

# Multiple Sequence Alignment

Bartek Wilczyński

March 24<sup>th</sup>, 2020

# How sequences evolve?

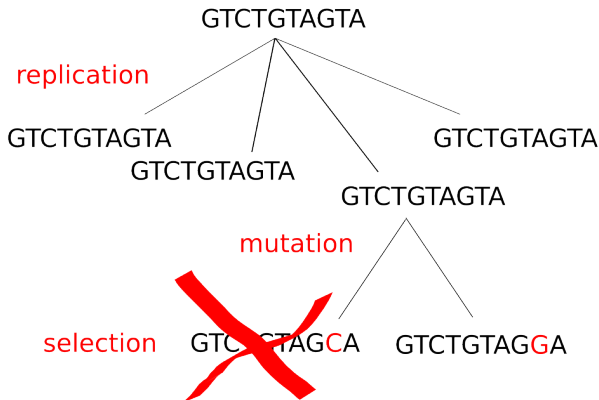


image (c)

BW

## Reconstructing alignments

Reminder on  
alignmentsMultiple  
alignments

		H	E	A	G	A	W	G	H	E	E
P	0	← -8	← -16	← -24	← -32	← -40	← -48	← -56	← -64	← -72	← -80
A	-8	↑ -2	↑ -9	← -17	← -25	← -33	← -41	← -49	← -57	← -65	← -73
W	-16	↑ -10	↑ -3	← -4	← -12	← -20	← -28	← -36	← -44	← -52	← -60
H	-24	↑ -18	↑ -11	← -6	← -7	← -15	← -5	← -13	← -21	← -29	← -37
E	-32	↑ -14	↑ -18	← -13	← -8	← -9	← -13	← -7	← -3	← -11	← -19
E	-40	↑ -22	↑ -8	← -16	← -16	← -9	← -12	← -15	← -7	← 3	← -5
A	-48	↑ -30	↑ -16	← -3	← -11	← -11	← -12	← -12	← -15	← -5	← 2
E	-56	↑ -38	↑ -24	← -11	← -6	← -12	← -14	← -15	← -12	← -9	← 1

HEAGAWGHE-E

--P-AW-HEAE

image (c) Durbin et al.

## Greedy tree inference – example

	A	B	C	D
A	0	17	21	27
B		0	12	18
C			0	14
D				0

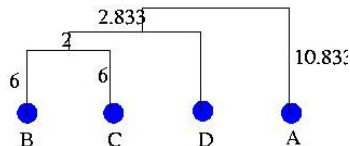


image (c) P. Winter

# Inconsistencies with pairwise alignments

Pairwise alignments used to calculate distances (and reconstruct a tree) may lead to inconsistent picture. For example consider alignment of all pairs of 3 sequences: CAAC, AACA, ACAA

- CAAC- AACA- ACAA-
- -AACA -ACAA -CAAC

Which C in CAAC was in the ancestral sequence?

## What is the solution to those inconsistencies?

Can we make a generalization of the pairwise alignment idea?

Q5E940	BOVIN	MPREDRATKSNYSFYKLTIIQLDDDPKCFYVGA	ADVNGVKQM00IMSLRGK	AVVLMGKNTMMRKATRGHLNN	-PALE	76	
RLAO	HUMAN	MPREDRATKSNYSFYKLTIIQLDDDPKCFYVGA	ADVNGVKQM00IMSLRGK	AVVLMGKNTMMRKATRGHLNN	-PALE	76	
RLAO	MOUSE	MPREDRATKSNYSFYKLTIIQLDDDPKCFYVGA	ADVNGVKQM00IMSLRGK	AVVLMGKNTMMRKATRGHLNN	-PALE	76	
RLAO	RAT	MPREDRATKSNYSFYKLTIIQLDDDPKCFYVGA	ADVNGVKQM00IMSLRGK	AVVLMGKNTMMRKATRGHLNN	-PALE	76	
RLAO	CHICK	MPREDRATKSNYSFYKLTIIQLDDDPKCFYVGA	ADVNGVKQM00IMSLRGK	AVVLMGKNTMMRKATRGHLNN	-PALE	76	
RLAO	RANBY	MPREDRATKSNYSFYKLTIIQLDDDPKCFYVGA	ADVNGVKQM00IMSLRGK	AVVLMGKNTMMRKATRGHLNN	-PALE	76	
Q7ZUG3	BRARE	MPREDRATKSNYSFYKLTIIQLDDDPKCFYVGA	ADVNGVKQM00IMSLRGK	AVVLMGKNTMMRKATRGHLNN	-PALE	76	
RLAO	ICTH	MPREDRATKSNYSFYKLTIIQLDDDPKCFYVGA	ADVNGVKQM00IMSLRGK	AVVLMGKNTMMRKATRGHLNN	-PALE	76	
RLAO	DROME	MYRHNKAQAQYIKVYIDFDFKCFYVGA	ADVNGVKQM00IMSLRGK	AVVLMGKNTMMRKATRGHLNN	-PALE	76	
Q5ALP0	DIAB	MSAG-SKKRVYIKATKLTITITDKMIAVAIV	YCS00IKKISIRGI	GAIVMGKMKIRKIVLADSKP	-PDEL	75	
RLAO	PICPB	MAKLSKQKQOMYIKLSLQIQSKSLIVYD	YDMSVASYKSLRGK	AVVLMGKNTMMRKATRGHLNN	-PALE	76	
RLAO	SULAC	HIGLAVTTTKKIAKRVDEVAIVIKKTKTIT	ITIANTFGPAADKIEIKKLRGK	ADIVYKKNLFIKAKNAG	-IDKX	79	
RLAO	SULTO	HRINAVITTKKIAKRVDEVAIVIKKTKTIT	ITIANTFGPAADKIEIKKLRGK	ADIVYKKNLFIKAKNAG	-IDKX	79	
RLAO	SULO	MKRILALAKQKRVKVAIVIKKTKTITIT	ITIANTFGPAADKIEIKKLRGK	ADIVYKKNLFIKAKNAG	-IDKX	79	
RLAO	AERPE	MSVSYLVGYMKHYKIPKWTMLRELE	LFPSKRVFLVADITITVYVYKIKLWK	YKVMIAKRTILIAKHAAGLE	-LDIE	80	
RLAO	PYRAE	HMALGKIKRYVIRBQYPAKRYKIVSEAT	IKQYFYFDFLGLSIRLIEYVIRLAY	GVIKYKTLFLKIAFTYKGG	-TPAL	85	
RLAO	METAC	MAERHHHTETQWKKDEINIKELIQSKYK	GMVGLIGLATAKQITRDLDV	AVLVKRTITLIEBALNQLQ	-ETIP	78	
RLAO	METMA	MAERHHHTETQWKKDEINIKELIQSKYK	GMVGLIGLATAKQITRDLDV	AVLVKRTITLIEBALNQLQ	-ETIP	78	
RLAO	ACBFU	MAAYVGS	DFYFVRAVEIKIMISSEYVAIVSY	YKGMOKHIEIRFGK	ADIVYKTLIEBALDAG	-GQFL	75
RLAO	METKA	MAAYVGS	DFYFVRAVEIKIMISSEYVAIVSY	YKGMOKHIEIRFGK	ADIVYKTLIEBALDAG	-GQFL	75
RLAO	METW	MAHVAHWKKKEVEELANKIKSVYAL	VDVDSMYPATYLSQMRILIREN	GOLLVRSNTILIELAIKKAAKGLPE	-LWFD	74	
RLAO	METTL	MAHVAHWKKKEVEELANKIKSVYAL	VDVDSMYPATYLSQMRILIREN	GOLLVRSNTILIELAIKKAAKGLPE	-LWFD	74	
RLAO	METVA	MAHVAHWKKKEVEELANKIKSVYAL	VDVDSMYPATYLSQMRILIREN	GOLLVRSNTILIELAIKKAAKGLPE	-LWFD	74	
RLAO	METJA	MAHVAHWKKKEVEELANKIKSVYAL	VDVDSMYPATYLSQMRILIREN	GOLLVRSNTILIELAIKKAAKGLPE	-LWFD	74	
RLAO	PYRAB	MAHVAHWKKKEVEELANKIKSVYAL	VDVDSMYPATYLSQMRILIREN	GOLLVRSNTILIELAIKKAAKGLPE	-LWFD	74	
RLAO	PYRBO	MAHVAHWKKKEVEELANKIKSVYAL	VDVDSMYPATYLSQMRILIREN	GOLLVRSNTILIELAIKKAAKGLPE	-LWFD	74	
RLAO	PYRPU	MAHVAHWKKKEVEELANKIKSVYAL	VDVDSMYPATYLSQMRILIREN	GOLLVRSNTILIELAIKKAAKGLPE	-LWFD	74	
RLAO	PYRKO	MAHVAHWKKKEVEELANKIKSVYAL	VDVDSMYPATYLSQMRILIREN	GOLLVRSNTILIELAIKKAAKGLPE	-LWFD	74	
RLAO	HALMA	MSAESEKRTETQWKKDEYDAIV	MTESYSGVYVNIAGIDPDLQ	DMHDLRGL	AAVLRVRSNTILIEBALDDVD	-DGLF	79
RLAO	HALVO	MSAESEKRTETQWKKDEYDAIV	MTESYSGVYVNIAGIDPDLQ	DMHDLRGL	AAVLRVRSNTILIEBALDDVD	-DGLF	79
RLAO	HALSA	MSAESEKRTETQWKKDEYDAIV	MTESYSGVYVNIAGIDPDLQ	DMHDLRGL	AAVLRVRSNTILIEBALDDVD	-DGLF	79
RLAO	HALAC	MSAESEKRTETQWKKDEYDAIV	MTESYSGVYVNIAGIDPDLQ	DMHDLRGL	AAVLRVRSNTILIEBALDDVD	-DGLF	79
RLAO	THEVO	MKKIDPKKKEIYSELADITKSKAVALIV	YKGVN00IAKAKRQK	ADIVYKTLFLKIAFALDSND	-EKIF	72	
RLAO	PICFO	MTEPDAKIDFYKNHIEINRSKRVAAIV	YSKLNH00IKSIRGK	ADIVYKTLFLKIAFALDSND	-NNYV	72	

*Note: the correspondence with edit distance is lost*

image (c) P. Winter

# How to score multiple alignments?

What is the natural way to score multiple alignment “quality”?

- Assume column independence (as usual)
- Sum of pairs (SP) score

$$S(m_i) = \sum_{k < l} s(m_i^k, m_i^l)$$

- Does it work well (for counting parsimonous mutations)?  
*hint: Consider a column with all characters different.*
- How much does it overestimate the number of necessary mutations?

# Can we use dynamic programming?

$$\alpha_{i_1, i_2, \dots, i_N} = \max \left\{ \begin{array}{ll} \alpha_{i_1-1, i_2-1, \dots, i_N-1} & + S(x_{i_1}^1, x_{i_2}^2, \dots, x_{i_N}^N), \\ \alpha_{i_1, i_2-1, \dots, i_N-1} & + S(-, x_{i_2}^2, \dots, x_{i_N}^N), \\ \alpha_{i_1-1, i_2, i_3-1, \dots, i_N-1} & + S(x_{i_1}^1, -, \dots, x_{i_N}^N), \\ & \vdots \\ \alpha_{i_1-1, i_2-1, \dots, i_N} & + S(x_{i_1}^1, x_{i_2}^2, \dots, -), \\ \alpha_{i_1, i_2, i_3-1, \dots, i_N-1} & + S(-, -, \dots, x_{i_N}^N), \\ & \vdots \\ \alpha_{i_1, i_2-1, \dots, i_{N-1}-1, i_N} & + S(-, x_{i_2}^2, \dots, -), \\ & \vdots \end{array} \right.$$

image (c) Durbin et al



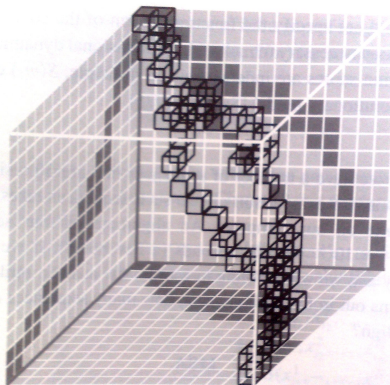
## Sum of Pairs is NP-complete...

- Dynamic algorithm has the cost of  $\mathcal{O}(n^k)$
- In general, the problem is NP-Complete
- We can still try to slightly improve its performance by looking at the lower bound of alignment score (Carillo-lipman)

$$\sigma(a) \leq S(a^{kl}) - S(\hat{a}^{kl}) + \sum_{k' < l'} S(\hat{a}^{k'l'})$$

$$\begin{aligned} S(a^{kl}) &\geq \beta^{kl} \\ \text{where } \beta^{kl} &= \sigma(a) + S(\hat{a}^{kl}) - \sum_{k' < l'} S(\hat{a}^{k'l'}). \end{aligned}$$

# Carillo-Lippman lower bound method



**Figure 6.3** Carrillo & Lipman's algorithm allows the search for optimal alignments to be restricted to a subset of the multidimensional programming matrix, shown here as three-dimensional. The sets  $B^{kl}$  are shown in dark grey, and the cells in the matrix to which the search can be confined are outlined in black.

# Feng-Doolittle ('87) greedy approach

Reminder on  
alignments

Multiple  
alignments

- We can use the greedy approach similar to UPGMA
- In each step we choose the nearest pair (using pairwise sequence distances)
- And we can merge alignments based on the pairwise alignment between the sequences
- Uses the principle of “once a gap, always a gap”.
- *We need to “elegantly” align alignments of more than one sequence*

In the process of incremental alignment we need to align profiles (alignments)

$$\begin{aligned}\sum_i S(m_i) &= \sum_i \sum_{k < l \leq N} s(m_i^k, m_i^l) \\ &= \sum_i \sum_{k < l \leq n} s(m_i^k, m_i^l) + \sum_i \sum_{n < k < l \leq N} s(m_i^k, m_i^l) + \sum_i \sum_{k \leq n, n < l \leq N} s(m_i^k, m_i^l).\end{aligned}$$

image (c) Durbin et al

## Algorithm: CLUSTALW progressive alignment

- (i) Construct a distance matrix of all  $N(N - 1)/2$  pairs by pairwise dynamic programming alignment followed by approximate conversion of similarity scores to evolutionary distances using the model of Kimura [1983].
- (ii) Construct a guide tree by a neighbour-joining clustering algorithm by Saitou & Nei [1987].
- (iii) Progressively align at nodes in order of decreasing similarity, using sequence–sequence, sequence–profile, and profile–profile alignment. ◁

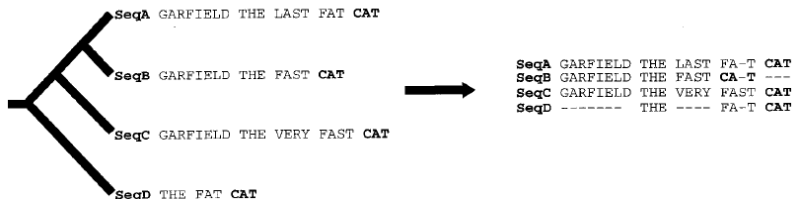
image (c) Durbin et al

# ClustalW improvements (1994)

- Sequences might be weighted in the profile alignments
- Different substitution matrices might be used at different levels of merging
- Gap scores are now dependent on the AA removed i.e.  
 $s(-, x) \neq s(-, y)$

# Incremental alignment problems

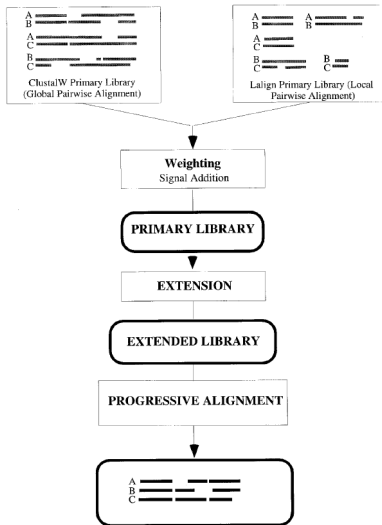
## a) Regular Progressive Alignment Strategy



# T-Coffee algorithm for the rescue (Notredame, 2000)

Reminder on  
alignments

Multiple  
alignments





## T-Coffee in more detail

## b) Primary Library

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

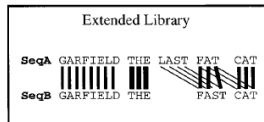
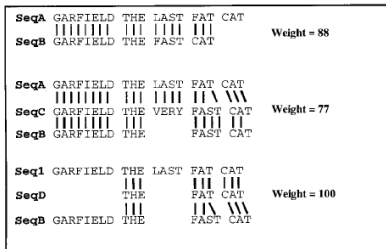
SeqA GARFIELD THE LAST FAT CAT Prim. Weight = 100  
 SeqD ----- THE ---- FAT CAT

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

## c) Extended Library for seq1 and seq2



Dynamic Programming

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

# Muscle algorithm (Edgar 2004)

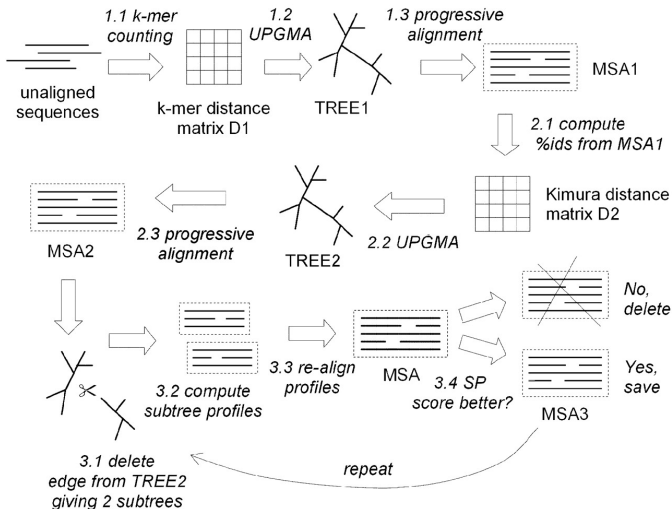


image (c) RC. Edgar, NAR 2004