Multiple Alignments

Laboratory of Bioinformatics I
Module 2

Emidio Capriotti
http://biofold.org/

Biomolecules Folding and Disease

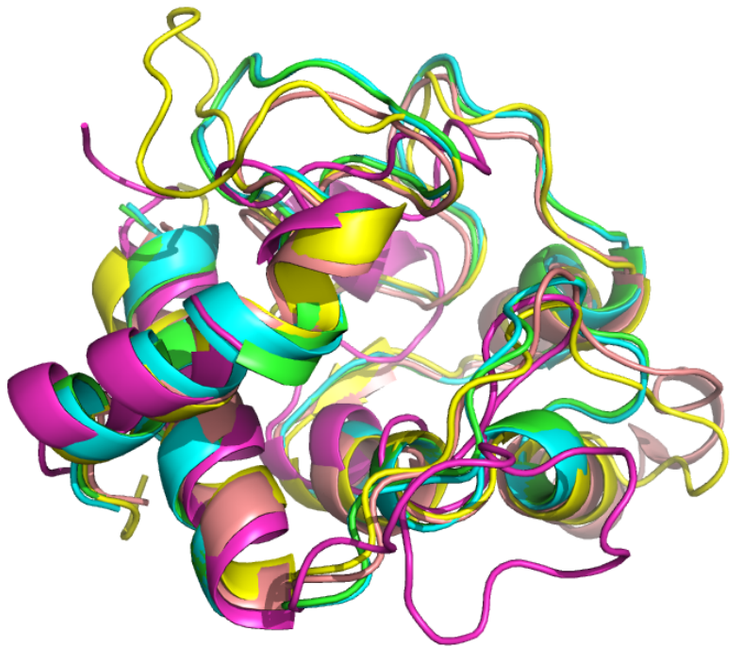
Department of Pharmacy and Biotechnology (FaBiT)
University of Bologna

Multiple Structure Alignment

Align the structure of 5 structures of Cytochromes

```
3zcf:A --GDVEKGKKIFIMKCsQCHTVEkgg-------khKTG--PNLHG--LfgRKTgqapgysyt----aank
3o20:A --GDVEKGKKIFVQKCaQCHTVEkgg-------khKTG--PNLHG--LfgRKTgqapgftyt----dank
2ce0:A --LDIQRGATLFNRACaACHDTGg----------nIIQpgATLFTkdL--ERN----------------
1cxc:A qeGDPEAGAKAFNQCQ-TCHVIVddsgttiagrnaKTG--PNLYG--VvgRTAgtqadfkgygegmkeag
1i8o:A --EDAKAGEAVFKQCM-TCHRADk----------nMVG--PALAG--VvgRKAgtaagftysp-lnhnsg

3zcf:A nkgiIW-GEDTLMEYLENPKkyi-------------pgTKMiFvGiK---------------KKEERAD
3o20:A nkgiTW-KEETLMEYLENPKkyi-------------pgTKMiFaGiK---------------KKTERED
2ce0:A ----GVdTEEEIYRVTYFGK-------------------GRM-PgF-GekctprgqctfgprlQDEEIKL
1cxc:A akglAW-DEEHFVQYVQDPTkflkeyt------gdakakGKMtF-KlK---------------KEADAHN
1i8o:A eaglVW-TADNIVPYLADPNaflkkfltekgkadqavgvTKMtF-KlA---------------NEQQRKD

3zcf:A LIAYLKKATne----
3o20:A LIAYLKKATne----
2ce0:A LAEFVKFQAdqgwpt
1cxc:A IWAYLQQVAvrp---
1i8o:A VVAYLATLK------
```

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.

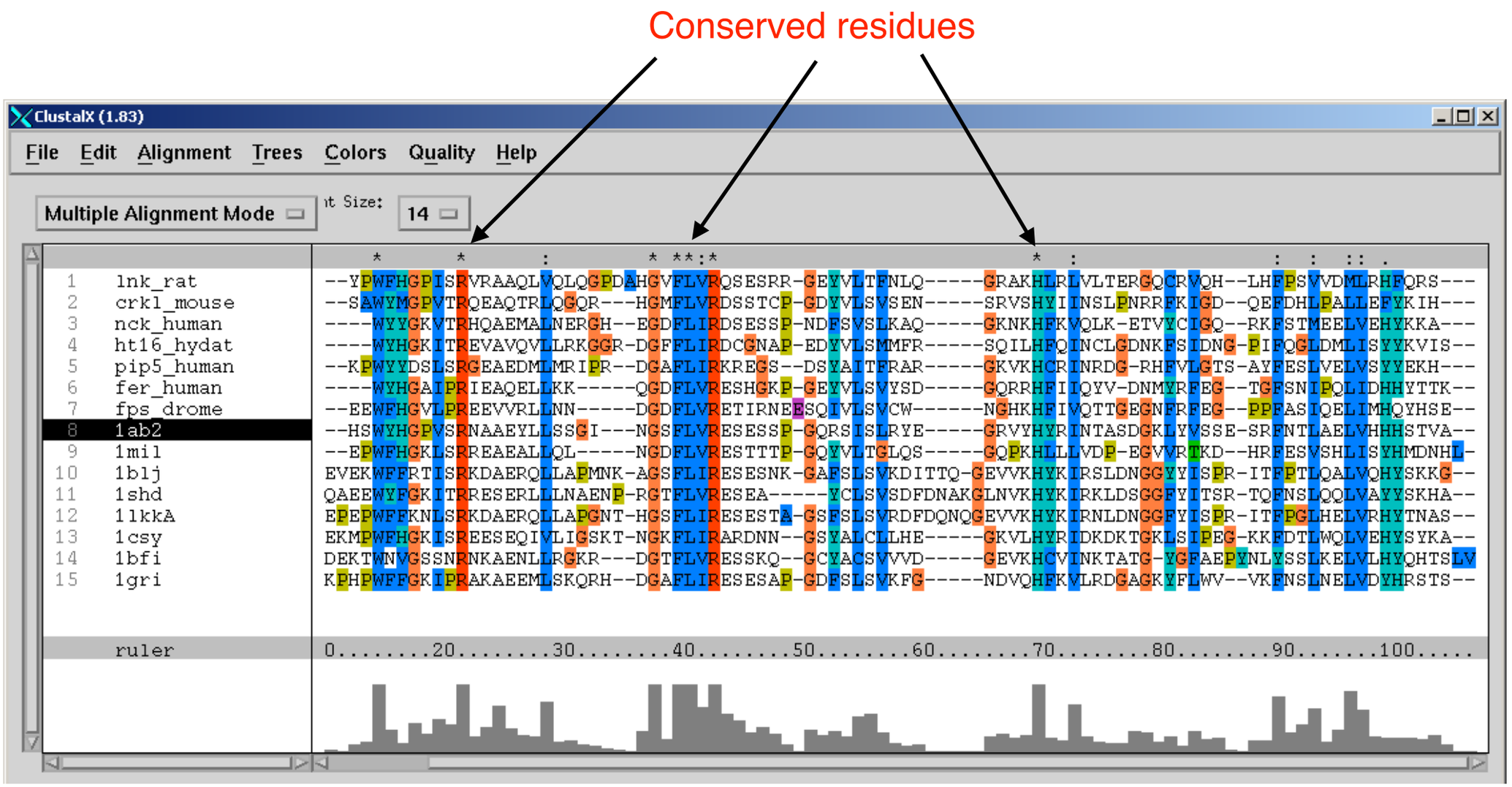
Part of an alignment of SH2 domains from 14 sequences

```
                  *     *     :       * **:*                          *  :                :  :  ::
lnk_rat     ----------YPWFHGPISRVRAAQLVQLQGPDAHGVFLVRQSESRR-GEYVLTFNLQ-----GRAKHLRLVLTERGQCRVQH--LHFPSVVDML
crk1_mouse  ----------SAWYMGPVTRQEAQTRLQGQR---HGMFLVRDSSTCP-GDYVLSVSEN-----SRVSHYIINSLPNRRFKIGD--QEFDHLPALL
nck_human   ----------WYYGKVTRHQAEMALNERGH--EGDFLIRDSESSP-NDFSVSLKAQ-----GKNKHFKVQLK-ETVYCIGQ--RKFSTMEELV
ht16_hydat  ----------WYHGKITREVAVQVLLRKGGR-DGFFLIRDCGNAP-EDYVLSMMFR-----SQILHFQINCLGDNKFSIDNG-PIFQGLDMLI
pip5_human  ----------KPWYYDSLSRGEAEDMLMRIPR--DGAFLIRKREGS--DSYAITFRAR-----GKVKHCRINRDG-RHFVLGTS-AYFESLVELV
fer_human   ----------WYHGAIPRIEAQELLKK-----QGDFLVRESHGKP-GEYVLSVYSD-----GQRRHFIIQYV-DNMYRFEG--TGFSNIPQLI
1ab2        ----------EEWFHGVLPREEVVRLLNN-----DGDFLVRETIRNEESQIVLSVCW------NGHKHFIVQTTGEGNFRFEG--PPFASIQELI
1mil        ----------HSWYHGPVSRNAAEYLLSSGI---NGSFLVRESESSP-GQRSISLRYE-----GRVYHYRINTASDGKLYVSSE-SRFNTLAELV
1blj        ----------EPWFHGKLSRREAEALLQL-----NGDFLVRESTTTP-GQYVLTGLQS-----GQPKHLLLVDP-EGVVRTKD--HRFESVSHLI
1shd        GSVAPVETLEVEKWFFRTISRKDAERQLLAPMNK-AGSFLIRESESNK-GAFSLSVKDITTQ-GEVVKHYKIRSLDNGGYYISPR-ITFPTLQALV
1lkkA       -------SIQAEEWYFGKITRRESERLLLNAENP-RGTFLVRESEA-----YCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSR-TQFNSLQQLV
1csy        --------LEPEPWFFKNLSRKDAERQLLAPGNT-HGSFLIRESESTA-GSFSLSVRDFDQNQGEVVKHYKIRNLDNGGFYISPR-ITFPGLHELV
1bfi        -------SHEKMPWFHGKISREESEQIVLIGSKT-NGKFLIRARDNN--GSYALCLLHE-----GKVLHYRIDKDKTGKLSIPEG-KKFDTLWQLV
1gri        -------HHDEKTWNVGSSNRNKAENLLRGKR---DGTFLVRESSKQ--GCYACSVVVD-----GEVKHCVINKTATG-YGFAEPYNLYSSLKELV
1gri        -------EMKPHPWFFGKIPRAKAEEMLSKQRH--DGAFLIRESESAP-GDFSLSVKFG-----NDVQHFKVLRDGAGKYFLWV--VKFNSLNELV
```

* conserved identical residues
: conserved similar residues

Why is important?

Multiple sequence alignment is important to find conserved residues



conservation profile

Sequence Profile

A multiple sequence alignment can be represented by sequence profile

MSA

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Y	**K**	**D**	**Y**	**H**	**S**	**-**	**D**	**K**	**K**	**K**	**G**	**E**	**L**	**-**	**-**
2	Y	R	D	Y	Q	T	-	D	Q	K	K	G	D	L	-	-
3	Y	R	D	Y	Q	S	-	D	H	K	K	G	E	L	-	-
4	Y	R	D	Y	V	S	-	D	H	K	K	G	E	L	-	-
5	Y	R	D	Y	Q	F	-	D	Q	K	K	G	S	L	-	-
6	Y	K	D	Y	N	T	-	H	Q	K	K	N	E	S	-	-
7	Y	R	D	Y	Q	T	-	D	H	K	K	A	D	L	-	-
8	G	Y	G	F	G	-	-	L	I	K	N	T	E	T	T	K
9	T	K	G	Y	G	F	G	L	I	K	N	T	E	T	T	K
10	T	K	G	Y	G	F	G	L	I	K	N	T	E	T	T	K

Sequence Position

Sequence Profile

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	70	0	0	0	0	60	0	0	0	0	20	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	70	0	0	0
F	0	0	0	10	0	33	0	0	0	0	0	0	0	0	0	0
G	10	0	30	0	30	0	100	0	0	0	0	50	0	0	0	0
H	0	0	0	0	10	0	0	10	30	0	0	0	0	0	0	0
K	0	40	0	0	0	0	0	0	10	100	70	0	0	0	0	100
I	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0
N	0	0	0	0	10	0	0	0	0	0	30	10	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	40	0	0	0	30	0	0	0	0	0	0	0
R	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	33	0	0	0	0	0	0	10	10	0	0
T	20	0	0	0	0	33	0	0	0	0	0	30	0	30	100	0
V	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0
W	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	70	0	0	90	0	0	0	0	0	0	0	0	0	0	0	0

Sequence Logo

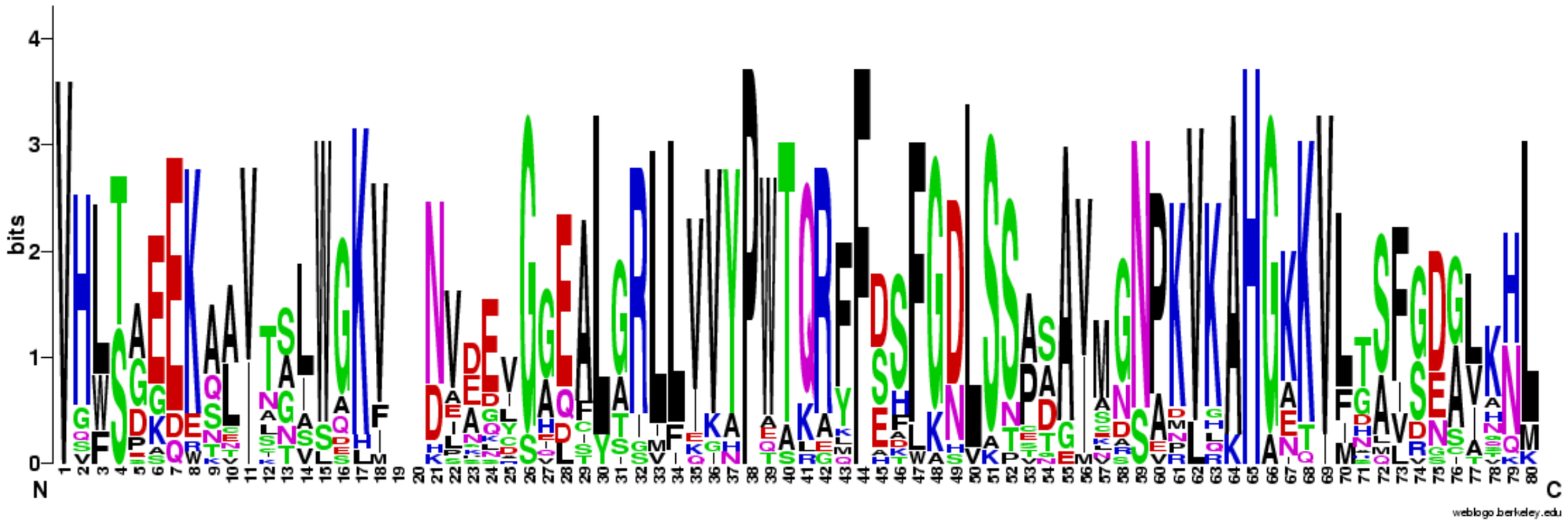
Plot drawn using a score derived from the information entropy (S)
calculating the information content (I) in each position

$$S(p) = \sum_{i=1}^{20} - p_i \log_2 p_i$$

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

$$h_i = p_i \times I \qquad h_i = \text{height } aa_i$$

$$e_n = \frac{k-1}{2n \times \ln 2} \qquad k = \#\ aa \text{ types} \qquad n = \#\ \text{sequences}$$



http://weblogo.berkeley.edu/

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

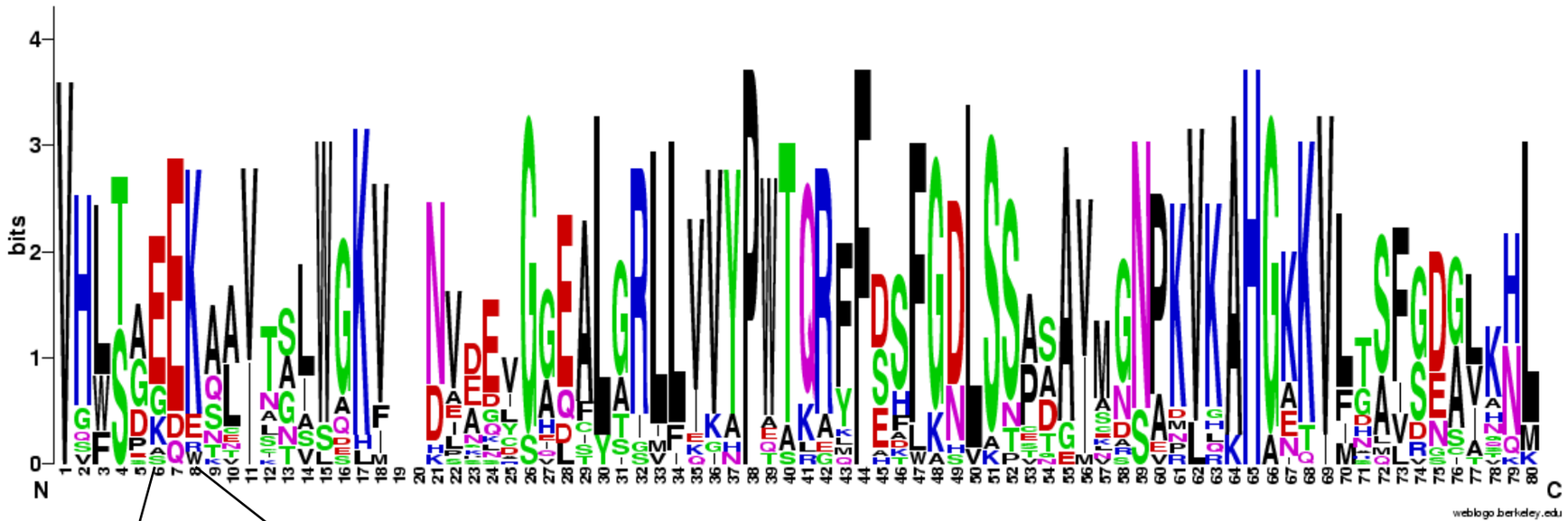
$$\boldsymbol{S} = \Sigma_{i<j}\, S(A_i, A_j)$$

$$S\begin{bmatrix} 1\text{>ASPTLPLSLA} \\ 2\text{>SS-TLPA--A} \\ 3\text{>SSPTLPA--A} \end{bmatrix} = S\begin{bmatrix} 1\text{>ASPTLPLSLA} \\ 2\text{>SS-TLPA--A} \end{bmatrix} + S\begin{bmatrix} 1\text{>ASPTLPLSLA} \\ 3\text{>SSPTLPA--A} \end{bmatrix} + S\begin{bmatrix} 2\text{>SS-TLPA--A} \\ 3\text{>SSPTLPA--A} \end{bmatrix}$$

Entropy Score

The multiple sequence alignment can be obtained minimizing the entropy

$$S = \sum_{j=1}^{Ncolums} \sum_{i=1}^{20} - p_{ji} \log_2 p_{ji}$$



The substitution score may depend on the position.

http://weblogo.berkeley.edu/

Profile-Based Alignment

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

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

S_i = “C”

= 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

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

	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	-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	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

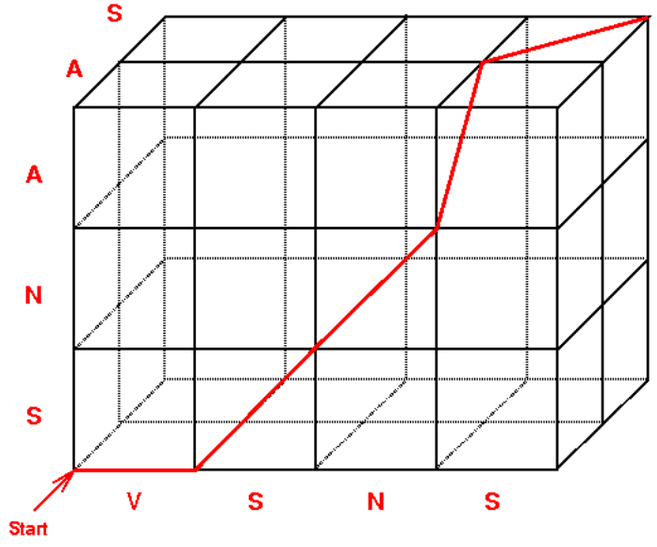
Iterative algorithms

Consistency-based algorithms

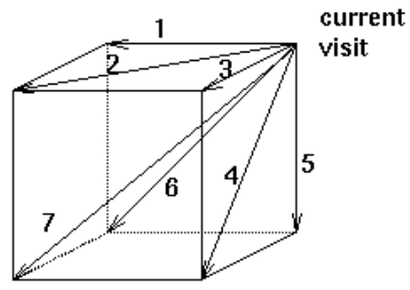
Optimal Alignment

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

Extension of dynamic programming for 2 sequences => N dimensions



V S N _ S
_ S N A _
_ _ _ A S



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?

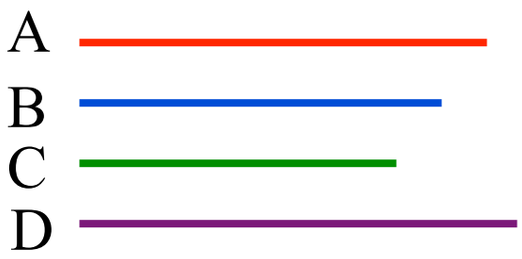
First align the most closely related sequences

How to measure the similarity of the sequences?

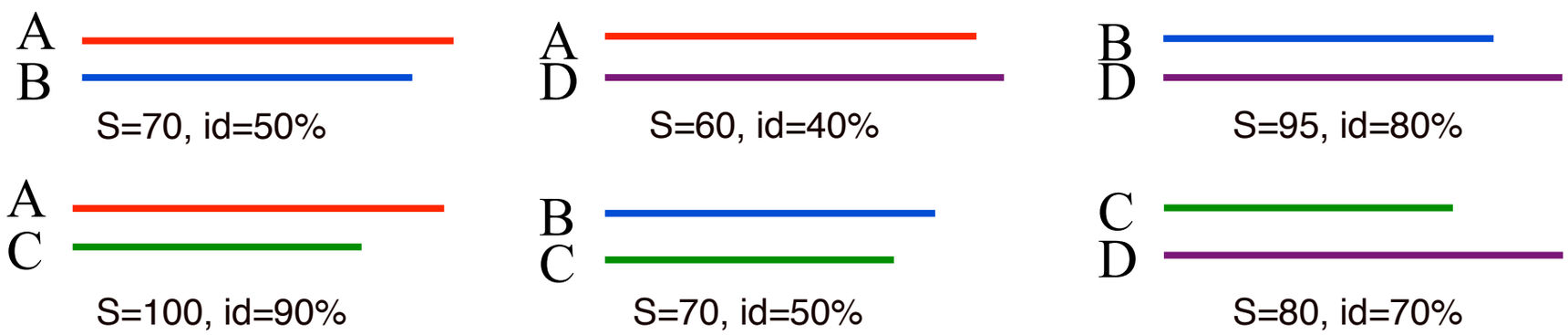
align all the sequences pairwise

calculate the similarity between each pair from the alignment

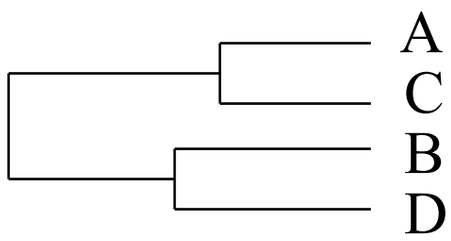
Progressive MSA - Start



Step 1: Pairwise sequence alignment: exact, all-against-all



Step 2: 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
$P^1_i = P^1_i(A)\ P^1_i(C)\ \ldots\ldots\ P^1_i(Y)$

The position j along the second sequence profile, it is represented by a 20-element vector
$P^2_j = P^1_j(A)\ P^1_j(C)\ \ldots\ldots\ P^1_j(Y)$

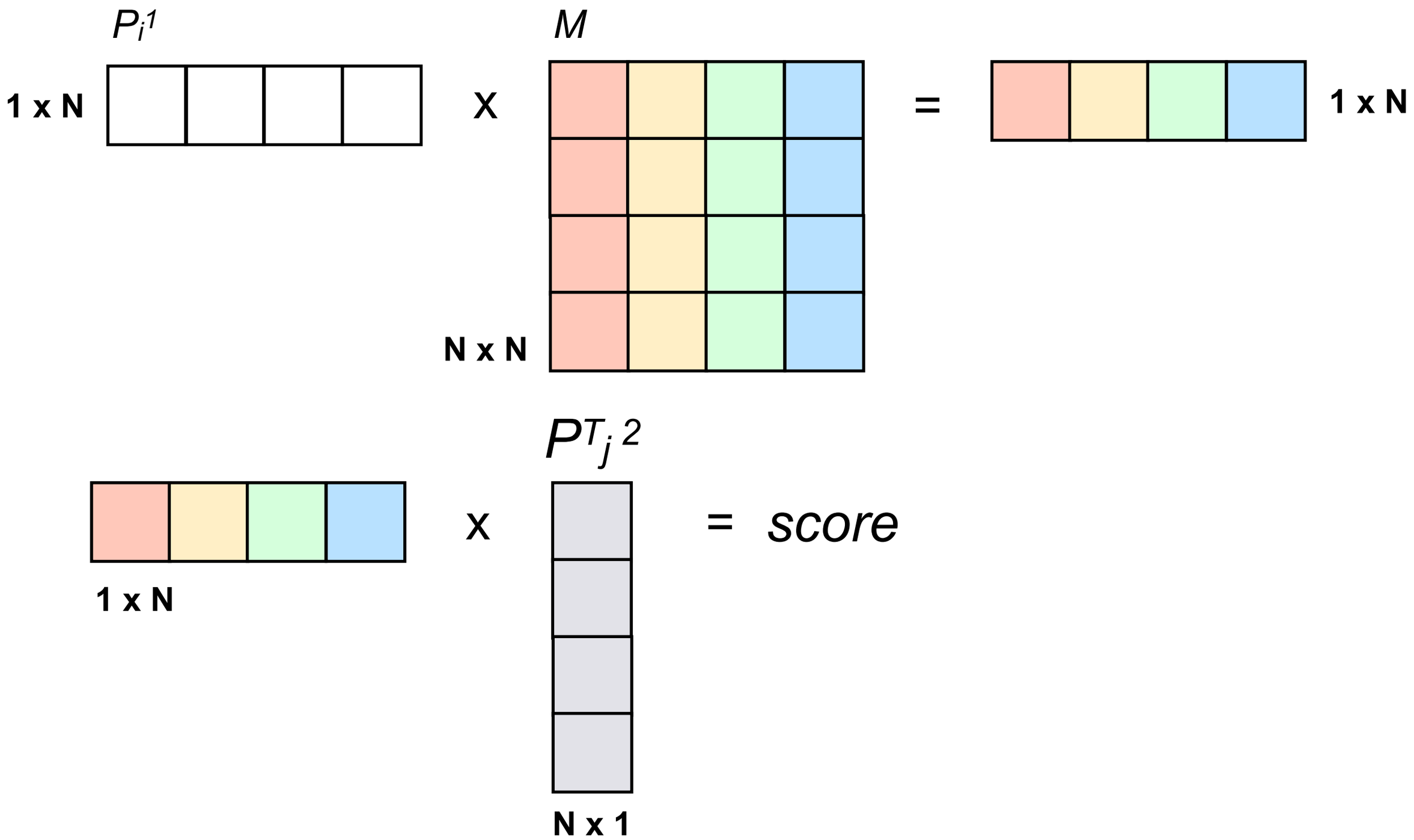
The score for aligning the two positions is:

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

where M is a matrix score (BLOSUM or PAM)
The score can be used in dynamic programming procedures (Needleman-Wunsch, Smith-Waterman)

Matrix representation

$$P_i^1 \times M \times P^T{}_j^2$$



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          -----------------------MGDVEKGKKIFIMKCS------QCHT 20
sp|P00004|CYC_HORSE          -----------------------MGDVEKGKKIFVQKCA------QCHT 20
sp|P0C0X8|CYC2_RHOSH         -----------------------QEGDPEAGAKAFNQCQTCHVIVDDSGT 27
sp|P00091|CYC22_RHOPA        ----------MVKKLLTILSIAATAGSLSIGTASAQDAKAGEAVFKQCMT 40
sp|Q93VA3|CYC6_ARATH_        MRLVLSGASSFTSNLFCSSQQVNGRGKELKNPISLNHNKDLDFLLKKLAP 50
                                                     *.   .                .  .

sp|P99999|CYC_HUMAN          VEKGGKHKTGPNLHG--LFGRKTGQAPGYS-YTAANKN---KGIIWGEDT 64
sp|P00004|CYC_HORSE          VEKGGKHKTGPNLHG--LFGRKTGQAPGFT-YTDANKN---KGITWKEET 64
sp|P0C0X8|CYC2_RHOSH         TIAGRNAKTGPNLYG--VVGRTAGTQADFKGYGEGMKEAGAKGLAWDEEH 75
sp|P00091|CYC22_RHOPA        CHRADKNMVGPALGG--VVGRKAGTAAGFT-YSPLNHNSGEAGLVWTADN 87
sp|Q93VA3|CYC6_ARATH_        PLTAVLLAVSPICFPPESLGQTLDIQRGATLFNRACIGCHDTGGNIIQPG 100
                                 .    ..*        .*:. .   . . :          *

sp|P99999|CYC_HUMAN          LMEYLENP-----------------KKYIPG----------TKMIFVGI 86
sp|P00004|CYC_HORSE          LMEYLENP-----------------KKYIPG----------TKMIFAGI 86
sp|P0C0X8|CYC2_RHOSH         FVQYVQDPTK---------------FLKEYTGD------AKAKGKMTFK-L 104
sp|P00091|CYC22_RHOPA        IINYLNDPNA---------------FLKKFLTDKGKADQAVGVTKMTFK-L 122
sp|Q93VA3|CYC6_ARATH_        ATLFTKDLERNGVDTEEEIYRVTYFGKGRMPGFG--EKCTPRGQCTFGPR 148
                                 : ::                       *    .          :  *

sp|P99999|CYC_HUMAN          KKKEERADLIAYLKKATNE-------- 105
sp|P00004|CYC_HORSE          KKKTEREDLIAYLKKATNE-------- 105
sp|P0C0X8|CYC2_RHOSH         KKEADAHNIWAYLQQVAVRP------- 124
sp|P00091|CYC22_RHOPA        ANEQQRKDVVAYLATLK---------- 139
sp|Q93VA3|CYC6_ARATH_        LQDEEIKLLAEFVKFQADQGWPTVSTD 175
                              :. :   :  ::
```

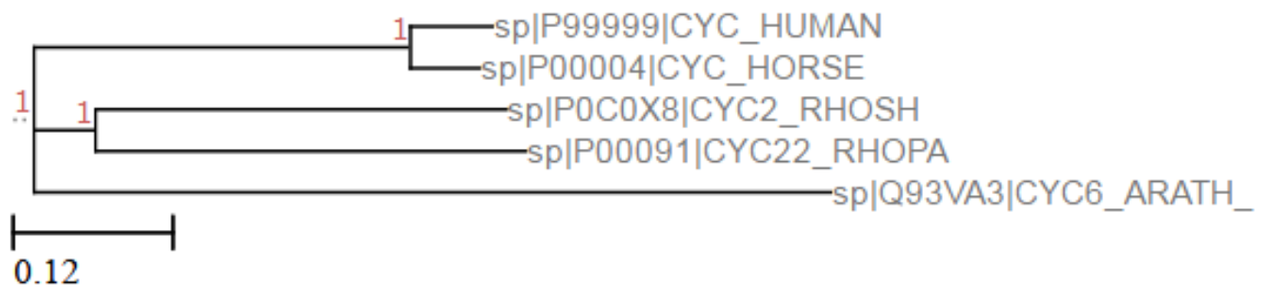
Alternative Alignment

How many conserved sites?

```
CLUSTAL 2.1 multiple sequence alignment


sp|P99999|CYC_HUMAN_Cytochrome      -MGDVEKGKKIFIMKCSQCHTVE-------KGGKHKTGPNLHGLFGRKTG 42
sp|P00004|CYC_HORSE_Cytochrome      -MGDVEKGKKIFVQKCAQCHTVE-------KGGKHKTGPNLHGLFGRKTG 42
sp|P0C0X8|CYC2_RHOSH_Cytochrom      QEGDPEAGAKAFN-QCQTCHVIVDDSGTTIAGRNAKTGPNLYGVVGRTAG 49
sp|P00091|CYC22_RHOPA_Cytochro      --QDAKAGEAVFK-QCMTCHR---------ADKN-MVGPALGGVVGRKAG 37
sp|Q93VA3|CYC6_ARATH_Cytochrom      QTLDIQRGATLFNRACIGCHDTG--------GNIIQPGATLFTKDLERNG 42
                                       * : *   *   *  **                .    *. *     .  *

sp|P99999|CYC_HUMAN_Cytochrome      QAPGYS-YTAANKN---KGIIWGEDTLMEYLENPKKYIP---------- 77
sp|P00004|CYC_HORSE_Cytochrome      QAPGFT-YTDANKN---KGITWKEETLMEYLENPKKYIP---------- 77
sp|P0C0X8|CYC2_RHOSH_Cytochrom      TQADFKGYGEGMKEAGAKGLAWDEEHFVQYVQDPTKFLKEYTGD----AK 95
sp|P00091|CYC22_RHOPA_Cytochro      TAAGFT-YSPLNHNSGEAGLVWTADNIINYLNDPNAFLKKFLTDKGKADQ 86
sp|Q93VA3|CYC6_ARATH_Cytochrom      VDTEEEIYRVTYFGKG-RMPGFGEKCTPRGQCTFGPRLQ---------- 80
                                       .    *              :  .   .         :

sp|P99999|CYC_HUMAN_Cytochrome      --GTKMIFVGIKKKEERADLIAYLKKATNE- 105
sp|P00004|CYC_HORSE_Cytochrome      --GTKMIFAGIKKKTEREDLIAYLKKATNE- 105
sp|P0C0X8|CYC2_RHOSH_Cytochrom      AKG--KMTFKLKKEADAHNIWAYLQQVAVRP 124
sp|P00091|CYC22_RHOPA_Cytochro      AVGVTKMTFKLANEQQRKDVVAYLATLK--- 114
sp|Q93VA3|CYC6_ARATH_Cytochrom      -----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)

ETEToolkit: http://etetoolkit.org/treeview

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:

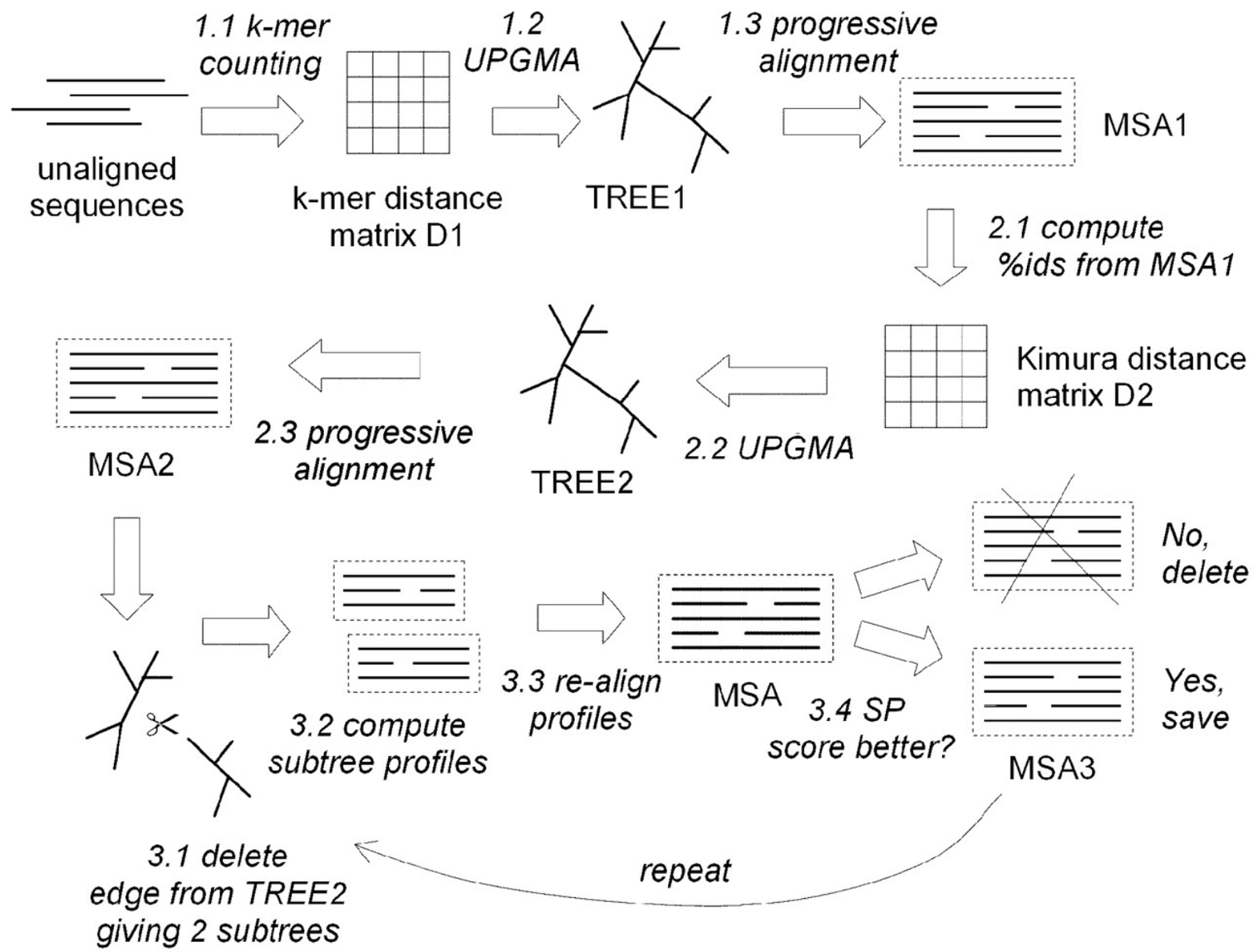
- use alignment to compute more accurate pairwise distance between sequences, Kimura distance: $-\ln(1 - D - D^2/5)$, where D is the fraction of identical bases between the pair of sequences.
- from new distance matrix, build the guide tree and a new alignment.

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.

from Lucia Moura

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)



Edgar (2004) PMID: 15034147.

Alignment with Muscle

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

```
sp|Q93VA3|CYC6_ARATH     MRLVLSGASSFTSNLFCSSQQVNGRGKELKNPISLNHNKDLDFLLKKLAPPLTAVLLAVS
sp|P99999|CYC_HUMAN      ------------------------MG----------------------------------
sp|P00004|CYC_HORSE      ------------------------MG----------------------------------
sp|P0C0X8|CYC2_RHOSH     ------------------------QEG---------------------------------
sp|P00091|CYC22_RHOPA    ----------MVKKLLTILSIAATAGSLSIGTASAQ-----------------------
                                                *

sp|Q93VA3|CYC6_ARATH     PICFPPESLGQTLDIQRGATLFNRACIGCH---------------DTGGNI---------
sp|P99999|CYC_HUMAN      ------------DVEKGKKIFIMKCSQCH-------TVEKGGKHKTGPNLHGLFGRKTG
sp|P00004|CYC_HORSE      ------------DVEKGKKIFVQKCAQCH-------TVEKGGKHKTGPNLHGLFGRKTG
sp|P0C0X8|CYC2_RHOSH     ------------DPEAGAKAFNQ-CQTCHVIVDDSGTTIAGRNAKTGPNLYGVVGRTAG
sp|P00091|CYC22_RHOPA    ------------DAKAGEAVFKQ-CMTCH----------RADKNMVGPALGGVVGRKAG
                                     * : *   *   *  **                .*  :

sp|Q93VA3|CYC6_ARATH     IQPGATLFTKDLER---NGVDTEEEIYRVTYFGKGRM--------PGFGEKCTPRGQCTF
sp|P99999|CYC_HUMAN      QAPGYS-YTAANKN---KGIIWGEDTL-MEYLENPKKYI------PG--------TKMIF
sp|P00004|CYC_HORSE      QAPGFT-YTDANKN---KGITWKEETL-MEYLENPKKYI------PG--------TKMIF
sp|P0C0X8|CYC2_RHOSH     TQADFKGYGEGMKEAGAKGLAWDEEHF-VQYVQDPTKFL-----KEYTGDAKAKGKMTF
sp|P00091|CYC22_RHOPA    TAAGFT-YSPLNHNSGEAGLVWTADNI-INYLNDPNAFLKKFLTDKGKADQAVGVTKMTF
                           .. . :          *:    :   : *. .                    :  *

sp|Q93VA3|CYC6_ARATH     -GPRLQDEEIKLLAEFVKFQADQGWPTVSTD
sp|P99999|CYC_HUMAN      VGIKKKEERADLIAYLKKATNE---------
sp|P00004|CYC_HORSE      AGIKKKTEREDLIAYLKKATNE---------
sp|P0C0X8|CYC2_RHOSH     -KLKKEADAHNIWAYLQQVAVRP--------
sp|P00091|CYC22_RHOPA    -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.

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

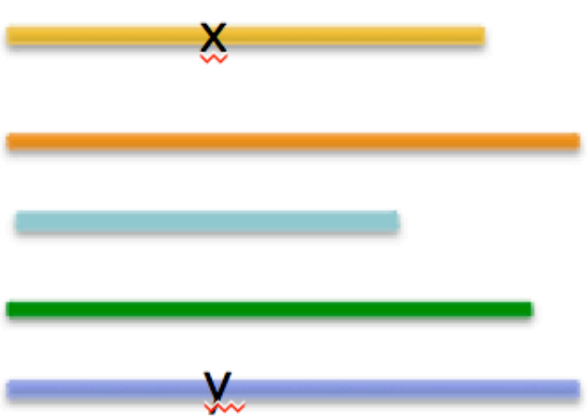
from wikipedia

Transitivity in Alignments

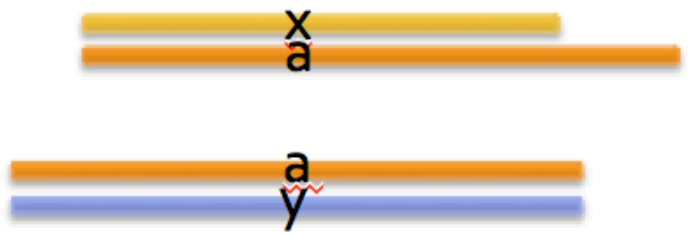
$$\forall a,b,c \in X : \left(aRb \wedge bRc\right) \Rightarrow aRc$$

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

Multiple Sequence Alignment

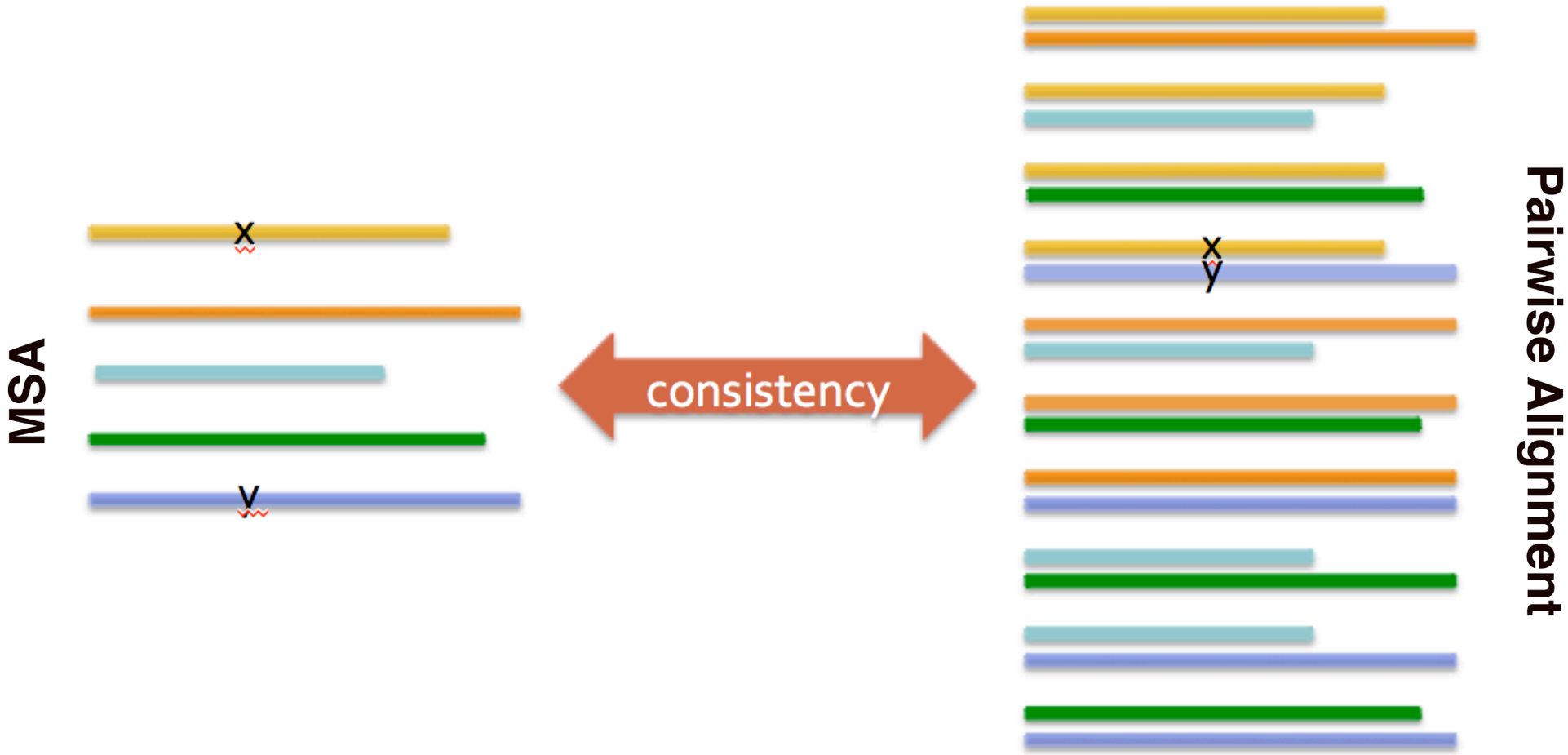


Pairwise Alignment



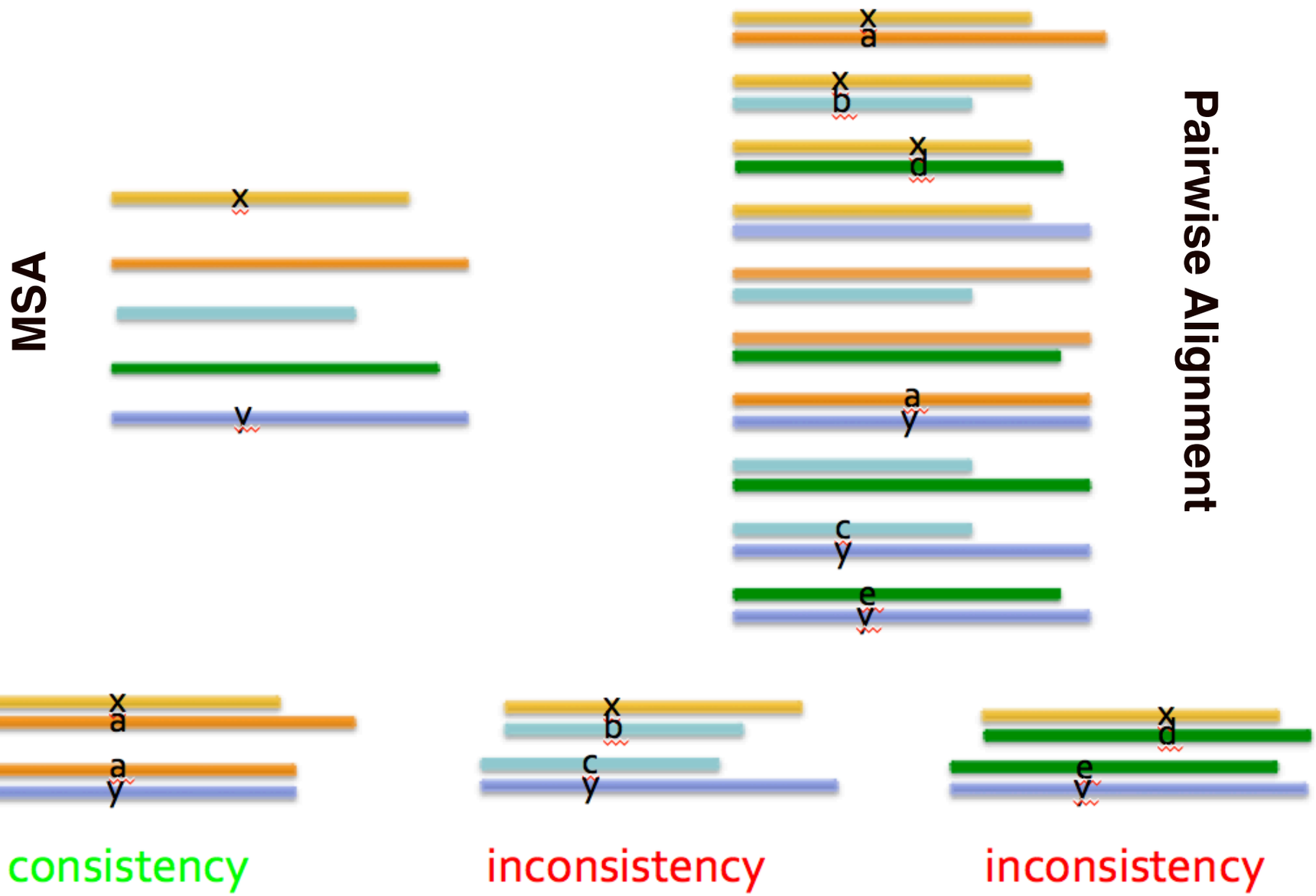
How it can be applied?

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

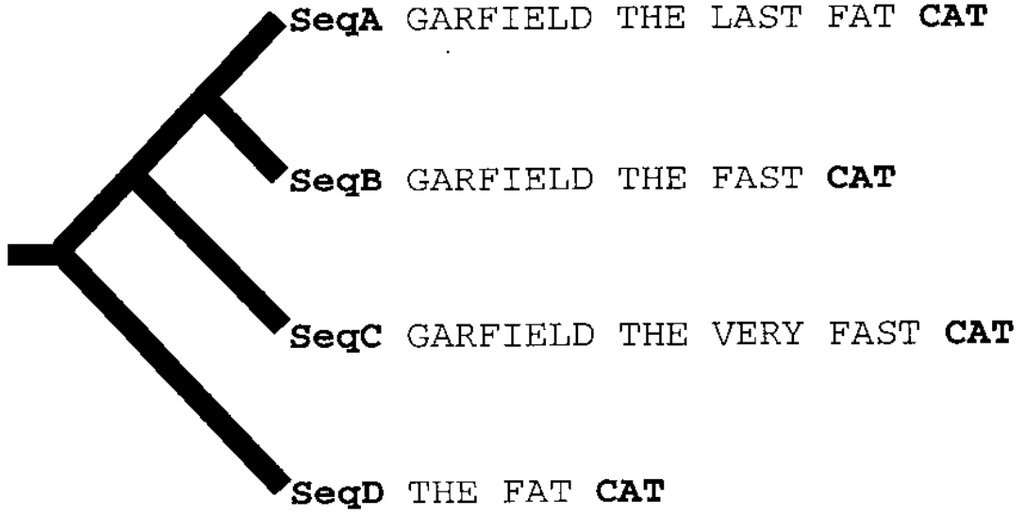
```
SeqA GARFIELD THE LAST FA-T CAT
SeqC GARFIELD THE VERY FAST CAT
```
Prim. Weight = 77

```
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 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

Notredame et al. (2000) PMID: 10964570

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
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

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

Extended Library

```
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
      ||||||||  |||      |||| ||
SeqB  GARFIELD THE       FAST CAT
```

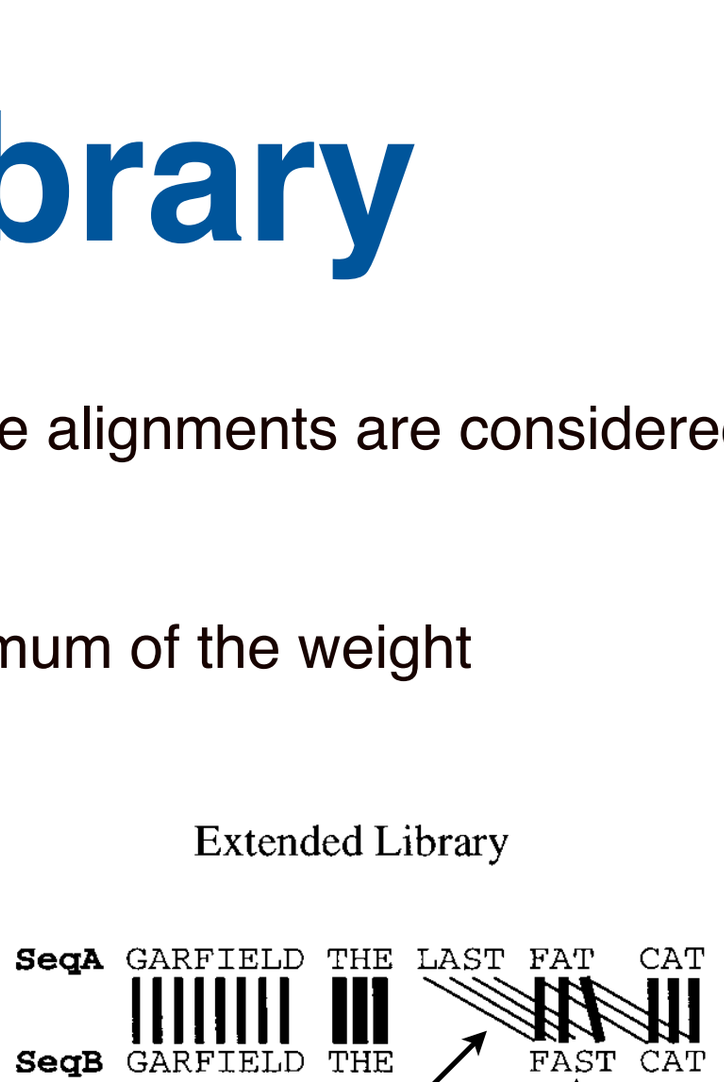
Weight = 77

Min (A-C=77, B-C=100)

```
SeqA  GARFIELD THE LAST FAT CAT
                  |||      ||| |||
SeqD              THE      FAT CAT
                  |||      ||\ \\\
SeqB  GARFIELD THE         FAST CAT
```

Weight = 100

Min (A-D=100, B-D=100)



Weight=88 Weight=177

Dynamic Programming

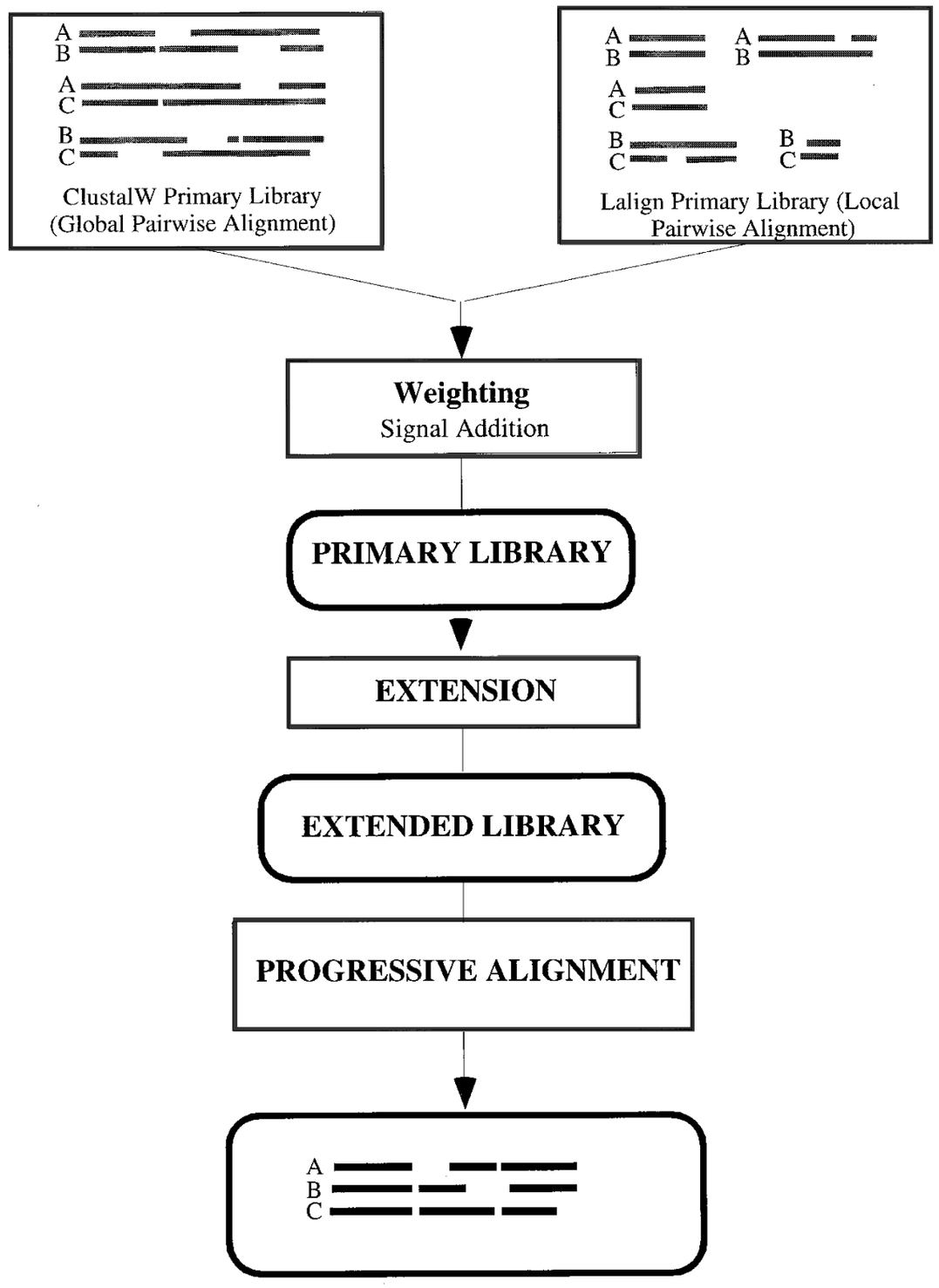
```
SeqA  GARFIELD THE LAST FA-T CAT
SeqB  GARFIELD THE ---- FAST CAT
```

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----------------------------------------KLLTILSIAATAGS
sp|P0C0X8|CYC2_RHOSH    Q------------------------------------------------------------
sp|P99999|CYC_HUMAN     ------------------------------------------------------------
sp|Q93VA3|CYC6_ARATH    MRLVLSGASSFTSNLFCSSQQVNGRGKELKNPISLNHNKDLDFLLKKLAPPLTAVLLAVS


sp|P00004|CYC_HORSE     ------------MGDVEKGKKIFVQKCAQCHTVE-------KGGKHKTGPNLHGLFGRK
sp|P00091|CYC22_RHOPA   ------LSIGTASAQDAKAGEAVFK-QCMTCHRA----------DKNMVGPALGGVVGRK
sp|P0C0X8|CYC2_RHOSH    ------------EGDPEAGAKAFN-QCQTCHVIVDDSGTTIAGRNAKTGPNLYGVVGRT
sp|P99999|CYC_HUMAN     ------------MGDVEKGKKIFIMKCSQCHTVE-------KGGKHKTGPNLHGLFGRK
sp|Q93VA3|CYC6_ARATH    PICFPPESLG--QTLDIQRGATLFNRACIGCHDTG----------GNIIQPG--------
                                   * : *   *   *  **                   *

sp|P00004|CYC_HORSE     TGQAPGFT-YTDANKN---KGITWKEETL-MEYLENPKKYI--------------PGTKM
sp|P00091|CYC22_RHOPA   AGTAAGFT-YSPLNHNSGEAGLVWTADNI-INYLNDPNAFLKKFLTDKGKADQAVGVTKM
sp|P0C0X8|CYC2_RHOSH    AGTQADFKGYGEGMKEAGAKGLAWDEEHF-VQYVQDPTKFLKEYTG------DAKAKGKM
sp|P99999|CYC_HUMAN     TGQAPGYS-YTAANKN---KGIIWGEDTL-MEYLENPKKYI--------------PGTKM
sp|Q93VA3|CYC6_ARATH    ------ATLFT---KDLERNGVDTEEEIYRVTYFGK--GRMPGFGE------KCTPRGQC
                               . :    ::    *:    :   : *. .    :                :

sp|P00004|CYC_HORSE     IFAGIKKKTEREDLIAYLKK---------ATNE
sp|P00091|CYC22_RHOPA   TFK-LANEQQRKDVVAYL-----------ATLK
sp|P0C0X8|CYC2_RHOSH    TFK-LKKEADAHNIWAYLQQ--------VAVRP
sp|P99999|CYC_HUMAN     IFVGIKKKEERADLIAYLKK---------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

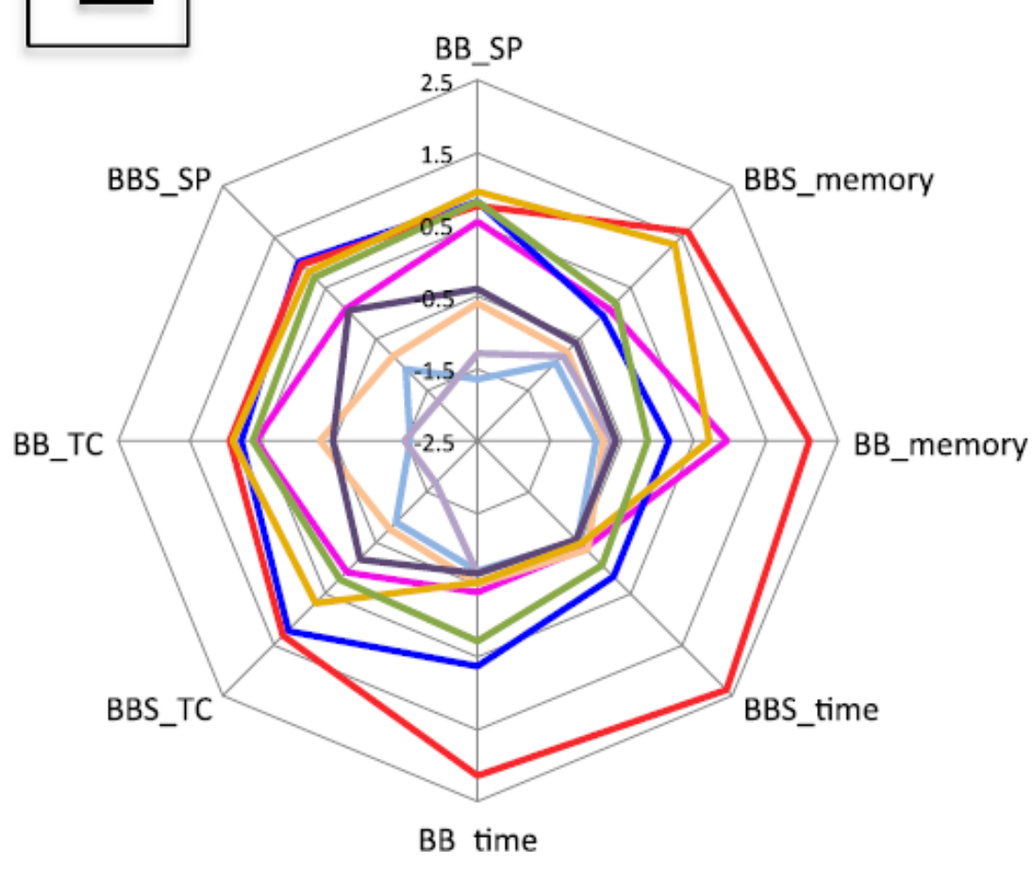
Time of execution

Peak memory

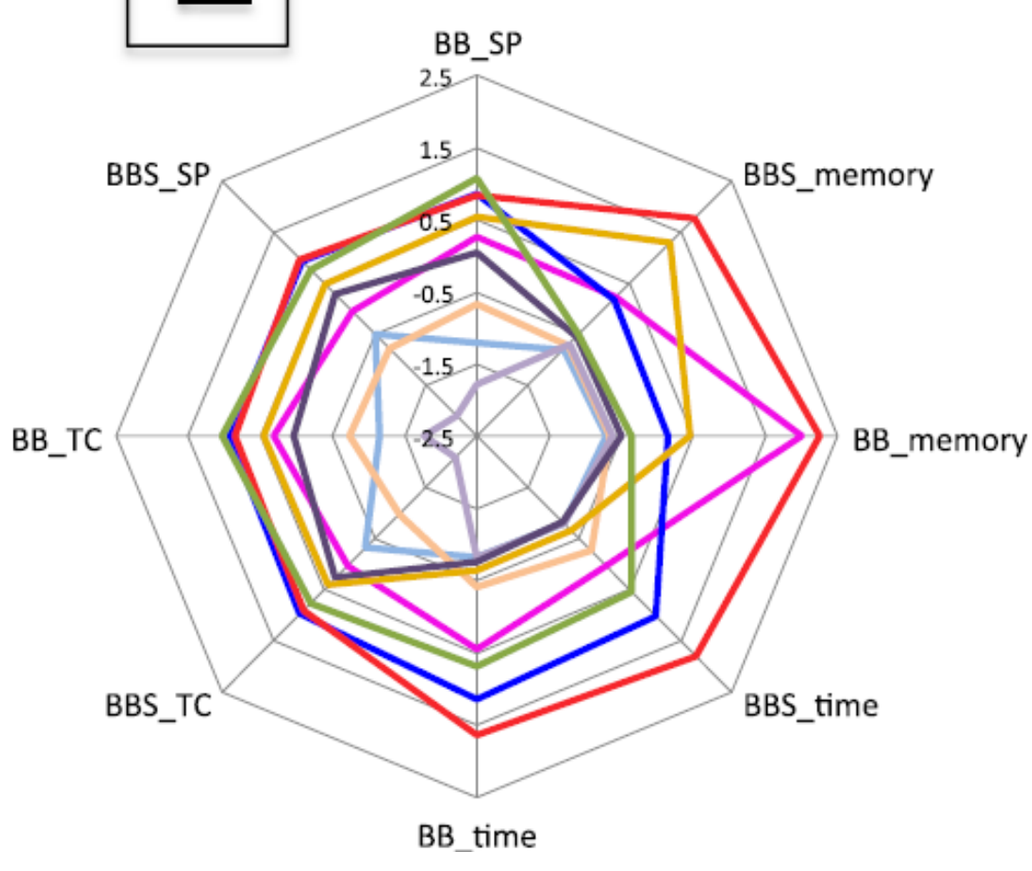
Pais et al. (2014) PMID: 24602402.

Performance (I)

A) Z-scores **%id < 20**



B) Z-scores **20 < %id < 40**

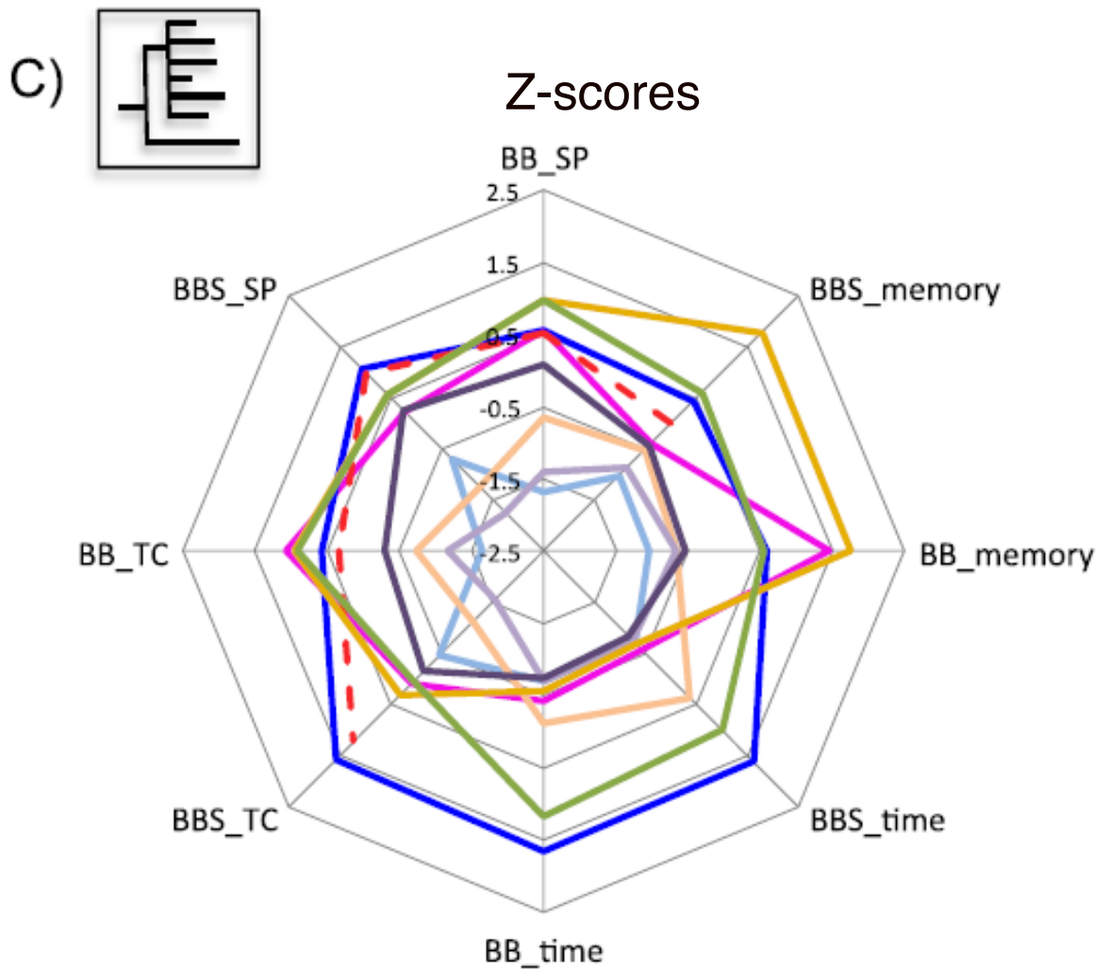


CLUSTALW CLUSTAL O DIALIGN-TX POA MUSCLE

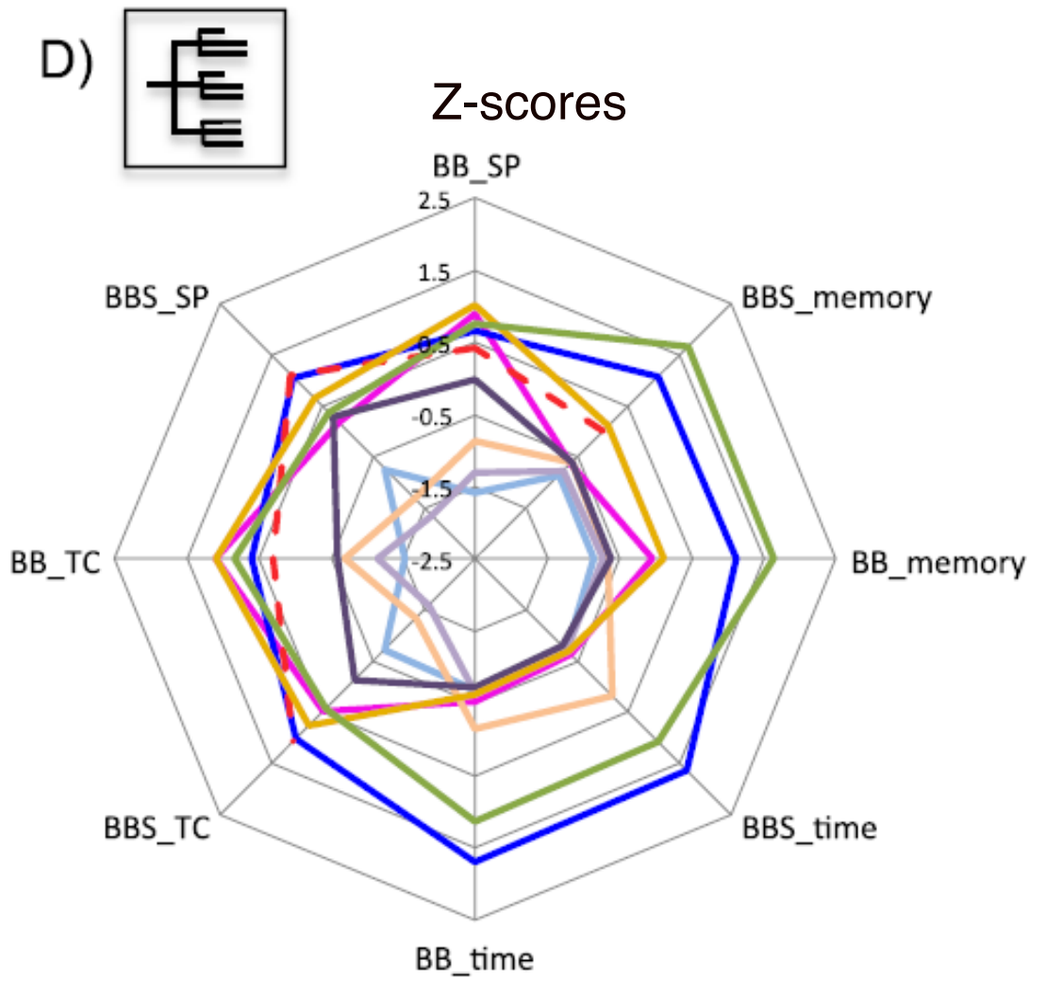
Probcons T-Coffee MAFFT Probalign

Performance (II)

C)



D)

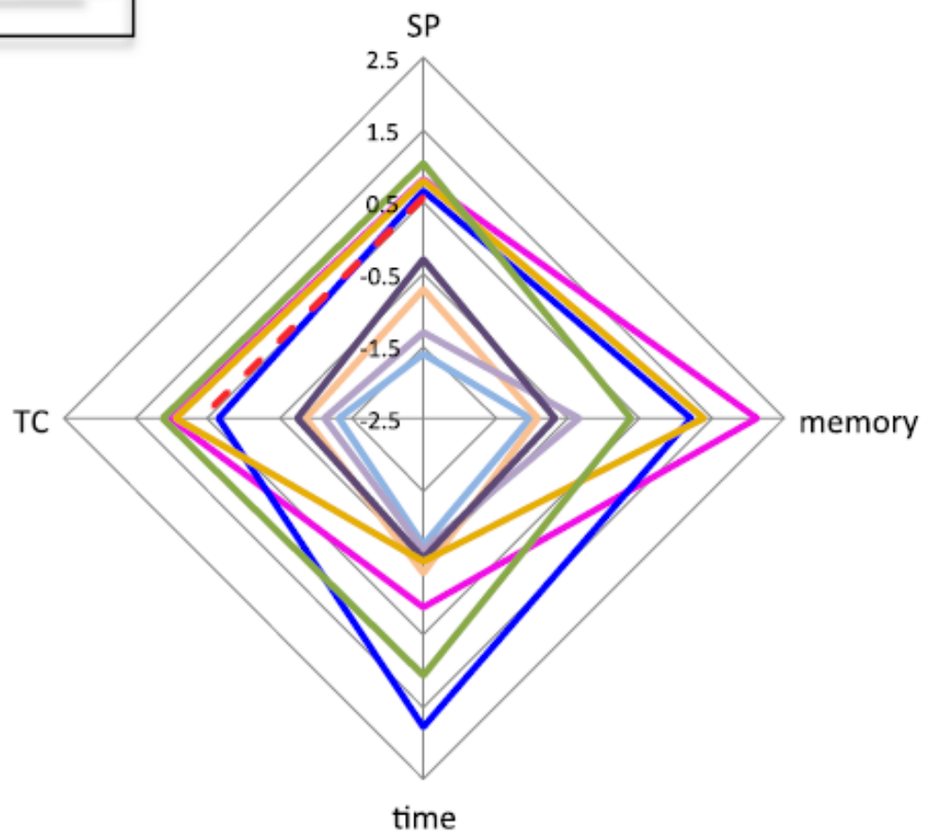


CLUSTALW
CLUSTAL O
DIALIGN-TX
POA
MUSCLE
Probcons
T-Coffee
MAFFT
Probalign

Performance (III)

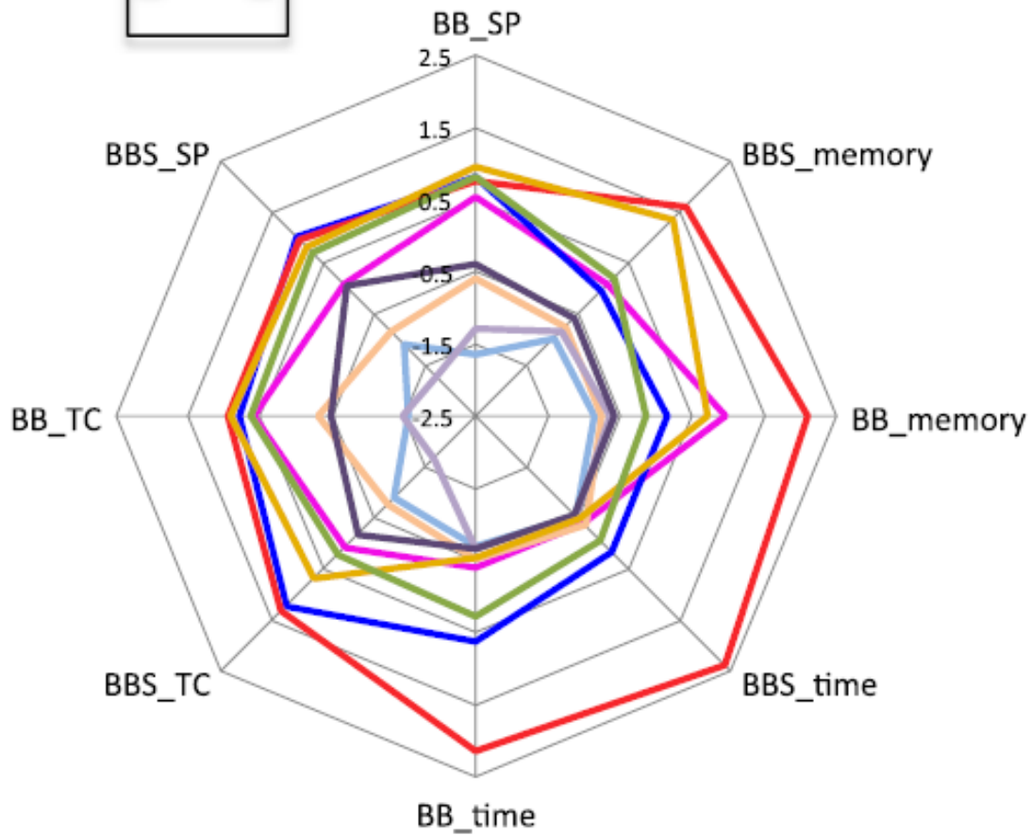
E)

Z-scores



F)

Z-scores



CLUSTALW
CLUSTAL O
DIALIGN-TX
POA
MUSCLE
Probcons
T-Coffee
MAFFT
Probalign

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.

Pais et al. (2014) PMID: 24602402.

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

- Analyze the conserved positions in the alignments
- Repeat with the Arabidopsis (Q93VA3) and the human (P99999) sequences
- Compare the results and in particular the pattern of conserved residues