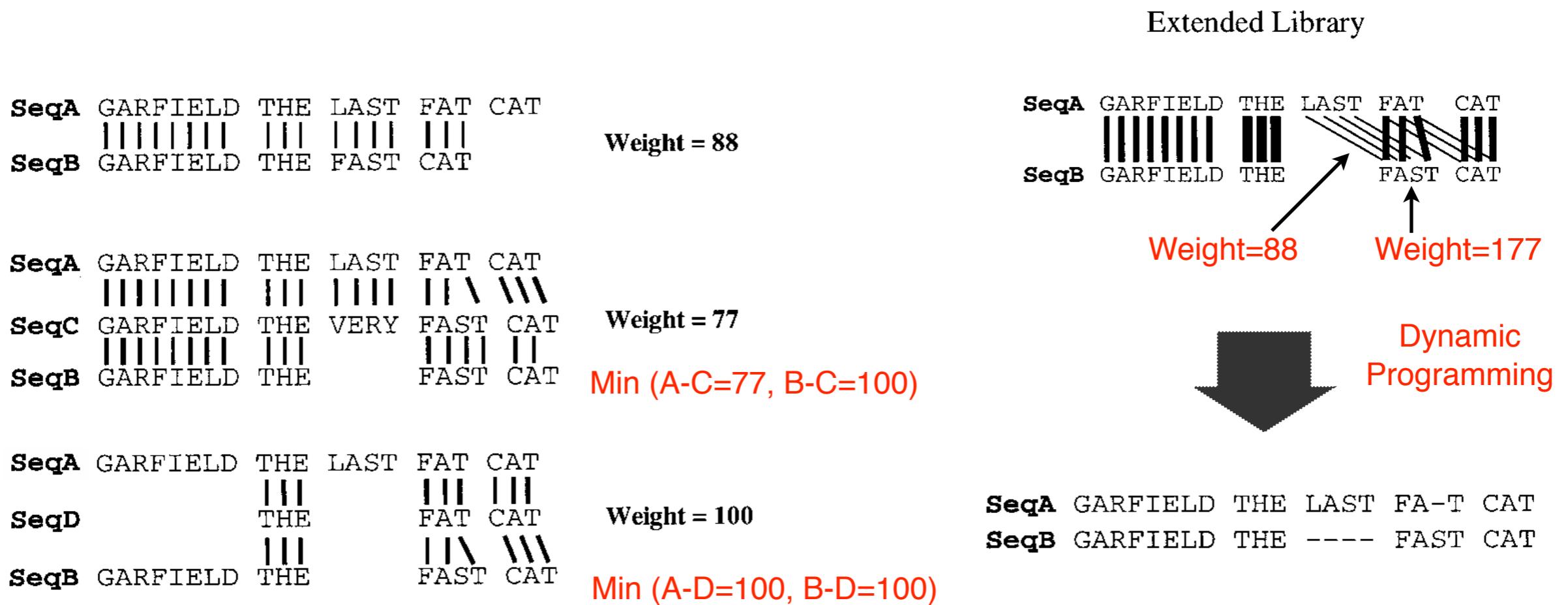
**SeqB** GARFIELD THE FAST CAT      Prim. Weight = 100  
**SeqD** ----- THE FA-T CAT

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

# Extended Library

In order to align sequence A and B, the three possible alignments are considered (A and B, A and B through C, A and B through D).

The weight associated to each alignment is the minimum of the weight associated to the pairwise alignments

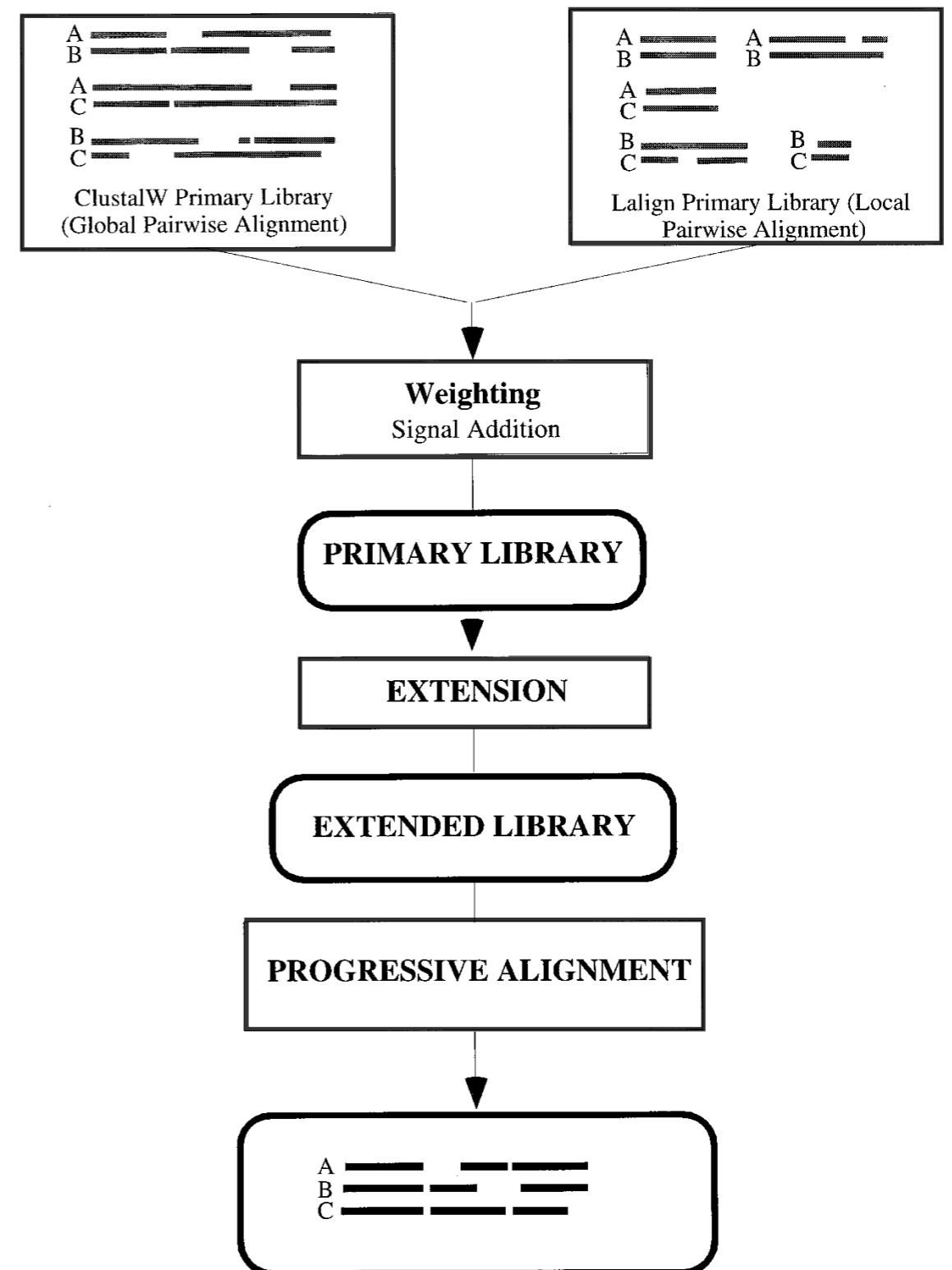


# T-Coffee Flowchart

The extended pairwise alignments are used to build a guide tree and progressive alignment procedure is then applied.

T-Coffee considers both global and local pairwise alignments.

It can also add supplementary information (domain, motifs....)



# Alignment with T-Coffee

Calculate the alignment of five sequences of the Cytochromes used before with T-Coffee

CLUSTAL W (1.83) multiple sequence alignment

sp P00004 CYC_HORSE	-----
sp P00091 CYC22_RHOPA	MVK-----
sp P0C0X8 CYC2_RHOSH	Q-----
sp P99999 CYC_HUMAN	-----
sp Q93VA3 CYC6_ARATH	MRLVLSGASSFTSNLFCSSQQVNNGRGKELKNPISLNHNKDLDFLKKLAPPLTAVLLAVS

sp P00004 CYC_HORSE	-----MGDVEKGKKIFVQKCAQ <b>CH</b> TVE-----
sp P00091 CYC22_RHOPA	-----LSIGTASAQDAKAGEAVFK-Q <b>CMT</b> <b>CH</b> R-----
sp P0C0X8 CYC2_RHOSH	-----EGDPEAGAKAFN-Q <b>CQT</b> <b>CH</b> VIVDDSGTTIAGRNAKTGPNLGVVGRT-----
sp P99999 CYC_HUMAN	-----MGDVEKGKKIFIMK <b>C</b> SQ <b>CH</b> TVE-----
sp Q93VA3 CYC6_ARATH	PICFPPESLG--QTLDIQRGATLFNRAC <b>I</b> G <b>CH</b> D <b>T</b> G-----GNIIQPG-----*

sp P00004 CYC_HORSE	TGQAPGFT-YTDANKN---KGITWKEETL-MEYLENPKKYI-----PGTK <b>M</b>
sp P00091 CYC22_RHOPA	AGTAAGFT-YSPLNHNSGEAGLVWTADNI-INYLNDPNAFLKKFLDKGKADQAVGVT <b>KM</b>
sp P0C0X8 CYC2_RHOSH	AGTQADFKGYGEGMKEAGAKGLAWDEEHF-VQYVQDPTKFLKEYTG-----DAKAKG <b>KM</b>
sp P99999 CYC_HUMAN	TGQAPGYS-YTAANKN---KGIIWGEDTL-MEYLENPKKYI-----PGTK <b>M</b>
sp Q93VA3 CYC6_ARATH	-----ATLFT---KDLERNGVDTEEEIYRVTYFGK--GR <b>M</b> PGFGE-----KCTPRGQC-----:

sp P00004 CYC_HORSE	IFAGIKKKTEREDLIAYLKK-----ATNE
sp P00091 CYC22_RHOPA	TFK-LANEQQRKVTVVAYL-----ATLK
sp P0C0X8 CYC2_RHOSH	TFK-LKKEAADHNIWAYLQQ-----VAVRP
sp P99999 CYC_HUMAN	IFVG IKKKERADLIAYLKK-----ATNE
sp Q93VA3 CYC6_ARATH	TFG-PRLQDEEIKLLAEFVKFQADQGWPTVSTD-----*

# Alignment Benchmark

**BAliBASE** was the first large scale benchmark specifically designed for MSA, providing high quality manually refined reference **alignments based on 3D structure superpositions**.

**BAliBASE** is divided into several reference datasets:

1. cases with small numbers of equidistant sequences, and was further subdivided by percent identity;
2. families with one or more “orphan” sequences;
3. a pair of divergent subfamilies, with less than 25% identity between the two groups;
4. sequences with large terminal extensions (N/C-terminal);
5. sequences with large internal insertions and deletions.

# Benchmark Evaluation

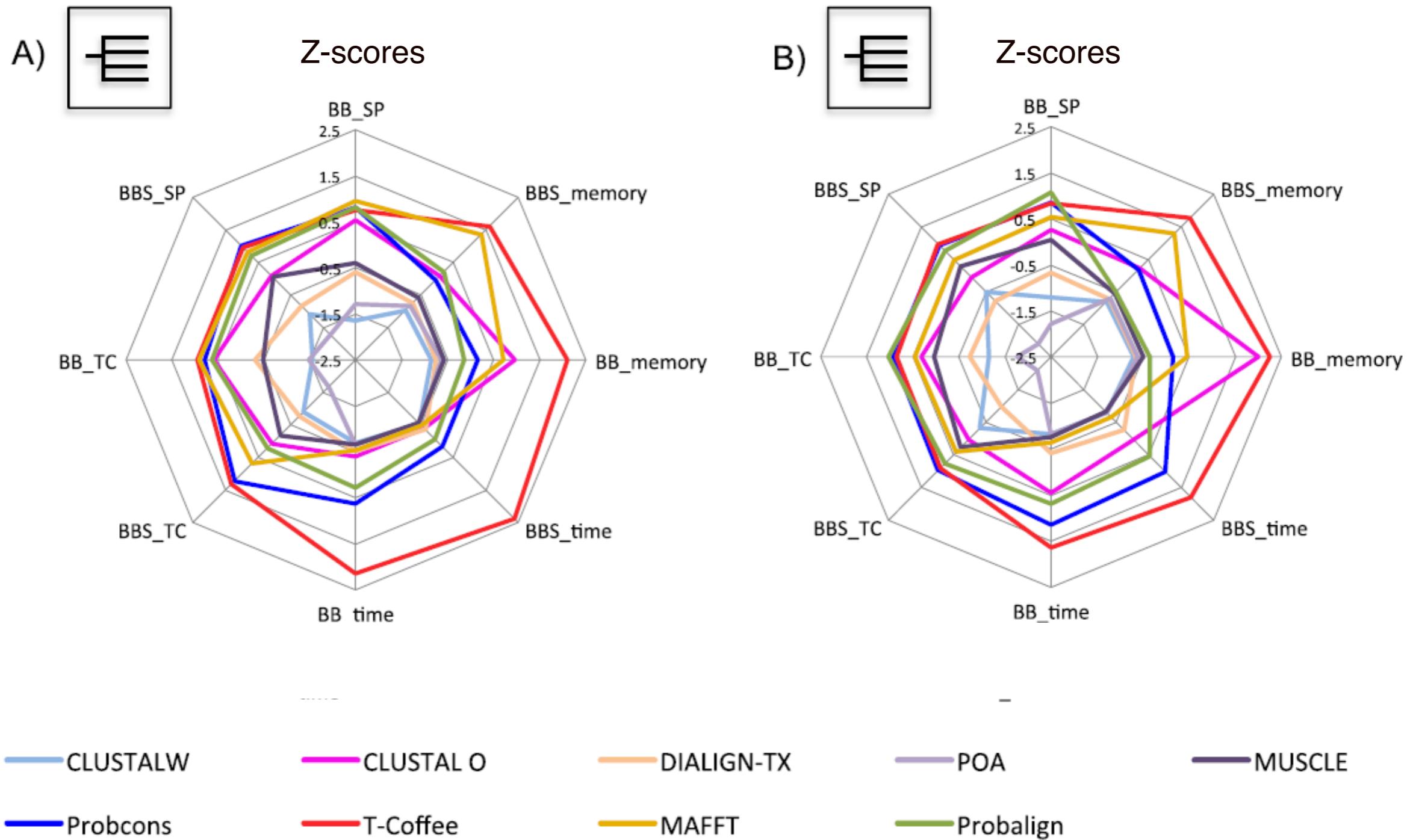
SP score determines the extent to which the programs succeed in aligning input sequences in an MSA. It is calculated as the ratio of the sum of scores  $p$  for all pairs of residues in every column of the alignment by the sum of scores in the reference alignment;  $p = 1$  if the pair of compared residues is aligned identically in the reference alignment, otherwise  $p = 0$ .

The TC score is calculated considering the ratio of the sum of scores  $c$  by the number of columns in the alignment, being  $c = 1$  if all residues in the column are aligned identically in the reference alignment, otherwise  $c = 0$

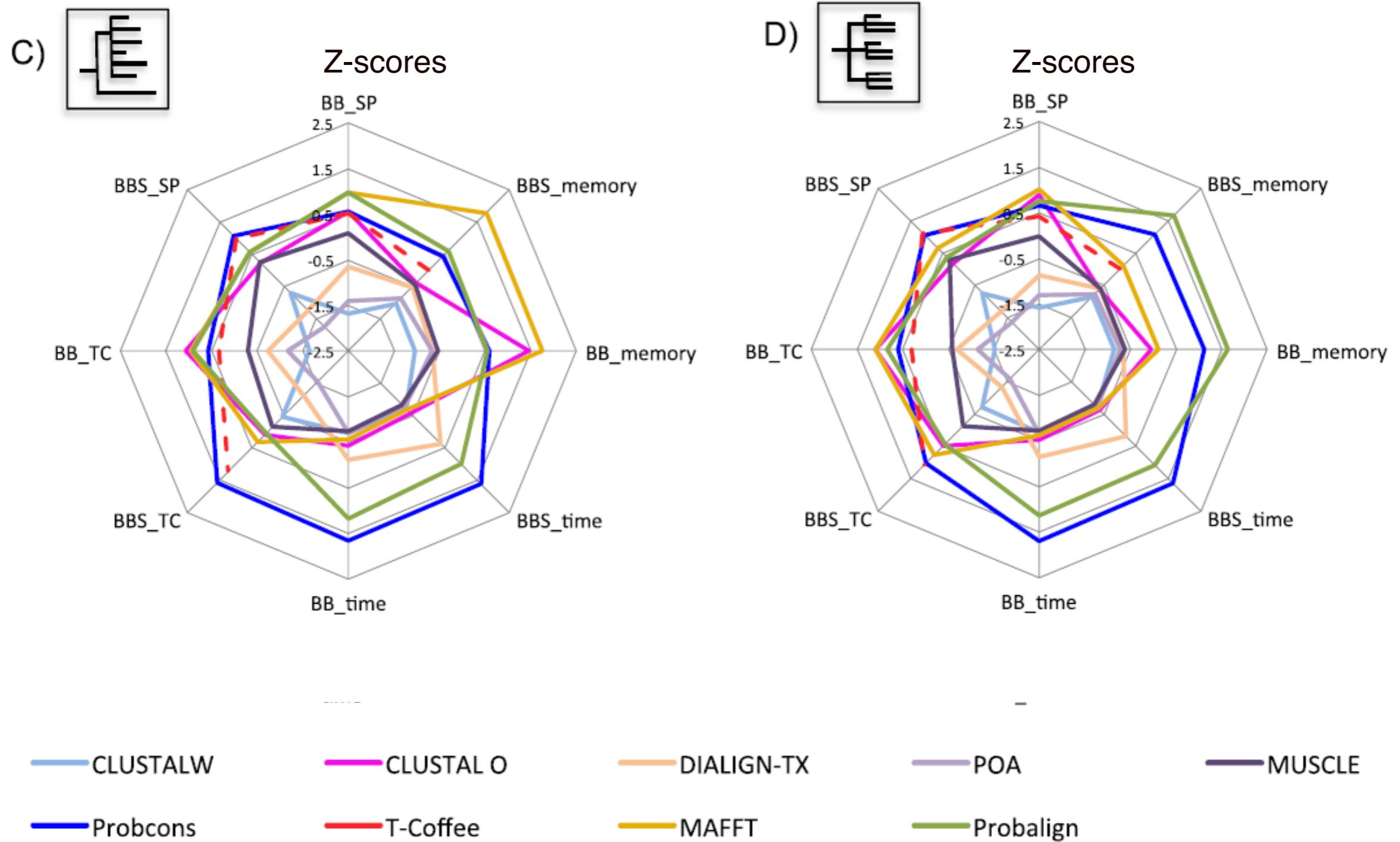
Time of execution

Peak memory

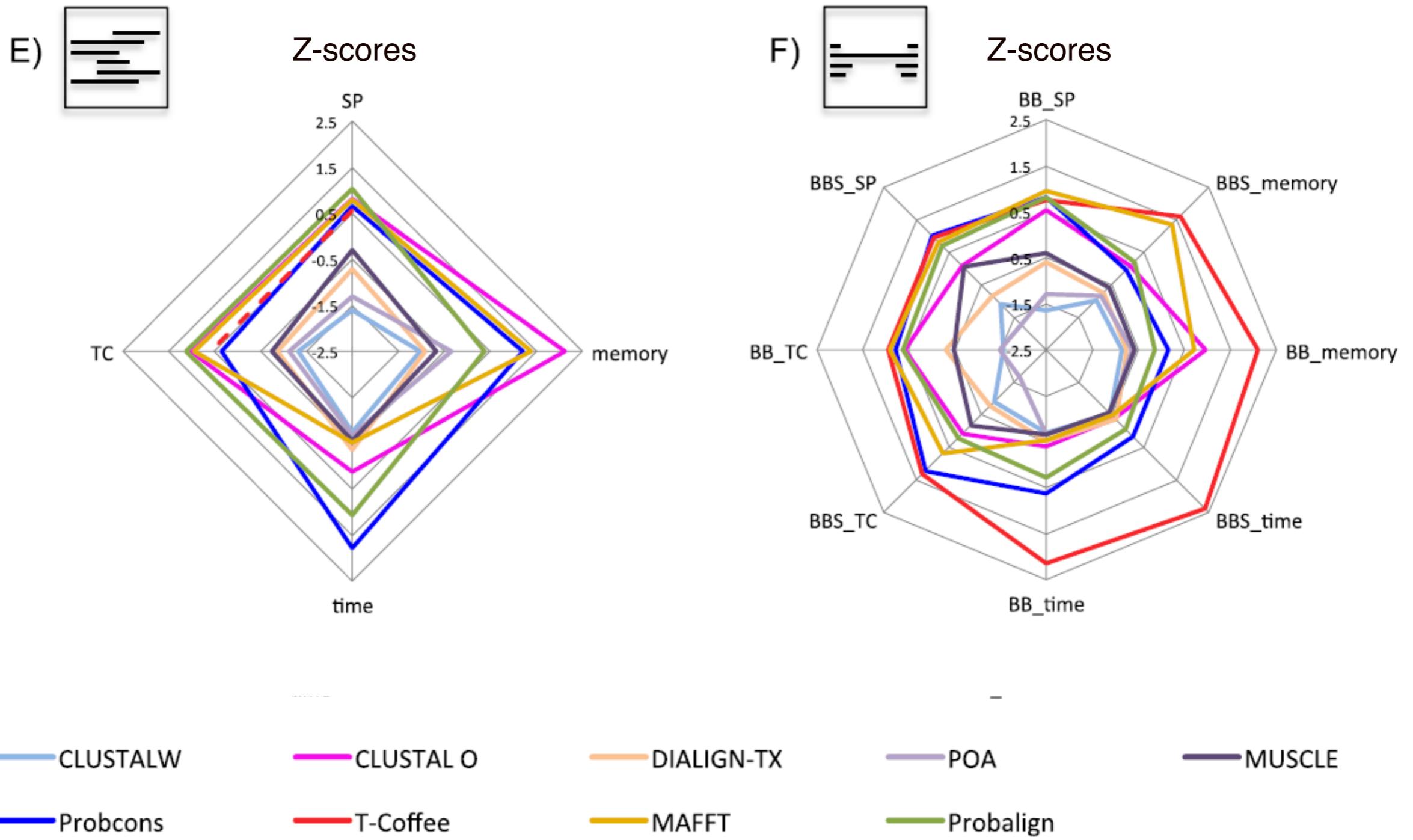
# Performance (I)



# Performance (II)



# Performance (III)



# Results and Conclusions

**Results:** Our results indicate that mostly the consistency-based programs Probcons, T-Coffee, Probalign and MAFFT outperformed the other programs in accuracy. Whenever sequences with large N/C terminal extensions were present in the BALiBASE suite, Probalign, MAFFT and also CLUSTAL OMEGA outperformed Probcons and T-Coffee. The drawback of these programs is that they are more memory-greedy and slower than POA, CLUSTALW, DIALIGN-TX, and MUSCLE. CLUSTALW and MUSCLE were the fastest programs, being CLUSTALW the least RAM memory demanding program.

**Conclusions:** Based on the results presented herein, all four programs Probcons, T-Coffee, Probalign and MAFFT are well recommended for better accuracy of multiple sequence alignments. T-Coffee and recent versions of MAFFT can deliver faster and reliable alignments, which are specially suited for larger datasets than those encountered in the BALiBASE suite, if multi-core computers are available. In fact, parallelization of alignments for multi-core computers should probably be addressed by more programs in a near future, which will certainly improve performance significantly.

# Exercise

Download from UniProtKB the sequences of the following proteins (in FASTA format)

P99999 (human)

P00004 (horse)

P0C0X8 (Rhodobacter)

P00091 (Rhodopseudomonas)

Q93VA3 (Arabidopsis)

Align with ClustalW @

<http://clustalw.ddbj.nig.ac.jp/>

<http://www.ch.embnet.org/software/ClustalW.html>

Write a script to calculate the information entropy of the MSA and for each column the most conserved residue and its frequency.

# Exercise

Using the BLAST tool at Uniprot, retrieve all the SwissProt sequences that are similar with an E-value <0,001 to the Rhodopseudomonas cytochrome C (P00091) .

Download the sequences in Fasta format and align with ClustalW, Muscle or T-Coffee

Analyse the conserved positions in the alignments

Repeat with the Arabidopsis (Q93VA3) and the human (P99999) sequences

Compare the results, an in particular the pattern of conserved residues